TÉCNICO LISBOA

Extraction of structural and semantic features for the identification of Psychosis in European Portuguese

Rodrigo Borges Pessoa de Sousa

Thesis to obtain the Master of Science Degree in

Information Systems and Computer Engineering

Supervisors: Prof. Helena Sofia Andrade Nunes Pereira Pinto
Prof. Alberto Abad Gareta

Examination Committee

Chairperson: Prof. Luís Manuel Antunes Veiga
Supervisor: Prof. Helena Sofia Andrade Nunes Pereira Pinto
Member of the Committee: Prof. Isabel Maria Martins Trancoso

November 2022

This work was created using LaTeX typesetting language
in the Overleaf environment (www.overleaf.com).

Acknowledgments

First of all, I would like to acknowledge and show my appreciation for the impressive work and availability of Doctor Daniel Neto, Doctor Joaquim Gago, and Doctor Ana Moreira without whom the execution of this work would have been impossible. Without these health professionals not only the work would not have been accomplished but we would have not gained important insights into the relevance and seriousness of the domain in which our work inserts itself into.

The work carried out also relied substantially on the help and support provided by the entire team of Centro Hospitalar de Lisboa Ocidental - Unidade de Saúde Mental de Oeiras. The entirety of the team facilitated our interaction with the patients and always showed full availability to discuss any pertinent topics and solve any problems that arose during our stay.

Additionally, I would like to thank professor Helena Sofia Pinto and professor Alberto Abad for the overview and guidance provided throughout the entirety of our work. Both, together, provided the attention to detail and technical skills required for our work to be accomplished. Without a doubt, without them together, as advisors, the work would not have reached as good and meaningful results as it reached.

Furthermore, I would not have endured and surpassed the entirety of this phase without friends and family support. My family always provided the unconditional love, affirmation, and personal skills needed to beat any challenge that might have arisen. Without all that you have imbued and conveyed into me, I would not be the person I am today and would not be where I am today.

Last but not least, I would like to show my particular appreciation for my girlfriend, Isabel Soares. Without her support and motivation, the work would not have been accomplished. My girlfriend helped me gain focus on what was important when I was overwhelmed and provided me with the self-assurance and tranquility that I needed when all seemed lost.

Abstract

Psychosis is a mental condition that affects the subject's behavior and perception of the world, impairing both his cognitive and speech capabilities. These impairments, when associated with the stigma that mental disorders carry in society, promote the subject's disconnection from reality and society. Psychosis, and other mental disorders, lack efficient and accurate diagnostic tools, relying mostly on self-reports from patients, their families, and specialized clinicians.

Studies identified by us, typically focus on the identification or prediction of psychosis through surface-level analysis of the subject's speech targeting audio, time, and paucity features. Recent studies have started focusing on high-level and complex language analysis such as semantics, structure, and content. A very limited number of studies have targeted the Portuguese language. To the best of our knowledge, no study has focused on structural or semantic features in European Portuguese, which is our main objective. Our work also aimed at expanding the *First European Corpus for Psychosis Identification* with more subjects and a more diverse control group, to better understand the impact that such alterations have on the results achieved by the previous studies.

Results obtained support future developments of models for the identification of psychosis, that rely on structure, coherence, and content analysis of discourse. However, it also suggested that further developments in Natural Language Processing techniques, especially for European Portuguese, are required for the progression and improvement of the results obtained. The models developed by our solution were verified to be reliable and robust.

Keywords

Psychosis; Schizophrenia; Coherence Analysis; Structure Analysis; Content Analysis; Valence Analysis; Natural Language Processing; Classification;

Resumo

Psicose é definida como uma pertubação mental que afeta o comportamento do paciente e a sua percepção do mundo, afetando tanto as suas capacidades cognitivas como de discurso. Estas perturbações e o estigma de que estão acompanhadas, promovem o isolamento do paciente do resto da sociedade. Distúrbios mentais carecem de métodos de diagnóstico eficientes e precisos, baseando-se em relatos dos pacientes, das suas famílias, ou clínicos altamente especializados.

Estudos identificados pela nossa equipa, focam-se na identificação e predição de psicose através de análises de baixo nível do discurso de partipantes, focando-se em features associadas com o som, ritmo, e pausas do discurso. Estudos mais recentes focaram-se em análises de alto nível, analisando semântica, estrutura, e conteúdo. Um número muito limitado de estudos focou-se na Língua Portuguesa. Tanto quanto a equipa sabe, nenhum estudo se focou na análise estrutural e semântica de Português Europeu, e é por isso, este o nosso objetivo. Também temos como objectivo a expansão do *First European Corpus for Psychosis Identification* com mais participantes e com um grupo de controlo mais diversificado, de forma a perceber o impacto destas alterações.

Os resultados obtidos no nosso trabalho suportam futuros desenvolvimentos de modelos para a idenficação de psicose, que se baseiem na análise da estrutura, coerência, e conteúdo do discurso. Por outro lado, também sugere que são necessários desenvolvimentos no ramo da Análise e Processamento da Língua, especialmente para Português Europeu, de forma a atingir melhores resultados. Os modelos desenvolvidos pela nossa solução verificaram ser fiáveis robustos.

Palavras Chave

Psicose; Esquizofrenia; Análise de Coerência; Análise da Estrutura; Análise do Conteúdo; Análise da Valência; Processamento de Língua Natural; Classificação;

Contents

List of Figures

List of Tables

Listings

Acronyms

API Application Program Interface

BERT Bidirectional Encoder Representations from Transformers

BP Brazilian Portuguese

BPRS Brief Psychiatric Rating Scale

CB Committed Belief

CBF Cerebral Blood Flow

CETEMPúblico Corpus de Extractos de Texto Electrónicos MCT/Público

CHLO Centro Hospitalar de Lisboa Ocidental

CHR+ clinical high risk

CHR- non-clinical high risk

CPU Central Processing Unit

CSSJD Casa da Saúde S. João de Deus

DT Decision Trees

EOS End of Sentence

EP European Portuguese

FEP first episode of psychosis

FFNN Multi-Layer Feed-Forward Neural Networks

FOC First Order Coherence

FTD Formal Thought Disorder

GeMAPS Geneva Minimalistic Acoustic Parameter Set

GPT Generative Pre-trained Transformer

GPU Graphical Processing Unit

GUI	Graphical User Interface
HLT	Human Language Technology Lab
HTML	Hypertext Markup Language
INESC-ID	Instituto de Engenharia de Sistemas e Computadores - Investigação e Desenvolvimento
IP	Internet Protocol
IST	Instituto Superior Técnico
LCA	Latent Content Analysis
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
LSCC	Largest Strongly Connected Component
LCB	Level of Committed Belief
MCT	Ministério da Ciência e da Tecnologia
MLP	Multi-Layer Perceptrons
MRIs	Magnetic Resonance Imaging
NA	Non-Attributable Belief
NB	Naive Bayes
NCB	Non-committed Belief
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
POS	Part of Speech
RAM	Random Access Memory
RoBERTa	Robustly Optimized BERT Pretraining Approach
RF	Random Forest
RNN	Recursive Neural Networks
ROB	Reported Belief
SANS	Scale for the Assessment of Negative Symptoms
SAPS	Scale for the Assessment of Positive Symptoms
SCCs	Strongly Connected Components
SOC	Second Order Coherence

SRL	Semantic Role Labeling
SVD	Singular Value Decomposition
SVM	Support Vector Machines
TLI	Thought and Language Index
UAR	Unweighted Average Recall

1

Introduction

Contents

In 2017, it was reported that at the time around twenty million people suffered from psychosis, which represents 0.3% of the world population [1]. Specifically, regarding Portugal, a recent study reported that around 3% to 4% of the Portuguese population has suffered at least once from psychotic disorders [2].

Definition 1 (Psychotic disorders)**.** *"Psychoses, including schizophrenia, are characterized by distortions in thinking, perception, emotions, language, sense of self, and behavior. Common psychotic experiences include hallucinations (hearing, seeing, or feeling things that are not there) and delusions (fixed false beliefs or suspicions that are firmly held even when there is evidence to the contrary)"* - World Health Organization[1].

From the definition 1 of psychotic disorders, we understand that such disorders are characterized by the loss and deterioration of one's cognitive function such that the subject becomes less tied to reality itself, prone to hallucinations and other cognitive deviations that are typically identified through speech. In extreme cases, the disorder can lead to the deterioration of one's integration into society of the stigma that it propagates.

1.1 Motivation

Psychosis is marked by various symptoms which can be grouped according to categories. The **first category** of symptoms describes behavioral symptoms. Patients diagnosed with psychosis report hallucinations, typically regarding imaginary voices speaking to them and imaginary entities, and sudden reclusiveness and unwillingness to talk or expand on questions. The **second category** concerns speech noticeable symptoms, such as the longer duration of pauses, where these pauses occur in terms of sentence structure, and frequent and unnatural repetitions. Finally, the **last category** concerns content symptoms such as disorganized or completely incoherent speech, excessive use of determiners and pronouns, with the latter sometimes not being clearly associated with a previously stated entity, and poor speech, marked by sentences that are short and have limited complexity.

Due to inherent aspects of the disorder and the stigma propagated, patients who suffer from psychosis feel isolated from the rest of society. Insel [3] reported that less than 20% of diagnosed patients are employed, more than 30% are homeless, and these patients are three times more likely to be incarcerated than the remaining population (statistics for the United States of America in 2010). The successful reintegration and recovery of diagnosed patients can be improved by early detection through preceding indicators [3], which justifies the need for changes in diagnostic tools in terms of their efficiency.

Currently, diagnosis relies on self-reports from patients regarding their symptoms and on trained clinicians' identification of speech metrics and abnormalities from these interviews. Self-reports from

[1] https://www.who.int/news-room/fact-sheets/detail/mental-disorders accessed on January 6th, 2022

patients are unreliable since they come directly from the perspective of the patient. It is imperative that diagnosis is made by trained clinicians, but this requirement entails that early diagnosis is more difficult. Therefore, new research should aim at supporting their diagnosis and possibly pre-screening patients, for later confirmation, through techniques that rely on general interviews to be carried out by either general practitioners, not fully trained specialists, or other entities, sensible and conscious of the disorder but without the full specialization that is required for a diagnosis.

Computerized-aided diagnoses have long been introduced into other medical specializations. Even though studies have proposed solutions to the domain of mental disorders and specifically psychosis, none has been employed in real diagnosis. Such solutions have distinguished control from diagnosed patients [4–10] and even predicted future psychosis [11–13]. Computerized solutions would improve diagnosis prevention efforts due to their: efficiency, lower requirements in specialized training needed for their use, and ability to detect humanly imperceptible speech deviations.

Previous research either differentiates controls from diagnosed patients [4–13], patients with varying degrees of the disorder [14–17], or sometimes, although rarely, patients with different disorders [18–20]. When using computerized solutions, it is important that they consider the various disorders and their prevalence in the world, distinguishing them, and that such solutions are studied worldwide. Only one study has focused on the European Portuguese language, Forjó et al. [5]. The authors collected the first European Portuguese speech corpus with patients diagnosed with psychosis and mentally healthy subjects, following a protocol that does not require any private or medical information, and automatically generated transcriptions. Forjó et al. [5] analyzed the data distinguishing patients from mentally healthy controls, by transcribing their recordings and processing audio features extracted with *GeMAPS* [21].

1.2 Approach

Our work took as a starting point the work of Forjó et al. [5]. Our purpose was to extend the aforementioned work by exploring discourse's coherence, semantics, and content to evaluate if results could be supported or improved with these features. We extended the current corpus with recordings from patients diagnosed with psychosis and subjects who are diagnosed with other mental disorders, in our case, bipolar disorder in its various stages. By expanding the current European Portuguese corpus for psychosis, with data from more subjects, we could conclude whether the results obtained for the corpus by Forjó et al. [5] reflected inherent differences amongst psychotic subjects and the rest of the population or whether the classifier adjusted to other factors. The authors' study did not reflect the diversity that exists in the world, since, in the study, a subject is either healthy, showing no other disorders or signs of thereof, or is diagnosed patient with psychosis. Therefore, it is possible that the developed classifier targeted other factors, possibly, side-effects of the medication, which have been shown to affect and be

detectable in diagnosis [22].

Our work did not aim to distinguish the various mental disorders, but only to identify which subjects are diagnosed with psychosis in a population of healthy controls, patients diagnosed with psychosis, and patients diagnosed with other mental disorders. In Portugal, there is no contact between clinicians and patients prior to the first episode of psychosis (FEP). For this reason, our work aimed only at identifying patients already diagnosed with psychosis and not at predicting psychosis.

1.3 Contributions

Due to the work carried out by our team, the *First European Portuguese Corpus for Psychosis Identification* was expanded by a factor of 86.96%. The original corpus was made up of 92 subjects whereas the final version, achieved by our work, is composed of 172 subjects. The corpus now approximates much more closely a real-world scenario of diagnosis, since it is made up of healthy controls, patients diagnosed with psychosis, and patients diagnosed with other mental disorders, in this case, bipolar patients. The corpus currently is made up of 35.91% healthy controls, 46.96% patients diagnosed with psychosis, and 17.12% patients diagnosed with bipolar disorder. By including this significant number of patients diagnosed with mental disorders other than the one being targeted by our classifiers, the distributions of socio-demographic factors were much closer between groups and we accounted for other confounding factors such as medication.

The classification models developed and their results also lead us to various interesting conclusions. We understood that the expansion of the corpus affected the results obtained, making it more difficult for classifiers to identify patients diagnosed with psychosis. Additionally, we also verified that the manual correction of the transcriptions generated for the audio recordings impacted the results. This manual correction even impacted the results for models that relied exclusively on audio features.

Finally, we understood the capabilities of the feature extraction techniques that focus on structure, coherence, and content. These techniques showed potential, mainly due to the confidence that these models provide. However, models that focused exclusively on this set of techniques, did not reach the results of our baseline for most of the tasks that make up the protocol. This indicated to us that further developments in these techniques are required, particularly for European Portuguese.

As part of our work, our team had the initial objective of writing and submitting a paper for a conference. This objective was accomplished by submitting a paper, which detailed part of the work developed, to the *IberSpeech 2022*[2] conference. *IberSpeech* is a yearly conference focused on the publication of papers that focus on the study of speech and language technologies, mainly for Iberian languages, such as European Portuguese. Our paper was peer-reviewed and accepted for publication by the conference.

[2] http://iberspeech2022.ugr.es/ accessed on October 16th, 2022

1.4 Document Structure

The following document is divided into several chapters. In chapter 2, we provide some basic notions and background regarding the clinical and technical domains. In chapter 3, we overview previous studies, describing their techniques, and achievements. This last chapter is essential to understand what had already been accomplished previously in our work and what could still be expanded on.

In chapter 4, we address the approach we used and the general architecture of the solution developed throughout our work. Chapter 5 discusses more in-depth the solution developed, discussing the various steps that contributed towards its development in its various stages.

Chapter 6 provides a deep analysis of the corpus data and the data generated from feature extraction during the execution of our models. We present tables and graphs descriptive of the data that aids the results' interpretation and discussion.

Then, in chapter 7, we reveal the results achieved by our model and do a brief analysis comparing the various scores achieved and how well they compare to a baseline previously established. Finally, in Chapter 8, we provide an analysis of the results obtained mentioning the impact that the results have on the domain, as well as revealing possible limitations in our work and what could be explored as future work in this domain.

2

Domain Background

Contents

This chapter presents concepts and definitions which serve as the basis for the developed work, and are therefore necessary for a complete understanding of the study and its implications on the domain.

Section 2.1 discusses what psychosis, schizophrenia, and other mental disorders are, as well as the current diagnosis techniques in the world and Portugal. Section 2.2 mentions the feature extraction techniques and classifiers that are referred to during the discussion of the related work, in chapter 3, in the solution's architecture, in chapter 4, and in the solution's implementation, in chapter 5.

2.1 Psychosis, Schizophrenia, and Mental Disorders

2.1.1 Psychosis

According to WebMD[1] and the National Institute of Mental Health[2], *"psychosis is not a mental disorder but a symptom of multiple disorders and conditions"*. This is denominated a symptomatic group since it associates these multiple conditions and disorders with a common symptom.

From the definition 1 we know that psychosis involves the loss of one's touch with reality, which can be manifested in various ways, such as speech abnormalities and in extreme cases hallucinations, and to various degrees but all complementing the idea that the subject's cognitive processes become impaired.

There are mainly four causes for the disorder: genetics, trauma, narcotics, and injuries/illnesses. **Genetics** could be responsible for an increase in the likelihood of psychosis or the extent of symptoms. **Traumatic** events in one's life might also lead to psychosis onset, which does not need to necessarily happen shortly after the event, triggering the onset much later in one's life. Recently, substance and **narcotics** usage has been seen as one of the causes on the rise for psychosis [23]. Lastly, **injuries and illnesses**, either recognized as physical or mental, have been linked to psychosis. Although the contributing causes identified are supported by literature it is still unknown what are the exact origins of psychosis.

2.1.2 Schizophrenia

One of the most common disorders directly linked to psychosis is schizophrenia. **Schizophrenia** is defined as a mental illness that similarly to the definition of psychosis distorts the process of thinking and affects how one interprets the world around. However, instead of describing a specific episode, it impairs the subject's day-to-day life. Although most diagnoses happen after the FEP, which usually occurs during adolescence [3]. There are usually signs identified previous to this first episode, defined

[1] https://www.webmd.com/schizophrenia/guide/what-is-psychosis accessed on December 7th, 2021

[2] https://www.nimh.nih.gov/health/topics/schizophrenia/raise/what-is-psychosis accessed on December 7th, 2021

in literature as the prodromal period. Both the active stage and the prodromal phase (to a lesser extent) of schizophrenia have been linked to a wide range of signs/symptoms [13]. These signs are typically divided into two categories: positive symptoms and negative symptoms [3]. The first is associated with exaggerated behaviors (e.g. hallucinations) whereas the latter is to a dwindle of behaviors and interactions both in quantity and power (e.g. poverty of speech).

Formal Thought Disorder (FTD) is the most common symptom identified in the literature for schizophrenia and is defined as the disruption of one's cognitive processes which ultimately influence speech and its construction. FTD can be identified through a set of various speech signs such as derailment commonly described in the literature as disruption of the flow of ideas and distractability [4, 7, 13, 14, 16, 24], poverty of speech [8, 11–14, 18, 20, 24], and tangentiality, typically described in the literature as loss of coherence and absence of topic [4, 7, 10–12, 15–17, 20].

However, there is a wide range of speech cues used in literature and supported by domain knowledge, that is used for identifying and possibly predicting schizophrenia and psychosis. Some of the mentioned cues for the disorder are: (i) **neologization**, which is the creation and use of non-existent words as if they are commonly well-known [16, 25], (ii) **loss of referential standards**, using pronouns which are not easily understood to which entity they map and sometimes mentioned as referential coherence [7, 16], (iii) **speech apathy**, where the subjects display almost no emotion during the discourse [4, 10, 12, 13, 15, 20, 25], and (iv) topic revolving around **sound or hallucinations** with voices or entities such as imaginary people [13].

2.1.3 Clinical Scales and Diagnostic Tools

Most of the clinical scales used for diagnosis rely on the fact, that schizophrenia is accompanied by FTD, but due to the more difficult evaluation of one's cognitive processes, it is more practical to focus on speech and content signs associated with FTD. Some clinical scales focus on the diagnosis of schizophrenia and others on the quantification and measurement of the extent of symptoms and cues typical of diagnosed patients sometimes even discretized between positive and negative symptoms. Scale for the Assessment of Negative Symptoms (SANS)[3] [18, 26], Scale for the Assessment of Positive Symptoms (SAPS)[4] [26], and Brief Psychiatric Rating Scale (BPRS)[5] [5] are all examples of clinical scales commonly used in the literature, that rely on trained clinicians ability to identify certain characteristics in the discourse of self-reports from subjects.

Curtis et al. [26] focused on identifying the origin of such cognitive deficits and subsequently speech abnormalities, through Magnetic Resonance Imaging (MRIs), during visual, paced, covert, and constrained verbal fluency tests. From the imaging, it was perceptible that the patients displayed attenuated

[3] https://en.wikipedia.org/wiki/Scale_for_the_Assessment_of_Negative_Symptoms accessed on December 7th, 2021
[4] https://en.wikipedia.org/wiki/Scale_for_the_Assessment_of_Positive_Symptoms accessed on December 7th, 2021
[5] https://en.wikipedia.org/wiki/Brief_Psychiatric_Rating_Scale accessed on December 7th, 2021

cerebral frontal region Cerebral Blood Flow (CBF), meaning that the patients exhibit a decrease in the amount of blood flow in the frontal region. However, the most common approaches used to extract features from the discourse of patients either involve the subject speaking as **freely as possible**, sometimes even reporting on **dreams** [8, 12, 20] or **long-term memory descriptions** [12, 20], or the already mentioned **verbal fluency tests** [5, 14, 26].

The idea behind **free speech** recordings is that by not structuring subjects' speech their discourse becomes more susceptible to possible deviations by cognitive processes that influence discourse, whereas when speaking more formally or even reading, derailment, poverty of speech, and tangentiality become more or even completely imperceptible. **Dream and long-term memory reports** have been used in part because of the aforementioned reasons and because they are more likely to exacerbate content signs associated with schizophrenia and psychosis such as a predominant topic of discourse involving sound and hallucinations.

On the other hand, **verbal fluency tests** are usually used as the methodology for the extraction of surface level and time features from discourse. Verbal fluency tests encompass a wide variety of tests that rely, at their core, on the subject enumerating as many words as possible in a given time frame (usually 60 seconds). These tests can be either **visual** or **auditory** triggered [26], that is the start and end are transmitted through auditory or visual senses. These tests can also be **unpaced** or **paced** [26]. In the first case, the subjects must enumerate freely as many words as possible in the specified time interval, and in the second, the subjects must enumerate words from time to time, for example stating a word every 15 seconds for 60 seconds. Verbal fluency tests can still differ in being **covert** or **overt** [26]. With overt tests subjects are required to enumerate the words out loud. In contrast, in covert tests subjects need only to think of the word, typically, researchers do not want to know the word that was enumerated but are interested in studying the cognitive processes that occur during such task, typically evaluating cognitive fluency using auxiliary tools such as MRIs [26]. Finally, tests can either be **constrained** or **unconstrained** [26], the first specifying a category to which words must belong (e.g. words starting with the letter *p*) and the second not requiring any specific category.

Insel [3] refers to four domains in which mental illness could be improved on. The **first** domain is early prevention of the disorder since a one-cure-all approach might not be achievable. Therefore being able to better understand the early signs of the disorder is crucial. **Second**, the author hopes that, in the future, medication is able to reduce cognitive deficits, instead of only focusing on positive symptoms. **Third**, the author hopes for better care integration, since at the moment, general medical and psychiatric care are two distinct and isolated domains. By integrating care for psychosis or other mental disorders in the way diabetes is dealt with, mental disorders could potentially be prevented. **Lastly**, the author also discusses that the stigma that surrounds mental disorders only further deteriorates the state of the patient and his separation from society, arguing that making efforts to reduce stigma for such disorders

and their fatality could facilitate the patient's reintegration into society.

Despite progress, improving preventive efforts for mental disorders is still needed, mainly focusing on creating diagnostic tools that can be easily integrated into medical practice. It is important to note that trained clinicians are capable of not only identifying but also predicting mental disorders and differentiating them from one another. Therefore, supported by the related work, in chapter 3, and its results, we believed that computerized solutions may contribute also for early detection and identification of possible on-set of several mental disorders.

2.2 Natural Language Techniques and Classifiers

This section overviews various techniques used in the related work, described in chapter 3. These techniques were used for the solution developed. These techniques are well-established in Natural Language Processing (NLP) [27], and applied to various domains.

2.2.1 Part of Speech Tagging

Part of Speech (POS) maps tokens to classes that share some grammatical function [10, 11, 15]. Although these classes are not interchangeable with grammatical classes, they do share some similarities with them. The technique maps tokens, not words, to classes since sometimes groups of words behave as a single element and cannot be subdivided to achieve the correct meaning.

Take as an example the following sentence and POS tagging (assuming tokenization):

> John isn't ready for the party.
> (John, proper noun singular), (is, verb 3rd person singular present), (not, adverb), (ready, adjective), (for, preposition or subordinating conjunction), (the, determiner), (party, noun singular or mass)

From this initial stage of NLP an immediate analysis of the content and structure of the sentence can be made. Typically in a NLP pipeline, this is one of the first stages that is required for most of the processes that follow.

2.2.2 Text Lemmatization and Stemming

Lemmatization removes word inflection by identifying each word's lemma, the dictionary form of the original word. Effectively, this process is grouping words according to their lemma, reducing the overall number of words in any text or corpus. This reduction is crucial for example when computing word embeddings. Word embeddings attempt to express the meaning of each word through an N-dimensional vector. We know for a fact that words such as `run`, `runs`, and `running` share the same meaning, and should be considered a single word. To add to this problem, most of the techniques used to compute

word embeddings rely on their occurrence. If the number of words is not reduced information becomes more sparse and embeddings harder to compute (for a constant-size corpus).

An alternative to lemmatization would be **stemming**. Stemmers are typically made up of rules, many times defined through state machines, that attempt to truncate words to a stem, which ideally, should express the word's meaning. Stemmers are easier to develop, implement, and use. Still, they may not be able of grouping words correctly like lemmas do. For example, `changes` and `changing` would be mapped to the same lemma `change` but to the stems `change` and `chang` respectively.

Notably, lemmatizers achieve more accurate representations/reductions of words, however, there is an added complexity for the acquisition of these models when compared with stemmers.

2.2.3 Stop Words Removal

Another NLP technique that is important to reduce the text size and the number of words that are processed and used is called stop word removal. This technique relies on the assumption that in any language exist words that carry no to little meaning. These words are not required for the vast majority of feature extraction techniques, since they carry no meaning and are therefore irrelevant.

This technique should be used with discretion, since, in some particular cases, the removal of these stop words can have a big impact on the results obtained. For example, transformers are trained with natural speech, and therefore when fine-tuning they should be maintained as well.

2.2.4 Extraction of Word/Sentence Embeddings

This group of techniques extracts representations of words or passages in the text, for later analysis or operations. These techniques usually output an N-dimensional vector composed of real values. Due to the important role that these techniques take in the domain of NLP, we define hereafter some concrete techniques for the achievement of word/sentence embeddings.

Some techniques, such as **Sent2Vec** and **Word2Vec**, use Neural Networks, in an unsupervised approach, for the acquisition of the word/sentence embeddings that express their meaning [7]. Sent2Vec typically achieves embeddings that better relate to the semantics of the text than Word2Vec, which typically captures word relations such as synonymy The neural networks although simple can extrapolate accurate embeddings for words/sentences through their neighboring words.

Latent Semantic Analysis (LSA) is another technique for the acquisition of word embeddings. This technique consists of the extrapolation of the latent meaning (not explicit) of a word or passage. Landauer [28] compares LSA to the way a child expands its internal vocabulary, stating that only one-quarter of the vocabulary retained by a child is gathered directly through spoken sentences, whereas the rest comes from associations between words that might not have even been expressed together.

LSA proposes a similar approach, words and passages are associated through their co-occurrences. For example, through a corpus, if the following two sentences are given: (i) *He has a cat as a pet*; (ii) *He has a dog as a pet*. We can conclude directly that $cat \leftrightarrow pet$ and $dog \leftrightarrow pet$ are related, but LSA goes one step further and suggests that there is a strong relation between $cat \leftrightarrow dog$ as well.

LSA uses, as its foundation the Singular Value Decomposition (SVD) algorithm. SVD states that any matrix can be decomposed into the product of three different matrices, with one allowing for dimensionality reduction to maintain only meaningful information. Choosing the correct dimensionality can be difficult. It is noteworthy that LSA computes embeddings for words that are independent of context and therefore are unable to capture correctly the various meanings that can be given to a word.

Latent Content Analysis (LCA) obtains the meaning of every word by analyzing its co-occurrences with every other word and expresses this meaning through a vector, similarly to LSA. However, it goes one step further, analyzing which and how many words are required to achieve the same sentence meaning and compares this sentence's meaning to the meaning of a set of probe words [13].

Finally, **Transformers** are generative deep-learning models that codify each item in a sequential input into an N-dimensional vector [29]. Their architecture allows for parallelization which speedups the embeddings' acquisition. Furthermore, they have a self-attention mechanism that mimics the human capability of retaining attention to some part of the input while evaluating another. Some of the most popular transformers are the Bidirectional Encoder Representations from Transformers (BERT), Robustly Optimized BERT Pretraining Approach (RoBERTa), and Generative Pre-trained Transformer (GPT).

For example, *XLM-RoBERTa* is a large multi-lingual transformer based on the original version of *RoBERTa* developed by *Facebook* and further trained with $2.5TB$ of data. This model has been trained in over 100 languages, one of which is European Portuguese. This model does not require the specification of the language to be used, instead, it grasps this information from the input that is fed to it. There are two available versions of *XLM-RoBERTa* available at *HuggingFace*, a `base` and `large` versions.

2.2.5 Vector Unpacking

Vector unpacking is characterized by the decomposition of a sentence's meaning into its various meaningful components, and word embeddings, and tries to map these various components into the words that originally constituted the sentence [13]. We can then ascertain if every word within the sentence can be reconstructed from the various meaningful components.

2.2.6 Clustering

Clustering involves the creation of sets of data points that are similar to one another. These sets are typically achieved through distance or similarity metrics and N-dimensional vectors, that represent the

various data points. In literature, specifically in the domain of this project, clusters are created based on similar or related content, which can be achieved through *probe words* [10, 13].

2.2.7 Semantic Role Labeling

Semantic Role Labeling (SRL) maps segments of the discourse to one or more semantic roles, which express the role that the segment takes in the sentence [10]. This methodology relies on a classifier already fitted to other data.

2.2.8 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a generative technique that maps segments of the discourse to one or more topics [10]. This methodology computes topics simply by the distribution and interrelation between words in the given data.

2.2.9 Level of Committed Belief

Level of Committed Belief (LCB) techniques measure the belief level of the subject while stating the propositions [10]. This level of belief is distinguished into four different types, all expressing different belief levels or types: (i) **Committed Belief (CB)** in which the subject believes the stated proposition; (ii) **Non-committed Belief (NCB)** when the subject might even believe the proposition, but does not believe it firmly; (iii) **Reported Belief (ROB)**, the belief stated by the subject is not his, believing it or not; (iv) **Non-Attributable Belief (NA)** when the subject is not stating a belief.

2.2.10 Sentiment Analysis Models

Sentiment Analysis Models measure the sentiment and the sentiment intensity of sentences. These techniques require corpora previously annotated with sentences' sentiment to correctly predict other sentences' sentiment [10, 30, 31]. Typically sentiments are analyzed in terms of arousal, which expresses the intensity of the emotion, and valence, which expresses the signal of sentiment and goes from negative to positive.

2.2.10.A Web Scraping for Sentiment Analysis

A big obstacle when attempting to find sentiment analysis models is that they may be nonexistent or severely limited in certain languages. Several tools exist for sentiment analysis for the English language, for example, however, this is not the case for the Portuguese language and especially for the European Portuguese language, for which models are severely limited. The models that exist for sentiment analysis

on European Portuguese typically work through dictionary-like objects, which map words, sequences of words, and/or lemmas to particular emotions or valence scores [32].

To aggravate the situation, in these particular languages, corpora annotated with sentiment scores are also limited. A solution that has been employed in other domains is the usage of **web-scraped information**. Web scraping relies on the extraction of public and readily available information from web pages to establish a corpus. For example, using reviews and their respectively annotated scores it is possible to develop a corpus for sentiment analysis that can then be used to train models capable of extracting sentiment/valence scores from the text [33].

2.2.11 Word Graph Analysis

Word Graph Analysis techniques allow for the assessment of topological structures of discourse [8, 12, 17, 20]. A connected and directed graph is created from transcribed speech samples where each node represents a given word, in the transcript, and each link represents temporal connections between words/nodes.

From the word graph, topological structures can then be evaluated and compared to others to identify possible deviations and reduced discourse coherence. The main advantage of this technique against the more usual and explored, LSA, is that it does not require a large corpus and is quite efficient.

Typically the most usual metrics in the literature that can be extracted from such structures are: (1) the number of nodes and edges, (2) the number of nodes in the various Strongly Connected Components (SCCs), (3) the probability of the various SCCs occurring, which is calculated by randomly shuffling the words, (4) the number of nodes in the Largest Strongly Connected Component (LSCC), (5) the probability of the LSCC occurring, which is also calculated by randomly shuffling the words, (6) the size and number of cycles.

2.2.12 Classifiers and Other Techniques

The focus of our work is not to contribute to the state-of-the-art classifiers or study their corresponding advantages and disadvantages, but rather, to use them to obtain predictions that can later be evaluated. Although such an endeavor would by itself constitute an important study, we are interested only in features to distinguish when to use each one of them. We explored classifiers in literature, the traditional classifiers, such as Decision Trees (DT) [34], Support Vector Machines (SVM) [35], and Naive Bayes (NB) [36], among others, and the more complex and recent ones, such as Random Forest (RF) [37], Multi-Layer Perceptrons (MLP) [38], Multi-Layer Feed-Forward Neural Networks (FFNN) [39], and Recursive Neural Networks (RNN) [40].

3

Related Work

Contents

This chapter presents insights into current and previous work to identify psychosis, schizophrenia, or otherwise improve efforts in the early prediction of mental disorders.

In section 3.1 we discuss research that relied on time or other surface-level features of speech. Section 3.2, addresses work that either analyzed overall speech structure or its coherence. In section 3.3, we focus on research that used speech semantics and pragmatics, which are critical for our work, since it is our focus. Lastly, section 3.4, provides a quick overview to easily identify the current state of the domain that is promising and more likely to produce better results.

3.1 Speech Fluency, Time, and Acoustic Analysis

Clemmer published one of the first studies [4] to defend the theory that mental disorders should be concretely defined through objective measures. Clemmer tried to correctly distinguish schizophrenic patients from controls, exclusively through speech surface attributes and time metrics. The author argued that such speech abnormalities might have as origin cognitive deficits, but since they are harder to evaluate, especially at that time, the author focused on measurable and objective speech metrics. Clemmer reported an 80% accuracy using as features the number of pauses, their duration, their type (silent or filled), their positioning (between or within constituent units), and the speech rate of the speaker.

Similar to the study of Clemmer, Alpert et al. [18] focused on the number of pauses, their type, and positioning. However, in this case, they studied possible correlations between such features and clinical scales used for diagnosis by clinicians. The authors' justification for the use of these features relied mainly on two negative symptoms of schizophrenia: alogia and the flat effect of speech. The study concluded that there is a correlation between clinical scales and such features and that clinicians while diagnosing, seem to have an intrinsic and instinctive perception of such features during patient interviews.

Kuperberg et al. [14] presented an attempt to compare, through objective metrics, different groups of patients diagnosed with mental disorders. The authors tested out the hypothesis that thought-disordered patients are less sensitive to linguistic aberrations. The hypothesis was confirmed. The reaction time for the detection of a given and specified word was recorded, for normal and altered versions of sentences. Although in general thought disordered patients have higher reaction times than non-thought-disordered patients, for both normal and deviated sentences, thought-disordered patients are the least affected by the aberrated sentences, displaying almost no difference in reaction times to their times in normal sentences.

Gosztolya et al. [19] published a study with a very similar methodology. The authors identified pauses, their types, and duration, and measured speech metrics such as the articulation rate and speech tempo. The difference of this study is that the objective was not the differentiation of psychotic disordered

patients from controls but instead the differentiation of psychotic patients from patients diagnosed with bipolar disorder. Although the sample was rather small (26 patients diagnosed with schizophrenia and 14 patients diagnosed with bipolar disorder) and might not offer confidence in generalized conclusions, the results supported that such separation of mental disorders is possible, supporting our work.

Lastly, the work by Forjó et al. [5] aims at establishing the first European Portuguese protocol and corpora for psychosis identification from audio recordings from diagnosed patients and healthy controls. The authors defined healthy subjects as subjects who had never experienced the FEP. The protocol consists of **seven tasks**:

1. Phonetic verbal fluency task in which the subjects enumerate words starting with 'p' for 60 seconds.
2. Categorical verbal fluency task in which the subjects enumerate animals for 60 seconds.
3. Reading of a well-known children's story.
4. Retelling of a well-known children's story.
5. Description of a positive affective image.
6. Description of a neutral affective image.
7. Description of a negative affective image.

The authors used the Geneva Minimalistic Acoustic Parameter Set (GeMAPS)[1] [21], a software for the extraction of audio features that can then be used by classifiers, and a transcriber for the acquisition of the speech and articulation rate. The authors achieved promising results, classifying with an accuracy of 87.5%. However, the authors did not experiment with a diverse control set and had a relatively big disparity in terms of education between both groups. Patients diagnosed with psychosis and the control group ideally should have a similar distribution to account for social-demographic factors. Even though the study had its limitations, it is important to mention that the authors' study was constrained by the COVID-19 pandemic and that the corpus acquisition was severely more arduous for this reason.

Notably, one aspect in common in all of the previously mentioned studies [4, 5, 14, 18, 19] is that their definition of psychosis or schizophrenia relies on cognitive deficits. All authors decided to focus on surface-level features of speech or on other time metrics for their simplicity, easy implementation, and quick results (possibly even live classification) [19], or because an analysis of thought processes, or lack thereof, is arduous, time-consuming, or even impossible [4, 5, 14, 18].

[1] https://audeering.github.io/opensmile-python/ accessed on December 12th, 2021

3.2 Coherence and Structural Analysis

Elvevåg et al. [6] were the first to apply coherence measures used in NLP, in this domain. The authors' focus was on the correct identification of diagnosed schizophrenic patients from controls, while, at the same time, researching whether family members of controls and diagnosed patients were identifiable as well. For classification purposes, the only feature used was the coherence score computed through LSA. The results obtained provide a proof-of-concept and basis for future studies and the use of NLP techniques for domains such as mental disorder identification.

Bedi et al. reported a longitudinal study [11], in which similar to the previous study, the authors applied coherence measures, defined in NLP literature, into the domain of clinical diagnosis of mental disorders. However, differently from the previous study, Bedi et al. [11] did not aim at distinguishing the diagnosed patients from controls but instead sub-categorizing patients. In this case, classifying patients as clinical high risk (CHR+) or non-clinical high risk (CHR-), effectively predicting future psychosis in patients. Besides using coherence metrics through LSA, the authors also used the number of determiners (normalized by phrase length) and the maximum phrase length as features to build a convex hull classifier that segregates patients into their respective groups. The study achieved 100% accuracy in predicting future psychosis on-set. Still, the sample size was rather small which impairs the generalization of the conclusions achieved. There were only 34 patients, five of whom transitioned into full-blown psychosis. Still, the authors tried to evaluate the likelihood of the results obtained occurring due to chance and achieved a probability of 0.05.

The study of Corcoran et al. [15] expanded on the study of Bedi et al. [11] by re-evaluating the findings obtained in the previous study with a larger sample size to obtain more meaningful results. The study followed the same approach: speech samples were elicited, by retelling and freely answering questions regarding a known story, Then, these samples were processed using LSA to achieve a model defined by a convex hull that accurately classifies patients for CHR+ or CHR-. However, the authors used the minimum coherence, maximum coherence, and coherence variability of the subject, which achieved an accuracy of 72% which seems to be more plausible than the 100% mentioned in the previous study. A possible critique of LSA studies, is that due to the very nature of the technique, it does not provide an easily explainable justification for the score achieved which makes it more difficult to use in future applications of the technique in clinical diagnosis.

Some of the most referenced speech characteristics traced to schizophrenia in the literature are the disruption of syntax rules and the disruption of the flow of ideas. A few studies [7, 16] even pointed, as an extension of the symptoms referenced, the incapability of correct reference-making from diagnosed patients. The study of Iter et al. [7] improved on previous implementations of LSA in the domain. The authors explored referential coherence measure where they extract the number of ambiguous pronouns, from interviews with patients and controls. A reference is ambiguous when it precedes the referenced,

denominated as a *cataphora*, or when it is not associated with any entity. The study by Iter et al. [7] had a fundamental problem, its sample size, the dataset used for modeling and classification consisted only of 19 subjects. Just et al. [16] expanded on this study by following the same methodology but with a sample size of 30, 20 diagnosed patients (10 with FTD and 10 without FTD), and 10 controls. Although the sample size was still small, it added some generalization power to the discoveries.

The application of NLP techniques to medical diagnosis domains has proved beneficial, however, they must be extensively and carefully discussed to not amplify already existing biases or create new ones. Iter et al. [7] called for awareness when applying NLP techniques to sensitive domains such as the identification of mental disorders. LSA does not take into account, for example, repetitions. Taking this problem to the extreme, one could repeat constantly a single sentence or even a word and it would obtain a high coherence score. This problem is especially important since word repetition is one of the speech cues associated with psychosis. The authors overcame this problem by first pre-processing the transcribed interviews to deal with discourse characteristics and removing stop words. Then, the authors' used state-of-the-art techniques for the computation of sentence meaning, by applying a *Sent2Vec* model to their study. Another problem with the technique is that it is biased towards longer sentences since it bags more words together, which in turn increases the likelihood of two words being semantically related and therefore increases (artificially) the sentences' coherence.

The study of Hitczenko et al. [41] showed that the coherence techniques, in this case, LSA, failed to identify psychosis. The authors' verified that LSA had identified and correlated with social factors from the subjects. The studies mentioned before also considered socio-demographic factors such as education level in their studies, for example by matching control and patients diagnosed. Therefore, we concluded that although it is possible that models adjust to social factors this can be prevented.

Mota et al. published three different studies [8, 12, 20] in which they employed word graph analysis, employed in other NLP domains [27], to our domain. The authors carried out semi-structured interviews of dream reports [8,12,20], memory, reports preceding the dream [12,20], reports of the earliest memory [12], and descriptions of affective images [12]. From the transcriptions of these interviews with controls, schizophrenic patients, and bipolar patients, the authors attempted to correctly identify each one of the mentioned groups. The authors used features extracted from the developed speech graphs. The metrics used allowed for the differentiation of the several groups and the word graphs, when visualized, were noticeably different. Similarly to the studies by Mota et al. [8, 12, 20], the study of Spencer et al. [17] employed the same metrics using speech graphs although now distinguishing between controls, patients with already the first episode of psychosis, and patients identified as CHR+. The authors also used a scale for formal thought disorder evaluation called Thought and Language Index (TLI) and used it as the baseline for the future comparison of the results obtained. Both Mota et al. [8, 12, 20] and Spencer et al. [17] concluded that speech graphs' measures correctly distinguish between groups, and

the latter, even concluded that these measures correlate with TLI scores. These studies effectively predicted psychosis by identifying patients labeled as CHR+

From the studies presented in this subsection, we identify that the most common approaches used were LSA and speech graph measures. LSA is better at directly measuring coherence through sentence embeddings, but, it requires a corpus, other than the one being evaluated, or previously computed word embeddings. On the other hand, speech graphs need no additional corpus other than the one being classified. Due to its high efficiency and simplicity, this technique could even be employed live during interviews to extract features that correlate with coherence. However, they do not measure directly coherence, basing themselves on the overall structure of speech, and not considering the word/sentence's meaning.

3.3 Speech semantics and pragmatics

The studies previously mentioned, focus either on surface-level features of speech or on its structure, since these closely relate to the features also used by clinicians in their diagnosis, but due to the very nature of mental disorders, such as psychosis, other studies have hypothesized that such cognitive deficits have an impact on speech semantics.

The leading study of Rezaii et al. [13] explored the application of semantic analysis for the prediction of psychosis. The authors stated that evidence of feature psychosis can be detected already in the prodromal phase and that early prediction can stop or at least, slow down the progression of psychosis. The techniques used by the authors have as supporting psychosis characteristics: the poverty of speech, which in turn means a low semantic density in speech and auditory hallucinations. Then the authors propose a technique to measure each one of the mentioned symptoms: LCA with vector unpacking and probe word clustering, for the poverty of speech and auditory hallucinations respectively. The authors achieved a 93% accuracy when predicting psychosis with the combination of both techniques, supporting that future psychotic patients have lower semantic density and have a particular focus on sound-related content. However, the sample included only 40 subjects with 7 converting to psychosis in less than 2 years and a half, and due to this reduced sample size, especially in converters, the models might not generalize for future studies and must therefore continue being explored.

McManus et al. [9] focused on the content of discourse to correctly identify schizophrenic patients from controls, instead using as corpus, micro-blogging posts, in this case, *Twitter* posts. The authors used as features the time of the day of the posts, the time between posts, the number of friends, emoticons used, and, more importantly, the number and words used that are related to schizophrenia. The authors adopted as criteria for the selection of schizophrenic users: if the user self-identified as schizophrenic in its description or on any of its status/posts, or followed *@schizotribe*, according to the

authors, a well-known community of diagnosed schizophrenics. The application of domain knowledge to the identification of psychosis through such an informal discourse is interesting. However, there are some problems with the overall study. The authors used the number of words related to schizophrenia and whether the subjects followed a community as justification for the labeling. These attributes are biased toward self-diagnosis or curiosity for the theme and do not necessarily mean that the subject is diagnosed with psychosis.

The study of Kayi et al. [10] was essential to understand the current state of content analysis in this domain, since it described an extensive in-depth study of discourse semantics and pragmatics, with 93 patients diagnosed with psychosis and 95 controls. The authors aimed to correctly classify any given subject through topic, confidence, and sentiment analysis. To achieve this, the authors solicited two different one-paragraph long essays from each subject answering the questions *"What is it like your Sunday?"* and *"What makes you angry?"*, intentionally not evoking and evoking, respectively, sentiment from the subject. The study involved various techniques, namely POS tagging, dependency parsing, SRL, LDA, GLoVE clustering, and LCB, as mentioned in section 2.2. The authors concluded that the most discriminating features are the syntactic features followed by syntactic in conjunction with semantic features, respectively an F-score of 78.92% and 70.29%. There are two main critiques of the article. Firstly, the authors did not provide an in-depth description and discussion of the techniques and metrics used. The authors provided only a theoretical description of each, possibly due to the vast amount of techniques used. Since the authors did not provide this description, replication by future studies becomes arduous. Secondly, the authors limited the study to written essays, which is not the usual approach in the domain and might explain the better performance with syntactic features. Clinical diagnosis relies on spoken interviews in part of their spontaneity. By neglecting such an intrinsic part of speech the results might have been impaired. The authors also described a parallel study, in the same article, in which they analyzed *Twitter* posts from self-identified diagnosed schizophrenic patients but fell short on the same pitfalls as the ones described in the last study [9].

3.4 Summary and Discussion

From the table 3.1, in which we summarized the various studies reported, we verified that most of the work to identify, predict, or otherwise involving psychosis or schizophrenia has focused on discourse surface level and time features and coherence and structural features with semantic and pragmatic analysis being a, relatively speaking, unexplored domain.

Table 3.1 only displays the overall sample size of the study not considering the number of groups involved, but in general, we can conclude that studies have, relatively speaking, small sample sizes with just a few exceptions [9, 10, 15]. [10] used a pre-developed corpus (LabWriting) and [9] used micro-

Table 3.1: Overview of the mentioned studies grouped by category, regarding the analysis approach, diverse controls, subject's language, sample size (in the following order: control, psychosis, other), and whether it required the use of subject's private information.

Study	Analysis Approach	Diverse Sample	Subjects's Language	Sample Size (C/P/O)	No Private Information
	Fluency, Time and Sound Analysis				
[4]	time / pause analysis	✗	English	20/20/NA	✓
[18]	pause analysis	depressive	English	20/19/17	?
[14]	reaction time / verbal fluency	non-FTD	English	10/27/NA	✓
[19]	time / pause analysis	bipolar	Hungarian	A/26/14	✗
[5]	time / GeMAPS	✗	European Portuguese (EP)	56/36/NA	✓
	Coherence, and Structure Analysis				
[6]	LSA	✗	English	30/53/NA	✗
[11]	POS Tagging / LSA	✗	English	29/05/NA	✗
[15]	POS Tagging / LSA	FEP/CHR+	English	90/40/NA	✗
[7]	LSA / referential analysis	✗	English	05/09/NA	✗
[16]	LSA / referential analysis	non-FTD	German	10/20/NA	✗
[8]	word graphs	bipolar	Brazilian Portuguese (BP)	08/08/08	✗
[20]	word graphs	bipolar	BP	20/20/20	✗
[12]	word graphs	✗	BP	21/21/NA	✗
[17]	word graphs	FEP/CHR+	English	37/16/NA	✓
	Semantics and Pragmatic Analysis				
[13]	LCA / Clustering	✗	English	28/12/NA	✗
[9]	count schizophrenia related words	✗	English	200/96/NA	✗
[10]	SRL / LDA / Clustering / LCB	✗	English	95/93/NA	✗

✗- negative, not done ; ✓- affirmative, done ; ? - unknown ; NA - not applicable

blogging posts from social networks (*Twitter*).

Table 3.1 also shows that, particularly for content-focused studies, there is little variability in the sample of subjects. The world is not neatly divided into isolated categories, and therefore when, for example, identifying psychosis we should consider that subjects with other pathologies might be evaluated with the developed models. In the future, studies must evaluate the achieved models with a more diverse sample of subjects, possibly subjects diagnosed with pathologies other than the one being targeted.

We also verify that only one study [5] targeted European Portuguese subjects, focusing on surface-level features, and another three [8, 12, 20], all by the same authors, focused on structural analysis of Brazilian Portuguese. Studies that target this population provide more support to the conclusions of the mentioned studies from previous studies and could potentially advance the current state of mental illnesses paradigm for diagnosis in Portugal and internationally.

Lastly, it is noticeable that most of the work developed required the test subjects to disclose private information to the authors or other elements part of the study. In the particular case of studies, when

consent is given, it is not problematic that private information is divulged and used to support the diagnosis. However, if such methodologies are to integrate diagnosis or early detection efforts, they must not require private information to maintain confidentiality, one of the core duties of medical practice[2]. Although the explicit requirement of private information is dispensable, it is noteworthy that a subject is identifiable through discourse and speech characteristics, and therefore, in any case, data must be treated carefully.

We conclude by stating that there is a gap in the literature, which we fill, for studies that focus on content analysis with a diverse and significant size sample in European Portuguese that does not require the disclosure of private information.

[2] https://depts.washington.edu/bhdept/ethics-medicine/bioethics-topics/detail/58 accessed on January 11th, 2022

4

Solution's Design

Contents

This chapter describes the developed solution. Section 4.1 defines the objectives of our solution. Section 4.2, defines the requirements of the expected solution, and section 4.3, specifies in detail the architecture for the solution to be developed. Finally, section 4.4, specifies the evaluation metrics used to properly evaluate the obtained results and consequently, our solution.

4.1 Solution Objectives

Our work originated partially from the future work proposed by Forjó et al. [5]. Right at the start of the development of our work, we profited from talking with the authors to understand the limitations of their work and its implications as well as possible avenues for expansion. Due to time constraints, [5] did not explore the structure and semantics of speech in their solution. This is one of the core techniques, commonly used by previous studies, that could uncover more meaningful features from recordings and therefore allow for the development of a better solution. Another limitation of the referred research work is the dimension and diversification of the recorded samples, as discussed previously, since it only considered test subjects that are either deemed healthy or diagnosed with psychosis.

In addition, the beginning of our project aimed at understanding from the perspective of the domain experts, psychiatrists, what should be explored according to their experience, in particular, what could be untapped in terms of relevant information helpful in diagnosis. As a result of these shared experiences, we decided to extend the already-developed work for psychosis by using a control group constituted of healthy subjects and patients diagnosed with bipolar disorder. Bipolar disorder has several stages, each with its symptoms and its speech cues. This high variability in terms of subjects' discourse would introduce more variability in terms of the control population which means more meaningful results. Furthermore, since the patients diagnosed with psychosis and bipolar disorder are both under medication we would be accounting for possible side effects from the medication and preventing models from adjusting and fitting to these confounding factors.

4.2 Solution's Requirements

The main requirement to achieve the solution proposed for the stated problem is the acquisition of a large and well-representative corpus, that should reflect accurately the domain. Since our solution is aimed at aiding the process of diagnosis currently undertaken solely by clinicians, it should be able to differentiate psychosis from other disorders that might share some similarities in terms of symptomatology and speech cues.

The establishment of a protocol is critical to the acquisition of the recordings, but since we aimed at extending the existing First European Portuguese Corpus for the Identification of Psychosis, we followed

the already developed protocol, mentioned in section 3 and displayed in appendix A.

The protocol established for the work of Forjó et al. [5] had already been accepted by the Ethics Committees of three different institutions: Instituto Superior Técnico (IST), Centro Hospitalar de Lisboa Ocidental (CHLO), and Casa da Saúde S. João de Deus (CSSJD). Therefore, an addendum to the ethics committees was submitted and accepted. It allowed us to continue with the study and the recordings, and to extend the target population to include patients with bipolar disorder. It is important to note that protocol does not require any private information from the participants besides general socio-demographic data, relying on general tasks, that in the future could be carried out by clinicians, with low resource requirements, and without compromising confidentiality or the trust between patient and medical practitioner.

4.3 Solution's Architecture

The overall solution is composed of numerous tasks and stages that have to be carried out correctly to allow the acquisition of correct and meaningful results that are interpreted and discussed, in chapter 7. Due to the inherent complexity of the developed solution, in this section, we describe the different parts of the developed solution being agnostic of their implementation. This allows the reader to be familiarized with the solution before implementation details are discussed in chapter 5.

Subsection 4.3.1 provides an overview of the needed steps for the correct and safe processing of the subject's recordings and data. Subsection 4.3.2 describes the central and fundamental part of the solution, where models capable of classifying subjects are developed, in an implementation-agnostic way, referring only the various feature sets extracted from recordings and transcriptions. Finally, subsection 4.3.3 describes an overview of the developed solution that includes all the parts previously described, detailing how they are interconnected with each other.

4.3.1 Recordings and Data Processing

Data and its treatment was the first and likely the most crucial step in the development of accurate and meaningful models. Through figure 4.1, we can observe an overview of this stage, which is composed of numerous sub-stages that when carried out in order, allow for the creation/expansion of the First European Portuguese Corpus for Psychosis Identification.

When a new set of subjects has completed the protocol we carry out two tasks almost immediately. First, the subject's id and socio-demographic information was saved accordingly in a subset of the First European Portuguese Corpus for Psychosis Identification with information on subjects with an unknown diagnosis. It is important to note that at this stage, in the case of patients diagnosed with either bipolar

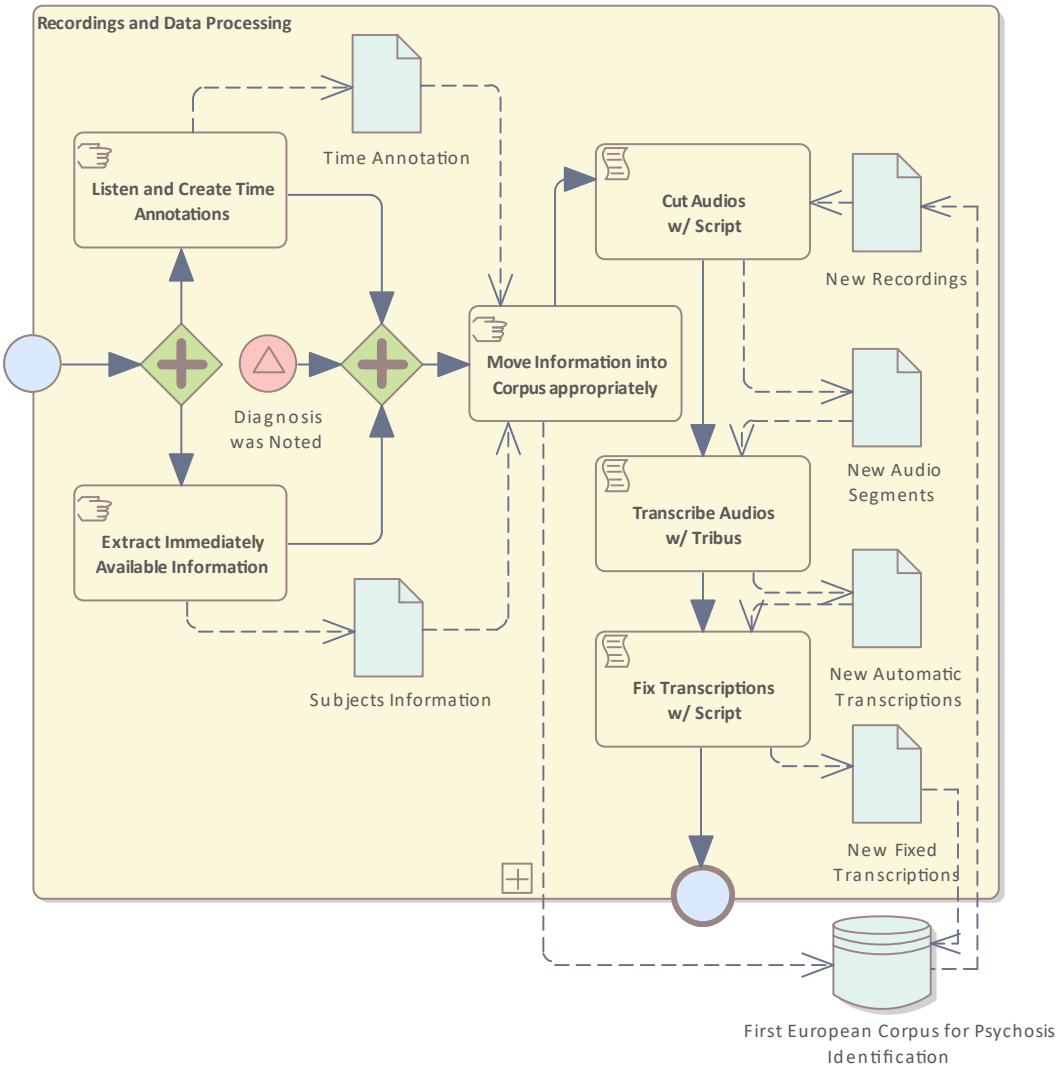


Figure 4.1: Diagram with a simplified representation of the various stages involved in processing the subject's recordings and data.

disorder or psychosis, the subjects' diagnosis is still unknown. This is why the information from these subjects is maintained separately from the rest of the corpus.

Then, we played each one of the recordings and manually identify the timestamps that mark the start and the end of the subject's response to each task, marking out possible interventions of the interviewer. This was required since we wanted to test how good each task was to identify psychosis and at the same time, allow the models to adapt to the content of each task. Furthermore, this process was carried out manually since full automation of such a task would be difficult. These timestamps were then also saved in the corpus but maintained apart from the rest of the data from the subjects who had their group already identified.

Eventually, the diagnosis of the subjects was disclosed to us and we could merge this information

with the previously extracted information and timestamps. At this point, these new subjects were formally added to the First European Portuguese Corpus for Psychosis Identification, considering exactly where they belong according to their diagnoses.

Afterward, we used the timestamps manually annotated to each one of the subjects' recordings and automatically cut the recordings into audio segments, ready to be transcribed and used by our models.

Following the segmentation of the recordings, we transcribed each audio sample. Most of the transcribers available were either free to be used but did not include European Portuguese, or included European Portuguese but were paid and/or required the data to be submitted to online tools. The first set was inadequate for the problem at hand. The second required financial support and was not advised since we were dealing with confidential and sensitive information especially when it came to the patient's information.

Therefore, we followed a slightly more arduous but more accurate methodology, using an already developed transcriber for European Portuguese, developed by Carvalho and Abad [42], *TRIBUS*. *TRIBUS* outputs accurate transcriptions, especially for recordings with good quality in which the subject speaks freely or slowly. However, *TRIBUS* sometimes cannot capture a specific word or words that are enumerated, not interconnected, or spoken too rapidly. This limitation might have had a large impact on content analysis since this analysis relies on the specific words being spoken. For the mentioned reasons, we manually analyzed the generated transcriptions while listening to the recording and made small corrections to miss-interpreted words from the transcriber.

At this point, the subjects' information and audio segments had been integrated into the already-developed First European Portuguese Corpus for Psychosis Identification, meaning that the stage responsible for the processing of the recordings and data stage has been completed.

4.3.2 Classification Models and Stages

Without classification models capable of identifying patients diagnosed with psychosis, from a corpus, through recordings and/or transcriptions, no contribution would have been possible in terms of computerized solutions in the mental health prevention domain.

It is important to stress that our work is an expansion of the work of Forjó et al. [5]. Consequently, we decided to use the model developed by Forjó et al. [5] as the baseline for all of our developed models. For example, by comparing the results obtained by Forjó et al. [5] sound and speech model with the original version of the First European Portuguese Corpus for Psychosis Identification against our expanded version we can conclude whether the model had identified psychosis or fitted to other characteristics of the population.

Notably, the model developed by Forjó et al. [5] served merely as a baseline, and therefore no in-depth analysis or explanation of this model was made. A more in-depth description of the various

classification models developed can be seen in subsection 4.3.2.A.

Besides the development of different classification models that relied on different features and extraction techniques, we also developed versions of these models for different stages, described in subsection 4.3.2.B. These stages eased our analysis and allowed us to have more confidence in the results obtained by being sure that we achieved a local maximum.

4.3.2.A Classification Models

Throughout our work, we developed two different models, that relied on different types of features, and consequently on distinct techniques for the extraction of features from subjects' recordings and/or transcriptions.

As mentioned before, the first set of models developed by us was almost identical to the one developed by Forjó et al. [5]. These models focus on speech and sound features, extracted and used by the developed models are displayed in greater detail in table D.1 and table D.2, respectively.

Note that the sound features presented in table D.2 do not have a description associated. For these features to be properly described further background knowledge would be required, and this is outside of the scope of our work. For us, it is only relevant to point out that these features are low-level audio descriptors that summarize vectors, formed by various points from audio recordings gathered at a fixed time rate, through statistical functions such as mean, standard deviation, or percentiles. For further comments and definitions of the individual sound features used, we refer the reader to the work of Forjó et al. [5].

The second set of models developed by us relied on structure, coherence, and content features extracted exclusively from the subjects' transcriptions. Note that the models that used these types of features were the focus of our work.

In these models, we extracted several features using several techniques previously mentioned in chapter 3. The techniques employed for feature extraction in our work were: *LSA* (section 2.2.4), *Word Graphs* (section 2.2.11), *Vector Unpacking* (section 2.2.5), *LCA* (section 2.2.4), and *Semantic Analysis* (section 2.2.10). Note that each one of these techniques may extract more than one feature from transcriptions. The specific structure/coherence and content features extracted can be seen in table D.4 and D.3, respectively.

The results obtained allowed us to evaluate whether psychosis can be identified exclusively through audio transcriptions and features that rely on the content and structure of what is being spoken.

4.3.2.B Classification Stages

Besides developing the various models that rely on different sets of features, our team also created several versions of these models, that behave differently from each other, calling these variations clas-

sification stages. These different model behaviors allow, at one point to explore extensively what might be the most appropriate classifier for each task, and at another moment to explore the full capabilities of the model and test the confidence in our models.

Our team developed two classification stages for our models:

- *Exploration Classification Task*: At this point, our focus was on exploring as many model variations as possible. This stage varied the feature set used, for example, testing with structure and content feature sets by themselves and together. Besides varying the feature set used, we also tested with various classification techniques and their various hyperparameters.
- *Assessment Classification Task*: After the initial stage has been carried out, we selected for each task the best performing model. Then, for each one of these task models, we incremented parameters such as the number of training epochs for neural models, to try to improve our results. Additionally, we ran each model multiple times to assess a model's confidence level and feature importance.

4.3.3 Overview

An overview of the entire solution, the models developed, and how the different stages are connected is shown in figure 4.2.

We started by recording subjects, healthy controls, patients diagnosed with psychosis, and patients diagnosed with bipolar disorder. These recordings and the data associated were then processed. This stage, described in detail in section 4.3.1 is repeated until the process of recording subjects is finished and there are no more recordings to process.

Once the data and recording processing stage was finished, we could develop our models. To this purpose, we started by developing our first and second models, the first relied on speech and sound features and the second relied on structure, coherence, and content features. Both followed the exploration classification stage in which we were interested only in understanding which were the promising classification techniques and their hyper-parameters.

Finally, from the results gathered in the exploration classification task, we developed more models that used structure, coherence, and content features. These models allowed for the assessment of the models' confidence and the relative importance of the various features.

4.4 Solution's Evaluation Criteria

Our solution relies on supervised learning since we know a priori the classification of each one of the recorded subjects and therefore, we could evaluate any given model with this information.

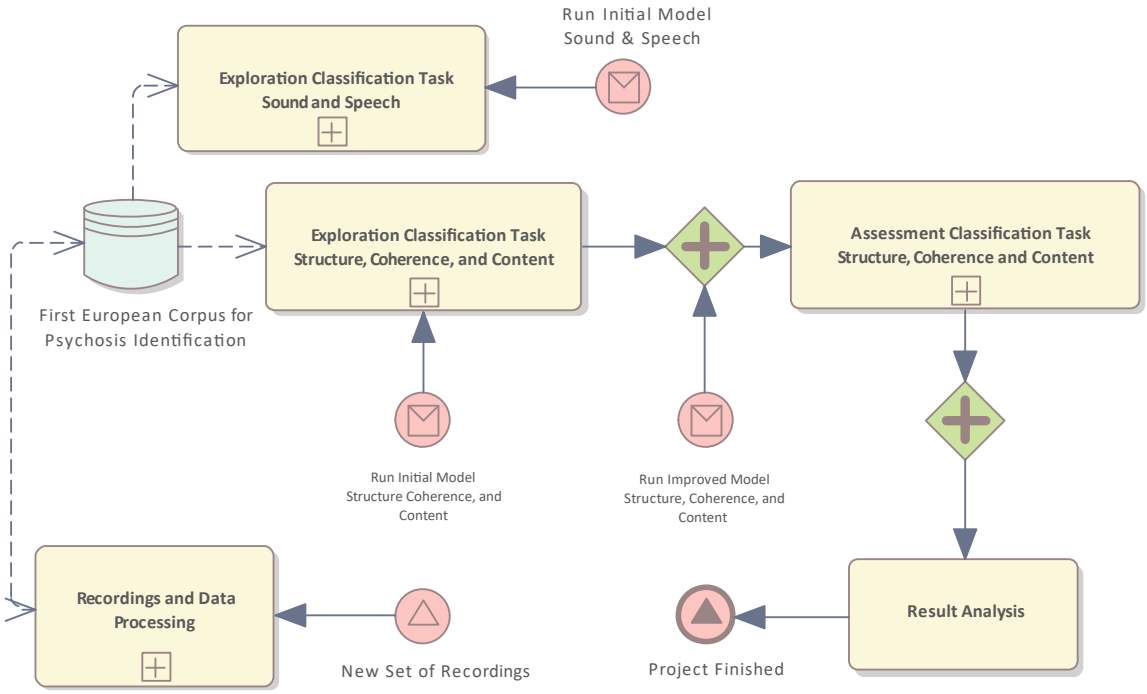


Figure 4.2: Diagram with an overview of the entire work developed, agnostic of implementation details, and the complexity of the stages mentioned before.

For this work, we heavily depended on the cooperation of the patients and on a small number of institutions and people that performed the recordings. For this reason, we did not expect a huge number of patients to be recorded. Therefore, we decided, right from the start, to use *leave-one-out cross-validation* to train and test the models developed.

The aforementioned labels were only used during the evaluation of the obtained results, and during the process of adjustment of the classifiers' parameters to improve the developed model or create it. We were mindful of data leaks, not using the labels for the process of classification, and not using the test set while training.

Several metrics could have been used to evaluate the results obtained by our supervised models: **accuracy**, **precision**, **recall**, or **F-measure**. Each has its advantages, for example, **precision** and **recall** are more useful in unbalanced corpora or when it is more important not to have false positives or false negatives, respectively. In our work, both classes were similar in size and we aimed to support clinicians' diagnoses. To this purpose, we concluded that it would be more relevant to prevent false negatives than false positives. Consequently, we believed that the best metrics to use were **accuracy**, **recall**, and the **F-measure**. The overall **accuracy** and **F-measure** provided us with a general estimation of the models' competence at correctly identifying psychosis. The **recall** also provided a general estimation of the models' competence, however, considering that false negatives should be avoided to not

exclude possible converters from the early detection stage. Additionally, we also used the **Unweighted Average Recall (UAR)**, a measure common in paralinguistic tasks, which computes the average of the **recall** of the positive class (**specificity**) and recall of the negative class (**sensitivity**). This was the main metric used throughout our work to evaluate models' performance.

In short, we evaluated the influence that (1) extending the corpora and (2) focusing on speech structure and semantics had on the results obtained. Lastly, we discussed possible justifications for the results obtained and the difference between test scores.

5

Solution's Implementation

Contents

This chapter provides further details about the solution development. In section 5.1, we describe in more detail the acquisition of subjects and their recordings. Section 5.2 describes the environments used and developed in order to carry out the mentioned studies. Section 5.3 details the processing of data and recordings, extending on what was discussed in section 4.3.1 with implementation details and discussion on the choices made. Section 5.4 is subdivided into subsections detailing the solution and all its requirements, expanding on what was discussed in section 4.3.2, and disclosing its implementation details.

5.1 Recordings Acquisition

This project development would not have been possible without the crucial cooperation from CHLO, especially from Oeiras MUental Health nit (Unidade de Sáude Mental de Oeiras).

Our work required us to go to this specific unit soliciting patients to take part in our study, which occurred after their respective medical appointments. The health professionals at this unit were aware of our work and our presence at the establishment, therefore, they were the first to approach the patient regarding their possible participation. This effort from the health professionals eased our work. Furthermore, it was also preferable since the patient felt more at ease with a professional with whom he/she is familiar and knows how to best approach the patient.

Notably, patients were free to choose to take part in the experiment. This was referred explicitly to each patient multiple times, specifying that our work was in no way part of their appointment. This step is essential in this specific scenario since many of the patients that go to this unit for their appointments are mandated by the court to do so.

For each patient, we started by providing a general overview of the study, so as to not immediately overload the patient with too much information. This information allowed the patient to immediately understand the scope of the study. This information included:

1. **What was asked of the subject during the study**, by making sure the subject understood that no private or confidential information was required during the execution of the tasks.

2. That at any point before, during, or after the execution of the various tasks, he/she could have asked for the **immediate and complete deletion of any data acquired** during the completion of the study.

3. That all the information acquired would be **maintained confidential** and used only for our work or work that comes as a direct extension of our work. Moreover, some patients expressed concerns regarding the analysis of recordings by health professionals at the unit. Therefore, we made

sure to explicitly state that such analysis would never happen, and individual recordings would be processed exclusively by our team which has no overlap with the health professional at the unit.

Then, the acquisition of recordings from patients and healthy controls follow the same procedure. We started by explaining the study, its scope, requirements, and the extent to which the subject has control over the data acquired during the execution of the protocol. Afterward, we asked the subject to fill out his/her consent form, in appendix B for healthy controls, and in appendix C for patients diagnosed with psychosis or bipolar disorder.

The first two pages of the consent forms provide the same information that was explained initially to the subject by us, but in writing, providing contact details for further explanation. Then the next two pages ask the subject and the researcher for their signatures, and explicit consent from the subject. One of the pages is given to the subject and the other is archived by the researcher. The last page asks the subject for demographic information, and in the case of the patients, for his/her diagnosis and other relevant information, such as the years since diagnosis, and their score on the BPRS.

It is relevant to note that some of the information required from the diagnosed patients was afterwards cross-validated with the health professionals at the unit. This validation was required since some of the patients are less aware of their diagnosis, and correct and accurate information is required for the development of a good model. This validation typically involved the verification of the subject's diagnosis, years since diagnosis, and score on the BPRS.

For the recordings, our team used a *Zoom H6 Handy Recorder* with an *Omnidirectional XLR Lavalier Microphone* and *XYH-6 Stereo Microphone Capsule* connected. Although the *XYH-6 Stereo Microphone Capsule* provided stereo recordings, we decided on mainly using the audio tracks recorded with the lapel microphone. This decision was made by listening to the recordings and verifying that, although the difference was small, typically better audio quality was achieved through this microphone. This divergence between microphones was especially noticeable with some diagnosed patients that talked in especially low voice tones.

The recordings were carried out in various places. **Healthy controls** were recorded mainly indoors, in rooms and buildings, still, on some rare occasions, recordings took place outside. Concerning **diagnosed patients**, these recordings took place exclusively at the health care unit, albeit in various of the rooms available. This high variability of recording locales, and consequently high variability of acoustic features, across groups, was beneficial since it assured us that any conclusions extracted from the models developed were not a result of the models fitting to particular locale audio characteristics. Furthermore, it is important to mention that audio quality was always assured by us, by listening immediately after a recording session to excerpts from the audio recordings.

5.2 Execution Environments

As mentioned before, during the development of our models, we tested several variations of the same models, fine-tuning hyperparameters and classification techniques used. Moreover, some of the techniques employed for feature extraction were complex and had a high demand for computational resources. Due to both of these facts, we immediately concluded that using personal and local computers exclusively would not be plausible.

Part of our team belongs to Human Language Technology Lab (HLT), a group from Instituto de Engenharia de Sistemas e Computadores - Investigação e Desenvolvimento (INESC-ID) that has powerful remote machines for computational intensive tasks. HLT servers are subdivided into two categories: Central Processing Unit (CPU) machines and Graphical Processing Unit (GPU) machines. There are 12 CPU machines, identified as "`X##`", all with 48 gigabytes of Random Access Memory (RAM). Finally, there are 6 GPU machines, identified as "`G##`", all with 247 gigabytes of RAM available.

We prioritized the usage of CPU machines over GPU machines. We made this decision because we wished to explore a large range of models without blocking other researchers from carrying out their respective studies. The use of GPU machines should be avoided if possible since they are fewer in number, and therefore are more sought after by researchers, and are more costly to upkeep and maintain. All the code developed by us was adapted to run exclusively on CPUs.

Another concern was that the development of our models should continue even if the connection between our local machine and the remote HLT machine is closed. The connection can be terminated for various reasons, from weak internet connection to inactivity. To solve this problem in an easy and accessible manner, we used *TMUX*[1]. *TMUX* is an open-source terminal multiplexer that allows the user to quickly detach and reattach terminal processes. By detaching a terminal process it will keep on running as long as the machine stays on and the process is not manually killed by another user or machine administrator.

Finally, due to the high number of variations of models developed, we understood that their development could not be carried out sequentially, as this would take significantly more time than we could allocate for their development.

Initially, we set out to use *HTCondor* [43]. *HTCondor* is an open-source software framework that allows for the easy management of coarse-grained distributed parallelization. Effectively, *HTCondor* provides intuitive tools, that allow for the parallelization of several tasks/scripts in several machines in an automated fashion. This allowed us to very easily distribute and manage tasks from one of the CPU machines into all the remaining CPU machines. Unfortunately, *HLT* machines had version `V.8.6.13` of *HTCondor* installed, which did not support an explicit limitation of the number of concurrent jobs running on a single machine. This was a problem since, through experimentation, we discovered that some

[1] `https://github.com/tmux/tmux/wiki` accessed on September 25th, 2022

of the processes carried out were so memory intensive that, if a lot of parallel instances of the same process were running on a single machine, the machines would run out of memory.

To solve this problem I developed *Python* [44] code that would effectively take the same responsibilities as of *HTCondor*, allowing the submission of various tasks/processes to be carried out and then managing their execution on various remote machines. The code we developed allowed for the explicit definition of the maximum number of concurrent jobs on a single machine. Our parallelization code uses *Paramiko*[2] a *Python* package that relies on the *SSHv2* [45] protocol to connect and execute commands in other remote machines. Further comments on parallelization, specifically regarding the code changes that are required for the development of models in a parallel fashion, are described in section 5.4.3.

5.3 Recordings and Data Processing

From the proposed design for processing the recordings and data acquired (section 4.3.1), we then developed, the various sub-stages displayed in table 5.1. These stages are displayed through a table to facilitate readers' later reference about concrete sub-stages and to more easily reference these sub-stages on text. This table provides information such as the environment where the sub-stage is executed.

Table 5.1: Initial version of the various stages required for the correct processing of the data and acquired recordings.

#	Stage Name	Execution		Input/Output
5.1.1	Extract subject's available information	Local	Manual	**Input:** - **Output:** Subject's Information
5.1.2	Listen and create task timestamps	Local	Manual	**Input:** - **Output:** Time Annotations
5.1.3	Cut audios according to timestamps	Local	Script	**Input:** Recordings, Time Annotations **Output:** Audio Segments
5.1.4	Transcribe audio segments	HLT	Script	**Input:** Audio Segments **Output:** TRIBUS generated models
5.1.5	Extract automatic transcriptions	HLT	Script	**Input:** TRIBUS generated models **Output:** TRIBUS Transcriptions
5.1.6	Fix transcriptions	Local	Script	**Input:** Audio segments, TRIBUS Transcriptions **Output:** Fixed Transcriptions

The sub-stage 5.1.1 took place locally and manually since we simply parse through the filled-out consent form and register in a spreadsheet the subject's information. Sub-stage 5.1.2 was also carried out locally and manually. Furthermore, this sub-stage can be performed in parallel with sub-stage 5.1.1. To accomplish this sub-stage, 5.1.2, the researchers listened through the recordings, using *VLC*[3], and annotated timestamps attempting, as best as possible, to delimit each subject's task execution. The

[2] https://www.paramiko.org/ accessed on September 25th, 2022
[3] https://www.videolan.org/vlc/ accessed on September 25th, 2022

researchers also identified timestamps that delimited their interventions, as to be excluded from the final audio segments.

Somewhere between sub-stage 5.1.2 and sub-stage 5.1.3, the diagnosis of the patients was disclosed to us and we could move the subject's information and recordings into the appropriate folder.

To complete sub-stage 5.1.3 we developed a *Python* script that took as input the subjects' recordings and their respective time annotations and split the recordings into audio segments. To achieve this segmentation, the annotated times were first converted into milliseconds, and then, using a *Python* package called *Pydub*[4], split into audio segments, and finally saved appropriately.

The next two sub-stages, 5.1.4 and 5.1.5, both took place in HLT remote machines. Sub-stage 5.1.4 transcribed the various audio segments, for all the existing audio tracks, using *TRIBUS*. As mentioned before *TRIBUS* is an automatic transcriber for European Portuguese. *TRIBUS* is deployed in HLT's machines, therefore, our only possibility was to carry out this stage here. This transcriber was also the most advisable choice since *TRIBUS*' generation of transcriptions is a memory and time-intensive task. *TRIBUS* then exported the models generated together with the resulting transcriptions. The latter must then be separated from the remaining files and saved accordingly, for which we developed a *Bash* script.

The transcriptions generated by *TRIBUS*, were not perfect, and after a careful analysis, we concluded that it would be beneficial to fix these transcriptions, as best as possible. As shown in figure 5.1 it is noticeable that although not completely wrong, *TRIBUS* still miss transcribed some words, which can have a large impact in all models, but especially, in models that focused on content analysis of speech. Figure 5.1 shows a concrete example of the difference between the automatically generated transcription and the manually corrected transcription, for a given subject, during the execution of *Task 4*. Note that the word "porquinhos" was miss transcribed as the set of words "por" and "quinze". This is the purpose of the last sub-stage, 5.1.6.

For this last stage, we implemented a slightly more complex *Python* script, that took as input audio segments and the generated automatic transcriptions, and, according to the researcher's input, would output a corrected transcription. The script started by finding out a subject and task which had an automatic transcription yet to be manually corrected. Once the subject and task had been selected, the application asked the researcher to select one of the audio tracks, for which an automatic transcription had been extracted, as the template on which alterations would be made. As mentioned before, whenever available, we selected the track created by the *Omnidirectional XLR Lavalier Microphone*. The script then entered its main execution which iterated between: (i) **playing the segment of the audio track** mapped to the next word identified by the automatic transcriber; and (ii) **presenting the main menu** to the researcher, asking if any changes are to be made that involve the last played audio segment or associated transcription.

[4] https://github.com/jiaaro/pydub/ accessed on September 25th, 2022

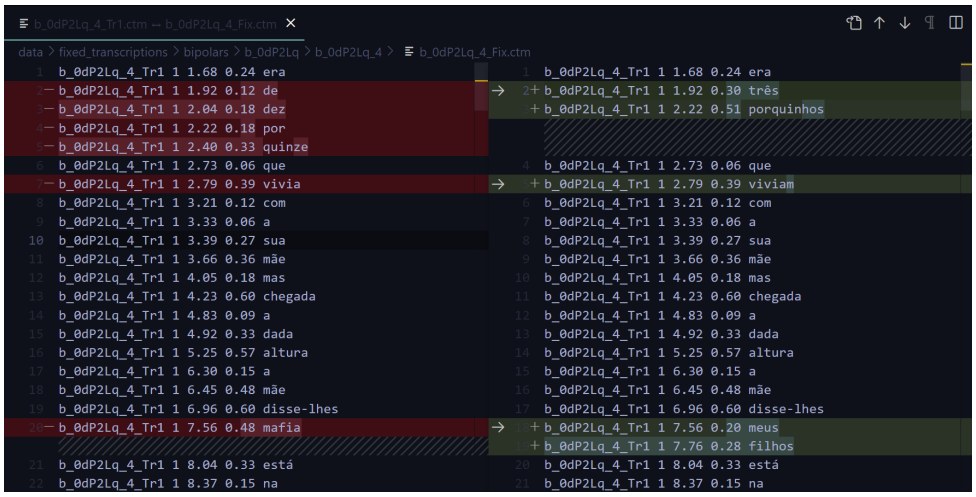


Figure 5.1: Effect of the process of manually adjusting and fixing automatically generated transcriptions.

The main menu of the application is shown in figure 5.2, and offers eight possible selections to the researcher: **(option 1)** replay the segment of the audio track just played; **(option 2)** move to the next segment of the audio track and its associated word; **(option 3)** move to the previous segment of the audio track and its associated word; **(option 4)** change word mapped to the segment of the audio track (researcher does not agree with the word but does agree with the time interval identified); **(option 5)** join the current segment with the next segment of the audio track (used when a single word has been misidentified as multiple words); **(option 6)** split the current segment of the audio track into multiple segments, each with its word associated (used when multiple words were miss identified as a single word); **(option 7)** place an End of Sentence (EOS) between the words associated with the previous and current segment of the audio track (we decided not to use this option for the identification of sentences in speech); **(option 8)** abort the application in its entirety not saving the progress made.

Once again, we feel it is important to note that our objective is not to develop a perfect transcription. This would require a larger team with knowledge and experience in this particular kind of task. Our purpose was to fix most of the obvious and identifiable errors, to minimize consequences as much as possible in the models developed that rely on content analysis.

Eventually, during the process of manually correcting the generated transcriptions, we identified a problem in our pipeline for recordings and data processing. The timestamps used to split the audio recordings into audio segments were not precise enough. Sometimes the audio segment would leave out a word spoken by the subject out of the audio segment or it would include a brief intervention of the researcher. These lapses happened both at the beginning and end of the audio segments. To solve this problem, we changed from the initial granularity of seconds to milliseconds, as should have been done initially.

However, this fix only would have solved the problem for future recordings to be processed, and at

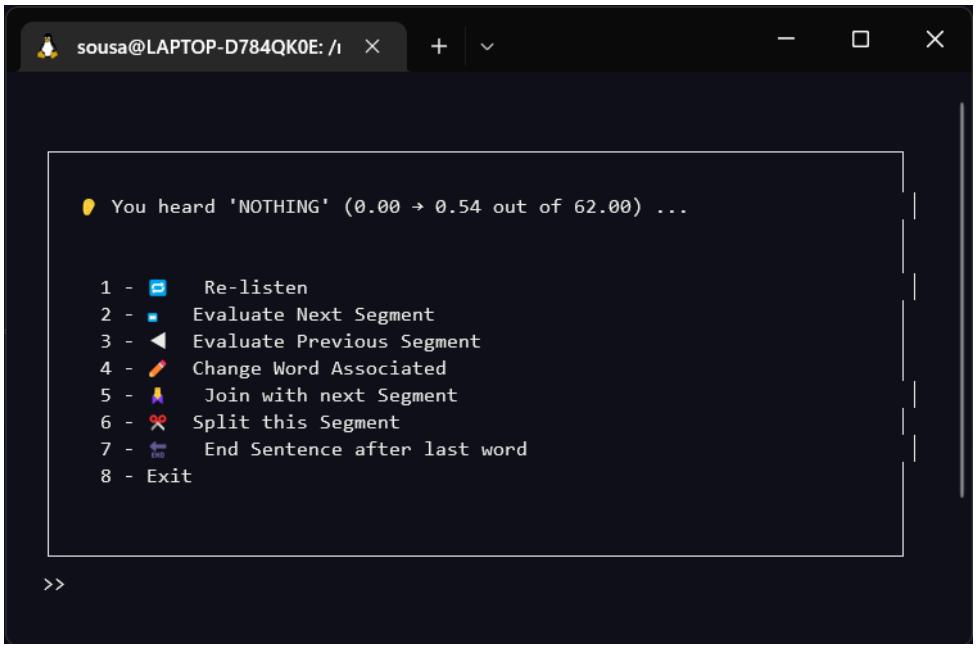


Figure 5.2: Main menu of the python script developed for the manual alignment and correction of the automatically generated transcriptions.

this point, we had already manually fixed most of the transcriptions. Simply rerunning the entirety of the pipeline with the new and more precise timestamps to delimit the subject's task execution, would mean that we would need to manually fix almost all the transcriptions once again. Due to time limitations, this was not a possibility. With this scenario in mind, we changed the pipeline from the one presented in table 5.1.

Table 5.2: Hotfix version of the various stages required for the correct processing of the data and acquired recordings. New developed stages of the pipeline are marked with bold.

#	Stage Name	Execution		Input/Output
	5.1.1 - 5.1.6			
5.2.1	**Check time annotations**	Local	Script	**Input:** Recordings, Initial Time Annotations **Output:** New Time Annotations, Changes Time Annotations
5.2.2	**Remove changed tracks**	Local	Script	**Input:** Changes Time Annotations **Output:** -
	5.1.3 - 5.1.5			
5.2.3	**Merge transcriptions**	Local	Script	**Input:** Audio segments, Initial Fixed Transcriptions, New TRIBUS Transcriptions **Output:** New Fixed Transcriptions

Note that stages 5.1.1-5.1.6 refer to the same stages carried out before. For stage 5.2.1 we developed yet another *Python* script that takes as input the various recordings, in their entirety, and the previously achieved time annotations for the various recordings. The script goes through the various

subjects and tasks, and for each subject-task pair identifies what we called the start and end fulcrum points, which are initialized with the initially identified start and end timestamps. Then, for each one of the fulcrum points, it plays the preceding two seconds, pauses for a second, and plays the following two seconds. If the researcher agrees with the fulcrum point, that it accurately separates the researcher's interventions from the subject's execution of the task, the execution ends, if not, an adjustment is made by the researcher and the entire process for the fulcrum point starts over.

Once the various time annotations had been adjusted, with a milliseconds granularity, we can carry out stage 5.2.2. For this stage, a script deleted every file that came as a result of the processing of an audio segment for which the time annotation had been changed. This was required to carry out stages 5.1.3-5.1.5 again, since these stages detect pairs subject-task for which they have yet to export their results to know what must be processed.

At last, for this pipeline, we required some sort of mechanism that would allow the merging of the initially developed manually corrected transcription (for which the time annotation was possibly wrong) with the new automatically generated transcription (for which the time annotation was corrected). Note that the only possible changes that could have happened, in terms of transcriptions, were either at the beginning, the ending, or both. Considering this, a *Python* script was responsible for iterating every subject-task pair for which a time annotation had verified changes, and then asking the researcher to adjust the transcription so that the two were in agreement. This was done by: **(i)** playing the affected parts of the audio segment; **(ii)** automatically adjusting the timestamps of all the words identified by the delay that the start timestamp had been shifted; **(iii)** automatically opening a text editor that displayed the changes between both files in an intuitive manner; **(iv)** asking the researcher, to manually correct the transcription, which had already been corrected for the initial time annotations.

Once this last stage has been completed for all the subject-task pairs that suffered changes, we know that we have **achieved meaningful manual corrected transcriptions**, at least to the best of our abilities. Nonetheless, the ***hotfix*** pipeline described is quite complex, since some of its stages are re-executed. If we had noticed that the time annotations were not precise enough before starting the manual correction of the various transcriptions, the process would have become much simpler.

Table 5.3 shows exactly the pipeline that would have been carried out if we had noticed sooner the time annotations problem. In fact, if more subjects, and therefore recordings, are to be added to the corpus, this pipeline can be followed. Notice that no stage is re-executed and that the ***improved*** overall pipeline is much simpler to follow and understand.

In this last pipeline, which we denominate as the *improved* pipeline, we also added a new stage. Sub-stage 5.3.8 assures the quality of the manually corrected transcriptions. This quality assurance is important since there is the possibility that whilst fixing the transcription other problems might have been introduced by the researchers. This sub-stage, 5.3.8, is carried out by another *Python* script that takes

Table 5.3: Improved version of the various stages required for the correct processing of the data and acquired recordings. New developed stages of the pipeline are marked with bold.

#	Stage Name	Execution		Input/Output
5.3.1	Extract subject's available information	Local	Manual	**Input:** - **Output:** Subject's Information
5.3.2	Listen and create task timestamps	Local	Manual	**Input:** - **Output:** Temporary Time Annotations
5.3.3	Check time annotations	Local	Script	**Input:** Recordings, Temporary Time Annotations **Output:** Time Annotations, Changes Time Annotations
5.3.4	Cut audios according to timestamps	Local	Script	**Input:** Recordings, Time Annotations **Output:** Audio Segments
5.3.5	Transcribe audio segments	HLT	Script	**Input:** Audio Segments **Output:** TRIBUS generated models
5.3.6	Extract automatic transcriptions	HLT	Script	**Input:** TRIBUS generated models **Output:** TRIBUS Transcriptions
5.3.7	Fix transcriptions	Local	Script	**Input:** Audio segments, TRIBUS Transcriptions **Output:** Fixed Transcriptions
5.3.8	**Check transcriptions**	Local	Script	**Input:** Fixed Transcriptions **Output:** Fixed and Corrected Transcriptions

as input the various transcriptions and checks them against a set of tests. These tests were developed in order to identify transcription errors, immediately correct transcription errors, or flag transcriptions suspected of being wrong, which we denominated error tests, fix tests, and warning tests respectively. The tests developed were:

- *Well Formatted (Error)*: Rather simple test, which verifies that the transcription follows a previously defined format. This format should be consistent across automatically generated transcriptions and manually fixed transcriptions. We decided on following the format in which *TRIBUS* exports its transcriptions, shown in listing 5.1.
- *Valid Duration (Error)*: Checks that every line in every transcription has a duration and that such duration is greater than 0.
- *Valid Timestamps (Error)*: This test verifies that a word at position x does not overlap with word $x + 1$. This is done by assuring that $start_x + duration_x <= start_{x+1}$. This is important since a given subject can't speak two words at once.
- *Valid Words (Error)*: Checks that every line from every transcription has exactly one word mapped to it and that such a word is lowercase.
- *Unusual Start/End (Warning)*: Signals transcription for which no word is mapped in the first 10 seconds or last 10 seconds. A transcription in this particular scenario is not necessarily incorrect. Nonetheless, it is worth analyzing if the time annotations and transcription itself are correct.

- *Check Word Sequences (Warning)*: Flags sequences of words that are commonly misinterpreted by the automatic transcriber. After a sequence of words has been flagged, the researcher verifies whether it should be fixed or not by listening once again to the recording. An example of this common misinterpretation was the word ‘porquinho‘ which was commonly mistaken for the sequence ‘por‘ and ‘quinze‘.
- *Fix Word Sequences (Error Fix)*: Automatically corrects words or sequences of words into other words or sequences of words. This test was mainly used to fix orthographic errors or inconsistencies across transcriptions. For example, in some transcriptions, ‘parakeet‘ was written like ‘piriquito‘ and in others ‘periquito‘.

Listing 5.1: Consistent format used in transcriptions, both automatic and mannual, across groups.

```
1 ...
2 <audio_track>  <start_seconds> <duration_seconds> <word_identified>
3 c_DdJefr_1_Tr1 41.64           0.84               carro
4 ...
```

Even after all the mentioned steps, stages, and pipelines, there is no assurance that the final achieved transcriptions are correct. Our focus was not on evaluating these methods' performance, or to assess manual transcriptions quality. Still, we understood that the models developed are only as good as the data used to train them. This is why data and recordings processing was such a crucial step for the development of our solution, described in such detail. We believe that the transcriptions achieved are significantly better than the ones achieved automatically. A deeper analysis of the transcriptions achieved is carried out in chapter 6.

5.4 Solutions Development

The following section describes the implementation and development of the models mentioned. We start by describing the various requirements that first must be met before any of our models are developed, in section 5.4.1. Section 5.4.2 describes the overall code structure of the project detailing the interactions between the various entities and modules developed. Finally, section 5.4.3, discusses the steps taken to parallelize the development of our models.

5.4.1 Solution Requirements

Some of the requirements in this section come as requirements from other requirements. These relations will be made explicit in the following subsection.

Note that this section is subdivided into three subsections. In section 5.4.1.A we describe the data requirements for the project. Section 5.4.1.B enumerates and details the various text processing applied during our work. Lastly, section 5.4.1.C describes the various models developed that were then used by the various feature extraction techniques.

5.4.1.A Data Requirements

Besides the more obvious requirement for a corpus composed of recordings of subjects diagnosed and not diagnosed with psychosis, which has already been extensively discussed, there are more requirements for data and corpora.

A – Corpus for Embeddings Several of the techniques mentioned require the existence of corpora, for which embeddings are developed. These techniques that rely on embeddings were described in detail in section 2.2.4. These embeddings are developed under the assumption that the corpora used for its development acts as a baseline for the expected discourse. Therefore, ideally, this corpora should be as close as possible, in topic and overall structure, to the discourse of the subjects during the execution of the protocol. On the other hand, our work is limited by the amount of data, and especially, available corpora. Generally speaking, data for European Portuguese is typically scarce or underdeveloped.

After researching freely available corpora, we understood that there was only one valid possibility. Most of the corpora available was for Brazilian Portuguese, or at best, for a mixture of European and Brazilian Portuguese. Even the corpora that was developed by a mixture of both languages typically rely more on data from Brazilian Portuguese. The only exception was Corpus de Extractos de Texto Electrónicos MCT/Público (CETEMPúblico) [46].

CETEMPúblico was developed by a partnership project between Ministério da Ciência e da Tecnologia (MCT), the Portuguese Ministry of Science and Technology, and Público, one the most renowned daily Portuguese newspapers. This corpus is made up of 1.485.828 extracts of articles published in Público. In total, this equates to a total of 190.6 million words and 234.5 million tokens. There are two versions of the corpus available:

- **Non-Annotated:** Extracts and their content are organized into a fixed hierarchical structure but without any NLP techniques applied to its content. This structure relies on Hypertext Markup Language (HTML)-like tags.
- **Annotated:** Extracts and their content are organized into a fixed hierarchical structure, with NLP techniques applied to their content. Note that this follows the same hierarchical HTML-like structure as the non-annotated version of the corpus. The content of the extract has been processed and the following NLP information has been extracted, among others: (i) POS tag of the word associated;

(ii) morphological information, such as verbal tense, gender, number, etc; (iii) lemma associated with the word.

B – Corpus for Valence Transformer Besides the corpus for the extraction of the embeddings, we also required data to fine-tune a transformer model for valence analysis. We decided on developing our own fine-tuned transformer model since, after some investigation by our team, we discovered that none of the available transformer models were fine-tuned specifically for valence analysis of European Portuguese. As explained in section 2.2.10.A it is possible to develop models that rely on *web-scraped information*.

We understood that the best available option was to extract reviews and their scores from users from various domains. The score is related to the written review, and therefore we could develop models that use this information, and that when fed text, would output a score that would express how positive or negative the text is.

To this purpose, we developed a *Python* project, which relied on abstractions, that would allow the researchers to more easily develop automatic scrapers for various web pages. The code relied mostly on two *Python* packages: (1) *Selenium* to open, access, and interact with web pages; (2) *BeautifulSoup* to parse the HTML source code, and more easily access, filter, and select HTML elements. In total we scraped five different domains for information:

- **Booking Scraper:** We scraped `https://www.booking.com` for booking reviews. *Booking* allows subjects to review their stays, identifying positive and negative aspects of their stay. We limited these reviews to the great metropolitan area of Lisbon and the European Portuguese idiom to approximate as much as possible the majority of the subjects that make up the First European Portuguese Corpus for Psychosis Identification.

- **CineCartaz Scraper:** We scraped `https://cinecartaz.publico.pt` for movie reviews. This domain is part of *Público*'s domain and allows users to write reviews and score movies that they have seen. There is a limitation to this scraper, the score is not directly linked to the user's review, and therefore users can score movies without necessarily writing a review for them. However, since this domain has limited traffic, and most of the movies have at most a couple of scores associated, we assumed that the general score of the movie would be approximate to the written review of any user for that movie.

- **Shein Scraper:** We scraped `https://pt.shein.com` for clothe reviews. *Shein* is a clothing online store that allows users to review products that they have bought and received. We restricted our scraping to a single category, dresses. We chose this category simply because it is the most popular of the online store, although others could have been used due to the enormous amount of

cloth categories, products, and reviews. *Shein* automatically displays reviews for the user's idiom, nonetheless, some reviews with the wrong idiom are displayed (mostly in English, and sometimes French). To solve this problem, we used a *Python* library called *LangDetect*[5]. This package can output the most probable language, and we used this output to filter out written reviews that are not in European Portuguese.

- **TrustPilot Scraper:** We scraped several review categories from `https://pt.trustpilot.com`. *TrustPilot* allows users to review various websites from various categories. TrustPilot automatically creates a default filter for reviews on the idiom of the user, so we did not need to filter out reviews based on the language review. We selected the following categories from *TrustPilot* to scrape: Money Insurance, Vehicles and Transportation, Jewelry Stores, Clothing Stores, Electronics and Technology, Fitness and Nutrition Services, Furniture Stores, Energy Suppliers, Real Estate Agents, Health and Medical Services.

While developing the scrapers and acquiring data from their scraping we faced several problems. We will not go into detail discussing these problems, since this is not the focus of our work. Nonetheless, we feel like it is important to mention them and their solution to allow the reproduction of our work. Many websites do not allow automatic access to their websites, so, one of the first steps when attempting to scrape a website was to verify their rules for automatic access and scraping, this can be accessed through the subdirectory `robots.txt`. Even when websites did not explicitly prohibit the scraping of their contents, some had implemented techniques to limit this access. For example, websites can detect the number of requests coming from a single Internet Protocol (IP) address and if no Graphical User Interface (GUI) is opened. To solve these problems we used selenium with:

- GUI opened up to simulate a real user.
- Automatic rotation of IP addresses after a certain number of requests. The IP addresses came from a freely available list of proxies that can be used to relay requests. Important to note that since this list is openly available to the public many of the IP addresses are already in use and overloaded or outright blocked by websites.
- Automatic rotation of user agents, which are also used by websites to detect and block requests. The list of user agents comes from a static list of possible user agents provided by a *Python* package.

5.4.1.B Text Processing Techniques

In the following table 5.4 we display the NLP techniques applied when processing the text obtained from the transcriptions generated for the execution of each task by each subject. This table also displays

[5] `https://pypi.org/project/langdetect/` accessed on October 1st, 2022

exactly which feature extraction technique required each one of the NLP techniques.

Table 5.4: Text processing techniques required by each feature extraction technique in order to achieve the structure, coherence, and content feature sets.

Support Technique	**Structure and Coherence**			**Content**		
					Semantic Analysis	
	LSA	Word Graph	Vector Unpacking	LCA	Dictionary	Transformer
Text Lemmatization	×		×	×	×	
Stop Words Removal	×		×	×	×	
Sentence Segmentation	×		×	×		

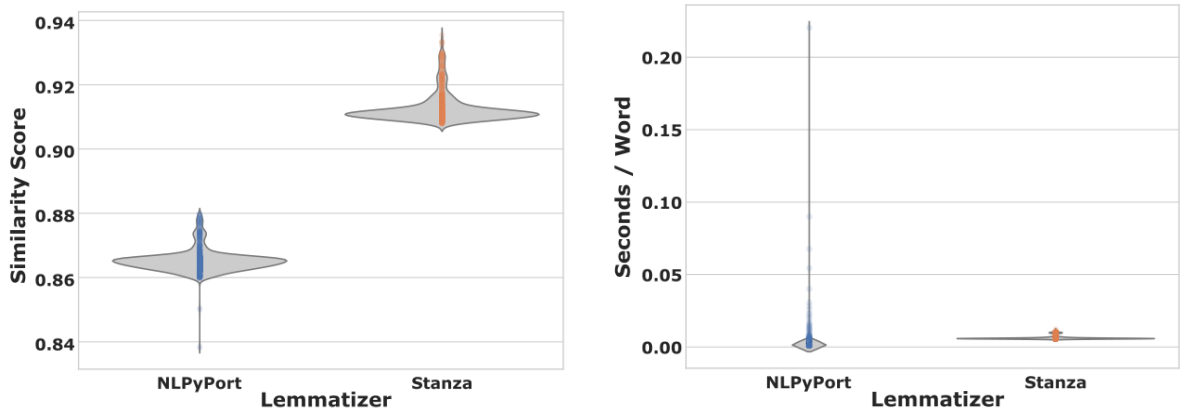
A – Text Lemmatization Once again, the problem that we faced was that lemmatizers for European Portuguese are scarce, paid, and/or underdeveloped. After some research, we identified these alternatives: (1) *NLPyPort*[6] which is a pipeline developed as a derivation of Natural Language Toolkit (NLTK) [47] pipeline, and (2) *Stanza*[7] [48], similarly to NLTK, offers a collection of methods developed into a pipeline, that allow for accurate and efficient analysis and processing of text.

To evaluate which one of the lemmatizers best fitted our requirements, we selected randomly 2500 extracts from the CETEMPúblico, and then fed the non-annotated version to each one of the lemmatizers and compared each one of the results to the annotated version of the corpus. Figures 5.3(a) and 5.3(b) express the scores and durations obtained for each one of the lemmatizers tested. The score was computed through the edit distance algorithm transformed into a similarity measure. With this measure, a score of 1 means that there is no difference between the lemmatizer output and annotated version of the corpus. The duration is expressed in words per second as to be independent of the extraction size.

Note that the results obtained do not express the quality of any of the lemmatizers, but instead express how well the lemmatizer adapts to our specific requirements. It is clear from the figures, that *Stanza* achieved the highest scores in the shortest amount of time. The score suggests that the format in which *Stanza* lemmatizes words coincides with the lemmas from CETEMPúblico. Therefore, our team decided on using *Stanza* for the lemmatization of the transcriptions.

B – Stop Words Removal Regarding the set of stop words, we decided on using the standard for the processing of text. As mentioned before, NLTK is one of the most well-known packages and a standard when it comes to text processing using *Python*. NLTK offers a vast set of tools, with one being sets of stop words. We used the default set of stop words for European Portuguese.

[6] https://github.com/NLP-CISUC/NLPyPort accessed on October 1st, 2022
[7] https://stanfordnlp.github.io/stanza/ accessed on October 1st, 2022



(a) Similarity score, computed through the edit-distance formula.

(b) Duration results expressed as seconds per word lemmatized.

Figure 5.3: Validation tests for each one of the lemmatizers selected.

C – Sentence Segmentation Once again, our team faced the recurrent problem of not existing models for European Portuguese, in this case regarding sentence segmentation. *Sentence segmentation* is an important NLP technique that some of the feature extraction techniques employed rely on.

In this particular case, no acceptable model was found, Consequently, we decided on employing a variation of a common NLP technique. This technique is called *n-grams* [49, 50] and subdivides words into sets of sequential words of equal length. *N-grams* creates all possible sequences of words of n-length. We were only interested in the non-overlapping groups of sequential words. Effectively, subdividing the transcription into sets of equal-length.

This technique is not ideal, since sentences are more than an exclusively structural entity. Sentences carry meaning, and their wrong delimitation can change the entire meaning of the text.

5.4.1.C Support Models

Some of these techniques require the previous development of models that serve as their support. The various support models developed and techniques that require such models are displayed in table 5.5. In the following subsections, we detail each one of the support models developed.

A – Word2Vec Embeddings We explored possible methods of developing a *Word2Vec* model to compute word embeddings. We decided on following the simplest strategy and avoiding '*reinventing the wheel*'. To this objective, we used *Gensim*[8] [51] a *Python* package that allows for the easy development of topic models such as *Word2Vec*.

[8] https://radimrehurek.com/gensim accessed on October 2nd, 2022

Table 5.5: Support models required by each feature extraction technique in order to achieve the structure, coherence, and content feature sets.

Support Technique	**Structure and Coherence**			**Content**		
					Semantic Analysis	
	LSA	Word Graph	Vector Unpacking	LCA	Dictionary	Transformer
Word2Vec Embeddings			×	×		
Latent Semantic Analysis Model	×					
Valence Dictionary (SentiLex)					×	
Valence Transformer (RoBERTa)						×

Gensim offers great features and advantages when compared to other alternatives. *Gensim* is fast, open source, and allows for the use of data streaming. The integration with data streaming structures was the main factor that made us choose *Gensim*. Without such capability, we would need to load the entire corpora, with which we want to develop the models, to memory. Loading the entire corpus is not plausible, at least with the computation resources that we had available.

Specifically, regarding the development of a *Word2Vec* model, we used `gensim.models.Word2Vec`, feeding, through streams, the various CETEMPúblico's extracts. When developing a *Word2Vec*, we must decide on a vector size to which the embeddings are generated. Due to time constraints, we decided on following the '*rule of thumb*' that is commonly followed when choosing a dimensionality, somewhere between 50 to 300, with 300 used for exceptionally large corpora. CETEMPúblico is a large corpus, still, not exceptionally big, consequently, we decided on using a dimensionality of 200.

B – Latent Semantic Analysis Similarly to *Word2Vec*, a model for *LSA* was developed using *Gensim*. *Gensim* also allows for the development of an LSA model through its developed class called `gensim.models.lsimodel`. This model also required a pre-established dimension to which the embeddings are formatted. However, this time, instead of following a '*rule-of-thumb*' value, we experimented with various dimensionalities, or topics as they are called by *Gensim*.

The following figure 5.4 summarizes the results obtained whilst studying the development of LSA models with varying dimensionalities. It is clear from figure 5.4, that there is a negative correlation between the dimensionality/number of topics and the score associated with the model. We scored each LSA model through its coherence score. Coherence is directly proportional to the semantic similarity between words that are rated highly in a specific dimension.

From this study, we decided on using a dimensionality of 20. This was the dimensionality for which we achieved higher coherence scores. It is important to note that this study served only as a mere indication of the value to follow. Once again, there is no guarantee that the best results will be obtained for this dimensionality, since: (i) the coherence of the various models is measured on a corpus (CETEMPúblico)

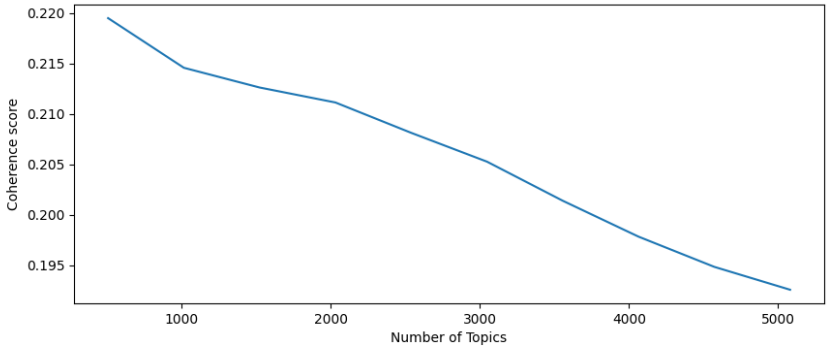


Figure 5.4: Results were obtained whilst studying the effect of varying the dimensionality in a *Gensim* LSA model.

which is not completely identical to the one where it is going to be applied (*First European Portuguese Corpus for Psychosis Identification*), (ii) and coherence does not express the quality of a model, simply how well separated its topics or clusters are from one another.

C – Valence Dictionary One of the techniques that we used for semantic analysis relies on the existence of a dictionary that maps words or lemmas to valence scores. After some research, we found *Sentilex*[9] [32].

Sentilex has valence scores associated with both inflected and lemmatized words. We used the subset with lemmatized words since we had this capability and by doing so we increased the likelihood of finding a match to the word in the dictionary. According to the authors of *Sentilex*, this dictionary is especially useful for sentiment and opinion mining of European Portuguese texts. This dictionary is available to the public. Consequently, we only needed to access it, download it, and convert it into a more appropriate format to which we could more easily access while extracting features from the subjects' execution of the tasks.

D – Valence Transformer The other technique that we used for sentiment analysis relies on a fine-tuned transformer. In section 5.4.1.A, we detailed the methods implemented to obtain the information that is to be used to fine-tune the transformer. Therefore, we only needed to fine-tune the model with this information.

In order to achieve this fine-tuned transformer we used a *Python* package, *transformers*[10], developed by *HuggingFace*[11] [52]. The package *transformers* provides an Application Program Interface (API) and tools to efficiently and easily download and train transformer models. *HuggingFace* provides access to various models and numerous fine-tuned versions of the same models. Nonetheless, none of the

[9] https://b2find.eudat.eu/dataset/b6bd16c2-a8ab-598f-be41-1e7aeecd60d3 accessed on October 2nd, 2022
[10] https://huggingface.co/docs/transformers/index accessed on October 2nd, 2022
[11] https://huggingface.co/ accessed October 2nd, 2022

fine-tuned versions satisfy our specific requirements, a valence model for European Portuguese.

The non-fine-tuned model that best fits our requirements is *XLM-RoBERTa*[12] [53]. We used the `base` version since the computational resources that we had available could not load and efficiently work with the `large` version.

Our model was fine-tuned for 100 ephocs and a batch size of 4. The batch size used is relatively small, however, a batch needs to be loaded in its entirety to memory. Due to limitations of the computational resources, we were incapable of increasing this batch size. To amend this problem, we set the gradient accumulation steps to 8. This means that the gradient is accumulated for 8 stages, effectively increasing the batch size during training to $4 \times 8 = 32$, and only after executing the backward propagation step, updating the transformer weights.

5.4.2 Models Development

This section not only details the steps taken but also concepts and definitions essential to a good understanding of the work completed.

5.4.2.A Entities and Abstractions Developed

The abstractions used throughout the project are detailed in the following section and were developed in the directory `modules_abstraction`.

A – Variation Represents a variation to be tested and for which results are then saved and displayed after the models have been developed. Note, that these variations either filter the First European Portuguese Corpus for Psychosis Identification into a subset which we aim to study, or define general attributes of the pipeline used to develop our models. A variation is defined by:

- **Target Task/s**: The protocol's task to be tested and for which the models were developed. This attribute filtered the original corpus into a subset that we aimed to study. Note that this attribute may refer to more than one task through an appropriate code. This attribute may be set with: *'Task 1'*, *'Task 2'*, *'Task 3'*, *'Task 4'*, *'Task 5'*, *'Task 6'*, *'Task 7'*, *'Verbal Fluency'*, *'Reading + Retelling'*, and *'Description Affective Images'*.

- **Target Gender/s**: The gender of the subjects to be tested. This attribute filtered the original corpus into a subset that we aimed to study. Note that this attribute may refer to more than one gender through an appropriate code. This attribute may be set with: *'Male'*, *'Female'*, and *'All Genders'*.

[12] https://huggingface.co/docs/transformers/model_doc/xlm-roberta accessed on October 2nd, 2022

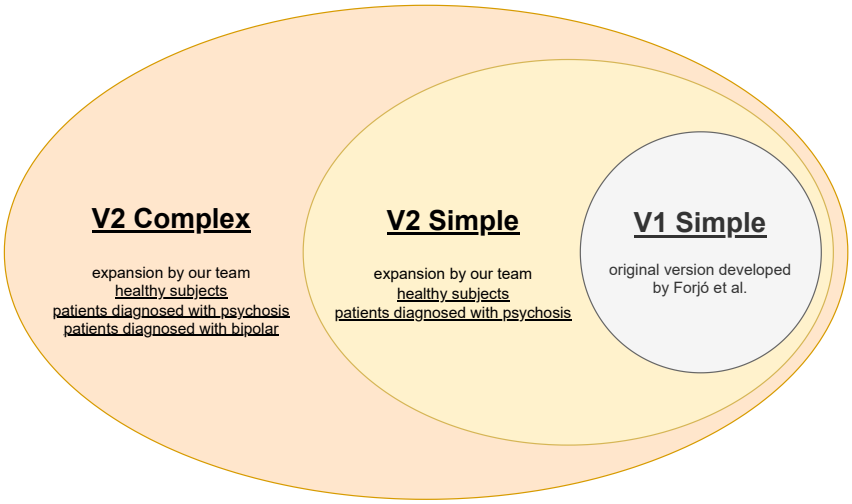


Figure 5.5: Target data variations were tested during model development.

- **Target Data**: The various data variations to be tested. We tagged each one of the subjects recorded into a set of categories, as displayed in figure 5.5. The variations allowed for the assessment of the impact of the corpus extension and of a more diverse control group (with patients diagnosed with bipolar) on the results obtained. This attribute filtered the original corpus into a subset that we aimed to study. This attribute may be set with: *'V1 Simple'*, *'V2 Simple'*, and *'V2 Complex'*.

- **Classifier Used**: The classifier technique to be used to develop our model. This attribute defined our pipeline, and consequently, affect the model developed but not the data that was used. This attribute may be set with: *'Naive Bayes'*, *'Decision Trees'*, *'Support Vector Machine'*, *'Random Forest'*, and *'Multi-Layer Perceptron'*.

- **Prepossessing Stages**: The prepossessing stages are to be carried out on the data before its fed to the classifier. This attribute defined our pipeline, and consequently, affected the model developed but not the data used. This attribute may be set with: *'DROP_ROWS_NAN'*.

Although our development would allow for the exploration of $10 \times 3 \times 3 \times 5 \times 1 = 450$ different variations out of which there would be $10 \times 3 \times 3 = 90$ different variations of the dataset used, we decided on following a simpler approach. The values set when following this simpler approach are displayed in table 5.6. By following this simplification we only explored $7 \times 1 \times 3 \times 5 = 105$ model variations, and $7 \times 1 \times 3 = 21$ dataset variations.

Finally, note that each one of these variations will be carried out for: (1) each one of the feature sets understudy, which will be further detailed ahead, and (2) each one of the classifier variations, which will be described in detail in the following subsection. This means that the number of variations explored will be much greater.

Table 5.6: Variations tested during the development of our models.

Variation Attribute	Values	#
Target Tasks/s	{'Task 1', 'Task 2', 'Task 3', 'Task 4', 'Task 5', 'Task 6', 'Task 7'}	7
Target Gender/s	{All Genders'}	1
Target Data	{'V1 Simple', 'V2 Simple', 'V2 Complex'}	3
Classifier Used	{'Naive Bayes', 'Decision Trees', 'Support Vector Machine', 'Random Forest', 'Multi-Layer Perceptron'}	5
Prepossessing Stages	{'DROP_ROWS_NAN'}	1
Total number of variations tested:		105

B – Variation Generator This abstraction was developed to manage and develop all possible variations. When fed the various variation attributes' possible values displayed in table 5.6, outputs all possible combinations of these values as variations. Moreover, it accepts extra arguments besides the ones directly fed to the variations:

- **Repetitions to perform**: Attribute used With the intent of repeating the development of the same model multiple times to obtain a confidence level, this attribute represents the number of repetitions to be carried out. This attribute is optional, by default each model was developed only once.
- **Study Feature Importance**: Attribute used With the intent of studying the importance of each feature on the models developed. By default, this attribute was set as `False`. When set with `True`, later on when the feature sets had already been fully developed, new variations were developed. For each variation, that had already been defined, and each feature that made up the feature set, a new variation was generated in which the importance of the selected feature was studied.

C – Classifier We experimented with five different techniques commonly used for the development of classifiers: NB, DT, SVM, RF, and MLP. We also mentioned that we wished to experiment with variations on the hyperparameters of these various techniques, to achieve at least a good local maximum for the models developed. For each one of the variations previously mentioned these hyperparameter variations were generated. Meaning that we exponentially increased the number of variations carried out.

We will not dive into detail on the meaning of the various hyperparameters and the reason behind the values chosen to experiment with. Nonetheless, to allow for the reproduction of the work developed, we provide a list of the hyperparameters, a brief description of the hyperparameter, and the values chosen to experiment with in the following tables. Tables 5.7, 5.8, 5.9, 5.10, and 5.11 display this information for NB, DT, SVM, RF, and MLP respectively.

The Classifier abstraction besides being responsible for the development of a classifier and its hyperparameter variations is also responsible for the development of train-test splits when given a feature set

Table 5.7: Hyperparameters variations of the Naive Bayes classifier tested during the development of our models.

Naive Bayes Variations			
Hyper Parameter	**Description**	**Values**	**#**
Algorithm	Algorithm used to compute likelihoods of the features	{'Gaussian', 'Bernoulli'}	2
Total number of variations tested:			2

Table 5.8: Hyperparameters variations of the Decision Tree classifier tested during the development of our models.

Decision Tree Variations			
Hyper Parameter	**Description**	**Values**	**#**
Criterion	Criterion used to measure quality of a tree node split	{'gini', 'entropy'}	2
Maximum Depth	Maximum depth to which the tree is allowed to develop	{1, 2, 4, 8, 16, 32, 64, 128}	8
Maximum Features	Maximum number of features when splitting a node	{'None', 'auto', 'sqrt', 'log2'}	4
Minimum Impurity Decrease	Impurity reduction threshold when splitting a node	{0, 0.1, 0.2, 0.4, 0.8}	5
Total number of variations tested:			320

as an argument. In our work, the number of subjects is still reduced, even after the corpus expansion. Consequently, we decided on employing *Leave-One-Out Cross-Validation* to assure that the models developed do not overfit the data and split provided.

D – Feature Set This abstraction represents a general-purpose feature set. This class was implemented to facilitate as much as possible the definition of new feature sets, focusing on the relevant and differentiating logic, and the methods utilized for feature extraction.

When defining a new feature set, three different methods must be defined, which represent three different stages in the development of the feature set. The names of the stages are merely representative. These stages are:

1. '**Basic Feature Set**': Columns of the feature set that are immediately developed when the feature set is instantiated. These columns typically extract information from subjects and their transcriptions and process it (for example lemmatizing and removing stop words). The generated informa-

Table 5.9: Hyperparameters variations of the Support Vector Machine classifier tested during the development of our models.

Support Vector Machine Variations			
Hyper Parameter	**Description**	**Values**	**#**
C	Regularization parameter	{0.25, 0.5, 1, 2, 4}	5
Kernel	Kernel type to be used in the algorithm	{'linear', 'poly', 'rbf', 'sigmoid'}	4
Total number of variations tested:			20

Table 5.10: Hyperparameters variations of the Random Forest classifier tested during the development of our models.

Random Forest Variations			
Hyper Parameter	**Description**	**Values**	**#**
Number of Estimators	Number of trees to be developed in forest	{10, 50, 100, 150}	4
Criterion	Criterion used to measure quality of a tree node split	{'gini', 'entropy'}	2
Maximum Depth	Maximum depth to which the tree is allowed to develop	{1, 2, 4, 8, 16, 32, 64, 128}	8
Maximum Features	Maximum number of features when splitting a node	{'None', 'auto', 'sqrt', 'log2'}	4
Minimum Impurity Decrease	Impurity reduction threshold when splitting a node	{0, 0.25, 0.50}	3
Total number of variations tested:			768

Table 5.11: Hyperparameters variations of the Multi-Layer Perceptron classifier tested during the development of our models.

Multi-Layer Perceptron Variations			
Hyper Parameter	**Description**	**Values**	**#**
Hidden Layers Structure	Number of neurons in each hidden layer	{(50,), (100,), (100, 50), (50, 100, 50)}	4
Activation Function	Activation function in each hidden layer	{'logistic', 'tanh', 'relu'}	3
Learning Rate Initialization	Value to which the learning rate is initialized	{0.001, 0.005, 0.025, 0.125}	4
Learning Rate Update	Update function for learning rate	{'constant', 'invscaling', 'adaptive'}	3
Maximum Iterations	Maximum number of iterations	{100, 500, 1000}	3
Total number of variations tested:			432

tion will typically be used by multiple feature extraction techniques. By doing this first development of the feature set, we compute only once each new column instead of once for each extraction technique. This part of the feature set can be carried out row by row, meaning that for a given row, the computed value for the new column is only dictated by the remaining values of that particular row.

2. '**Static Feature Set**': Columns of the feature set that are created to carry out a specific feature extraction technique. This part of the feature set is also carried out row by row, meaning that for a given row, the computed value for the new column is only dictated by the remaining values of that particular row. This part of the feature set is independent of the separation of the feature set into train and test sets.

3. '**Dynamic Feature Set**': Columns of the feature set that are created to carry out a specific feature extraction technique and that rely on the separation of the feature set into train and test sets. Some of the techniques applied for feature extraction assume a previous establishment of a baseline or an auxiliary model, and this must be done exclusively with the train set. We were mindful of this restriction to prevent data leakages.

Another abstraction that inherits from this abstraction was also developed. *'Merged Feature Set'* inherits from *'Feature Set'*, and allows for the quick definition of feature sets that are made of other feature sets. This new abstraction is implemented in a way that prevents the information from being redeveloped multiple times during a single execution.

5.4.2.B Feature Sets Developed

During the development of our study, we developed the following feature sets:

- **Speech Feature Set**: Feature set developed with the intent of replicating the work of Forjó et al. [5]. This feature set focuses on speech features that are extracted directly from the audio segments and their respective transcriptions.
- **Sound Feature Set**: Feature set developed with the intent of replicating the work of Forjó et al. [5]. This feature set focuses on sound features extracted from audio segments using *eGeMAPS* [54].
- **Speech + Sound Feature Set**: Merged feature set that aggregates the speech and the sound feature sets. This feature set was developed with the intent of replicating the study of Forjó et al. [5].
- **Structure / Coherence Feature Set**: This feature set focuses on structure and coherence features extracted from the transcriptions associated with the subjects' execution of the various tasks.
- **Content Feature Set**: This feature set focuses on content features extracted from the transcriptions associated with the subjects' execution of the various tasks.
- **Structure / Coherence + Content Feature Set**: Merged feature set that aggregates the structure/-coherence and the content feature sets. This feature set and the results obtained from models that utilize these feature sets were the main focus of our work.

The implementation of these feature sets, and their support methods, are all grouped under the project directory `modules_features`. In directory `modules_features > support`, we implemented support classes and methods for the definition and development of the feature sets. Generally speaking, in this directory, each *Python* file is associated with a feature extraction technique that is later used when developing the feature set. In the following sub-sections, we provided further details into the implementation of the various feature extraction techniques employed when developing *Structure / Coherence* and *Content* feature sets.

A – Extraction of Word Graph Features In order to simplify the development of word graphs from subjects transcriptions, we utilized a *Python* library called *NetworkX*[13] [55] . This library allows for

[13] https://networkx.org// accessed on October 10th, 2022

the easy definition of graphs by specifying their various nodes and edges. This definition is done by initializing a *NetworkX*'s *MultiDiGraph* and then feeding it every sequential pair of two words from a given transcription. The mentioned structure facilitates the identification of LSCC and SCCs and the extraction of the relevant word graph features.

B – Extraction of Latent Semantic Analysis Features We started by subdividing the transcribed, lemmatized, and filtered words into groups of 15 words, following the mentioned strategy for sentence segmentation. We chose 15 words as the length for the equal-length sentences since on average sentences are made up of 15 to 20 words, and during the execution of the protocol's tasks, we identified sentences, from both patients and healthy subjects, as being remarkably small.

Once the transcriptions' words had been split up into 'sentences' or groups, we mapped each one of these words into an appropriate embedding. This embedding is obtained through the LSA model whose development has been detailed in section 5.4.1.C. Then, the various word embeddings of each word group are averaged out, as to achieve 'sentence' embeddings.

We replicated the studies mentioned in chapter 3 by extracting First Order Coherence (FOC) and Second Order Coherence (SOC) as features for the identification of psychosis. These features were computed by averaging the cosine similarity of each 'sentence' with the sentence one and two positions ahead, respectively.

C – Extraction of Vector Unpacking Features Similarly, we started by subdividing the transcribed, lemmatized, and filtered words into groups of 15 words, following the mentioned strategy for sentence segmentation, described in subsection 5.4.1.B.

The various words that made up the various word groups were then mapped to the corresponding embeddings using a *Word2Vec* model, which development was previously detailed in section 5.4.1.C. Additionally, embeddings for the various groups of words were also computed by averaging the various word embeddings previously acquired.

Then, following the methodology of Rezzai et al. [13], the various word embeddings and corresponding group embeddings were then fed through the custom neural network as input and expected output, respectively. This custom neural network was composed of two layers, the input, and output layers, connected as displayed in figure 5.6. The neural network had the objective of minimizing the sum of squared errors by updating the various word weights. At any time, if the weights had fallen beneath a certain threshold defined through the following equation 5.1, the weight value was set as 0.

$$\frac{iteration\,number}{\tau \times max\{iterations\}} \tag{5.1}$$

with τ being a constant with a value of 100, achieved through experimentation.

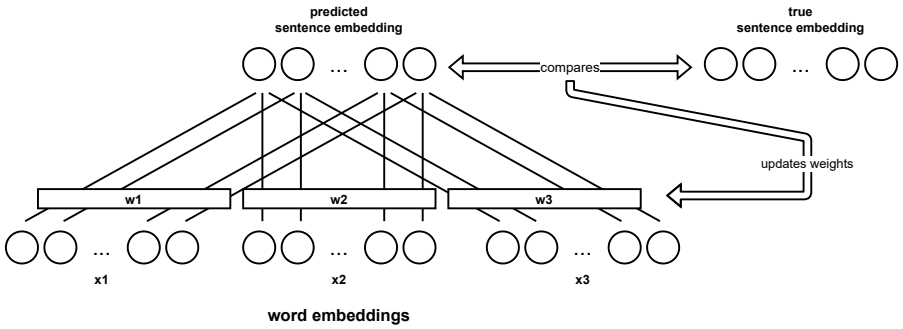


Figure 5.6: Structure of Rezaii's Neural Network used for the acquisition of Vector Unpacking Features.

When the neural network had finished updating the weights and minimized the sum of squared errors, we extracted the various vector unpacking features used for classification.

D – Extraction of Latent Content Analysis Features As preparation for the extraction of LCA features, we started by selecting the 10% most common words for European Portuguese. This selection of the most common words was done through CETEMPúblico [46]. We selected the 10% most common words instead of the 95% most common words, as reported by Rezaii et al. [13], to minimize space complexity and since we understood, from experimentation, that results suffered almost no impact from this reduction.

Afterward, we subdivided the words into sets of equal length, each group of words comprised of 15 words, and computed word embeddings and, by averaging out these word embeddings, 'sentence' embeddings. The word embeddings were obtained through a *Word2Vec* model, which implementation was previously described.

Then, for each transcription and each one of the most frequent words, we computed the highest cosine between any of the transcription groups and the frequent word. Subsequently, we selected the top 50 words that maximize the difference of cosines between groups, whilst considering their prevalence amongst all the documents by applying *TF-IDF*. Effectively we achieved the 50 words most prevalent in meaning in one group and least in the other.

From these words, we developed clusters, choosing the number of clusters that maximize their silhouette coefficient. An example of the clusters developed can be seen in Figure 5.7.

As features for the classification task, we required a value that expressed the similarity of the transcription with each one of the clusters. To this goal, we computed the highest cosine value from the various group embeddings, of a given transcription, with each one of the cluster centers computed. We expected these clusters to differ according to the group, and the distances to these clusters to express this difference in topic.

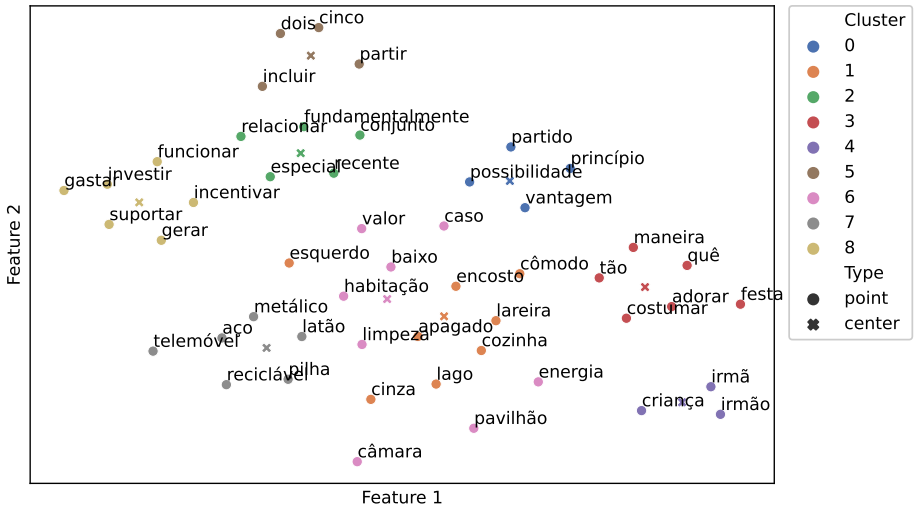


Figure 5.7: Word clusters developed for patients diagnosed with psychosis on Task 6 using Latent Content Analysis and KMeans clustering. Manifold TSNE was used to visualize high-dimensional data in two dimensions.

E – Extraction of Valence Features with Sentilex We have already detailed the acquisition of Sentilex as a corpus with valence scores associated with lemmas. Therefore, to compute a valence score associated with each transcription, we lemmatized the transcriptions and then found out if each lemma was contained by Sentilex's dictionary. The various scores identified were then averaged and counted, and each one of these characteristics was used as a feature for models development.

F – Extraction of Valence Features with RoBERTa The development of a RoBERTa model for the acquisition of valence scores has already been detailed. Since this model was initially developed and fine-tuned with unfiltered and non-lemmatized text, the 'raw' text from the transcription was fed to the developed fine-tuned model. The output from this model was then used as a feature for the classification task.

5.4.2.C Studies Developed

In the previous chapters and sections, we defined in detail the various studies developed, meaning, the type of classifier tasks carried out throughout our work. Even though this difference between studies exists, code-wise this difference is less apparent. In `modules_abstraction > module_models.py` we define variation generators, which directly map to the studies mentioned and developed, exploration classification task and assessment classification task.

With regard to the assessment classification study, in which we study in depth the most relevant

features for the results obtained and establish a confidence level for the models developed, we consider important to further describe how these objectives were accomplished. To study the models' confidence level, we re-developed each one of the best models achieved, one for each task, multiple times in order to understand how reliable and robust they are. We decided on carrying out 10 repetitions, as it provided us with an understanding of the underlying consistency of the models without overloading machine resources, both in space and time.

Note that some of the features extracted were dynamic in the sense that a varying number of features might be extracted from transcriptions, dependent on the specific train-test split. This variability in the number of features extracted happens with LCA, in which the number of clusters developed was chosen so that the silhouette coefficient was maximized. For this reason, when studying the confidence of these features, we exclusively analyzed the feature which relies on the first developed cluster for target and non-target. No further study was possible, since nothing guaranteed that all the tasks and all the train-test splits would have such a feature.

Regarding the study of features importance, we researched valid techniques that would estimate feature importance on the model. We established three different approaches: (i) the iterative removal of features from the dataset, (ii) the iterative addition of features to the dataset, (iii) and the randomization of the values on a particular feature.

The first and second would allow for an understanding of the impact of the addition and removal of features. However, these techniques would generate too many combinations and the results would become much more difficult to analyze. We decided on following instead the third approach, the randomization of the values on a particular feature to explore its importance [56, 57]. Through this technique, we could understand how dependent the models developed were on the particular feature. If the results were impaired by the randomization it means that the model adjusted, during training, to a particular feature, which in the test set revealed to be less capable of classifying subjects correctly. Consequently, feature importance was calculated through the following equation 5.2.

$$\frac{initial\ score - feature\ randomized\ score}{initial\ score} \tag{5.2}$$

5.4.3 Models Parallelization and Efficiency

As mentioned in section 5.2, it would not have been plausible to develop all of the model's variations sequentially. Instead of running the models as a whole, we ran each model variation separately. Once all of the variations had finished, we merged the information from their development into a single document. We can concretely identify four different stages for the parallelization of models:

1. Initial acquisition of the variations to carry out. At this stage the model was selected, the variations to be carried out were computed, and the number of variations was saved to a temporary file.

2. Creation of *bash* scripts, one for each variation, to be carried out by our parallelization script. We followed this strategy of developing different *bash* scripts since it was the standard approach for *HTCondor*, and therefore was also the standard approach for our parallelization script.

3. Execution of all the *bash* scripts, by our parallelization script, in a parallelized manner.

4. Merging off the results obtained from each one of the variations carried out into a single file which correctly expressed the capability of the models and their results.

Even with the parallelization described the model's execution was unreasonably time-consuming for the time allocated for the execution of our work. Consequently, we implemented further mechanisms for the improvement of the overall efficiency of the models' development. Necessarily, we were forced to increase its space complexity in order to reduce its time complexity. This was achieved by saving all of the feature sets developed, in all their different stages, in temporary files (*checkpoints*). These temporary files can then be read by other variations and therefore prevent the development of the same feature sets multiple times. This process is not as straightforward as one would imagine, since: different variations can have or not have the same target feature set and *'dynamic'* feature sets rely on the specific train-test split for which it was developed, implying that in a set with N records, N feature sets are developed all of which must be properly stored and identified. Nonetheless, by creating meaningful codes that uniquely identify each dataset, we solved this problem.

The selection of an appropriate code for the identification of the dataset was crucial. The **selection** of variation attributes which **do not impact** the dataset acquisition would have caused an exponential increase in the time taken to develop each feature set and space taken to store the checkpoints. The **omission** of variation attributes which **do impact** the dataset acquisition would have meant that the results obtained were incorrect and did not represent correctly the variation defined and used.

For example, if we had not inserted `Repetition XX` into the dataset code, all of the repetitions would reuse the same dataset, meaning that they would all be identical. Since we wished to study solutions' confidence level, nothing should be shared between variations to make sure that the solution was consistent due to the techniques employed and not due to caches or temporary files generated.

Lastly, take as an example the `Feature Importance`. If we had selected this attribute, then a new checkpoint would have been created for each feature under study. However, since in order to study feature importance, we simply randomized the values for the columns, it means that all variations take as starting point a common dataset. We concluded, that there was no need to checkpoint each and every randomized feature set, and could instead checkpoint the common dataset, and then randomize the column's values.

6

Corpus and Feature Sets Analysis

Contents

This chapter provides an in-depth analysis of the data acquired throughout our study. We start by describing the acquisition of the corpus and the corpus itself in section 6.1. In section 6.2 we analyze the feature sets acquired and comment on possible details that might influence the models developed from these feature sets.

6.1 Corpus Recordings and Analysis

An integral part of our work included the acquisition of recordings of patients from CHLO - Unidade de Saúde Mental de Oeiras. In total our team went to the clinic 36 days, recording in total 85 patients, which averages out to 2.36 recordings per day. This relatively low number further corroborates the idea that patients, especially patients diagnosed with psychosis, have a reluctance to take part in the study. Many of the patients that declined our invitation to take part in the study, suggested that they did not trust us or the protocol to safeguard their best interests.

Although 85 recordings were carried out at the clinic, the corpus was extended with 80 patients. This six patients' difference is due to: (1) one patient wished for his recording to be deleted, (2) three patients diagnosed with psychosis had already been recorded by Forjó et al. [5], and (3) two patients with bipolar disorder were Brazilian and did not speak European Portuguese.

The final extension of the corpus, with subjects already filtered out, either by choice or to maintain corpus consistency, is shown in figure 6.1. The original version of the corpus was made up of 92 subjects: (i) 56 healthy-control subjects, and (ii) 36 subjects diagnosed with psychosis. Our expansion allowed us to increment the *First European Portuguese Corpus for Psychosis Identification* with: (i) 9 healthy subjects, part of the control group, and (ii) 49 subjects diagnosed with psychosis, part of the target group, and (iii) 31 subjects diagnosed with bipolar disorder, part of the control group.

After our expansion, the final corpus achieved is composed of: (i) 65 healthy subjects, with no prior mental disorders, part of the control group, (ii) 85 subjects diagnosed with psychosis, that make up the target group, (iii) and 31 subjects diagnosed with bipolar disorder, part of the control group.

One of our main objectives was on approximating the age and schooling distributions between groups whilst recording more subjects. As noticeable from figures 6.2(a) and 6.3(a), the distribution between group types is considerable. The healthy group is considerably younger and more schooled. The target group has distributions that seem more natural and representative of the overall population, the problem being the healthy group which is biased toward a more specific academic and young population. These high contrasts between groups in socio-demographic attributes may impact the models developed and consequently the results obtained through these models.

The age and schooling distributions on the extended version of the corpus are visible in figures 6.2(b) and 6.3(b) respectively. The overall improvement in these two attributes is visible in figures 6.2

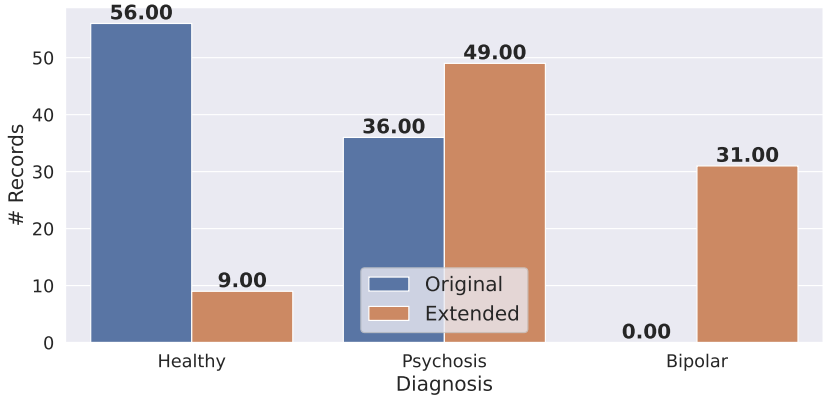


Figure 6.1: Corpus original size against our expansion of the corpus organized according to subjects' group.

and 6.3. Although the problem discussed remains, the improvements are noticeable, especially in age distribution. The age distribution of the control group is much more similar to the distribution of our target group and mirrors a normal distribution.

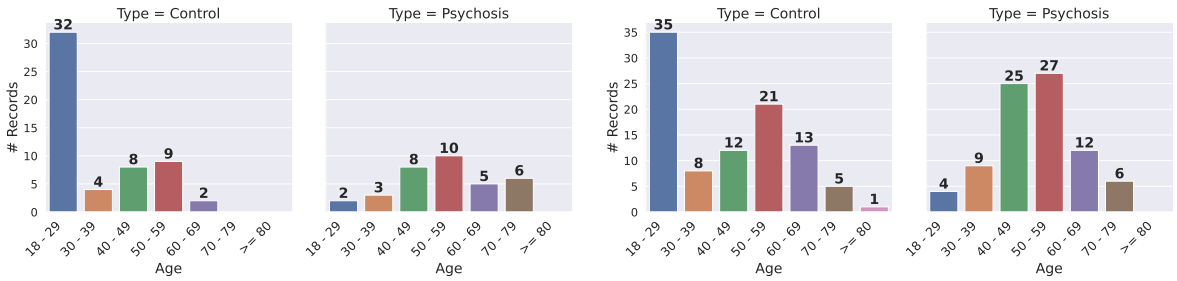


Figure 6.2: (a) Initial and **(b)** extended versions of the corpus age distribution, on top and bottom respectively.

Although our team recorded healthy controls that enhanced the mentioned distributions, the improvements reported are in the vast majority due to the recording of patients diagnosed with bipolar disorder. These patients are part of the control group, however, these subjects were recruited in the same location as patients diagnosed with psychosis. Consequently, it was to be expected that these two groups are much more similar in their socio-demographic distributions.

We recorded other information and characteristics from subjects, such as their gender and the usage of masks. Any of this information can impact discourse and must be disclosed and discussed to be sure that any of the effects on discourse, and features extracted from it, are accounted for.

Regarding **gender**: (i) out of the 65 healthy-controls, 37 are male and 28 are female, (ii) out of the 85 patients diagnosed with psychosis, 32 are male and 53 are female, and (iii) and out of the 31 patients diagnosed with bipolar disorder, 23 were male and 8 are female.

Regarding the **usage of masks**, note that this information was not registered for subjects on the

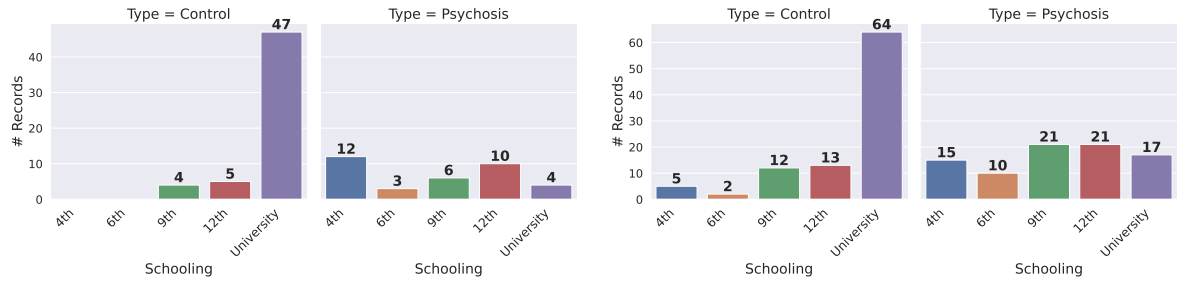


Figure 6.3: (a) Initial and **(b)** extended versions of the corpus schooling distribution, on top and bottom respectively.

original version of the corpus. At the time the usage of masks was mandatory at hospitals and health clinics in Portugal and advised in closed indoor spaces, so: (i) out of the 9 healthy-controls recorded by us, 2 used a mask and 7 did not, (ii) out of the 49 patients diagnosed with psychosis and recorded by us, 1 we did not register the information, 45 used a mask, and 3 did not, (iii) and out of the 31 patients diagnosed with bipolar disorder and recorded by us, 30 used a mask and 1 did not.

Finally, we also wanted to analyze the impact that our manually fixing the generated transcriptions had on the transcriptions themselves. Evaluate how much did we change the original transcriptions. To this purpose, we computed the similarity between the various transcriptions, using the edit-distance formula [58] applied to the entire transcription: (i) `Tr` transcription, automatically generated from the *XYH-6 Stereo Microphone Capsule* recording, (ii) `LR` transcription, automatically generated from the *Omnidirectional XLR Lavalier Microphone* recording, (iii) `Fix` transcription, manually generated from fixing the `LR` automatically generated transcription.

These similarities, which range from 0 to 1, with 0 being no similarity at all and 1 being identical, were then plotted through a violin and swarm plot, shown in figure 6.4. One of our concerns while fixing the transcription was that we would be changing the original transcriptions too much. Our purpose was only to fix major and meaningful words that had been misinterpreted by the transcriber, and that could significantly alter the content analysis. We did not aim to create a manual transcription of the various recordings, since this would require specific expertise. From figure 6.4 we can conclude that the changes made by our team were small since the difference between the transcriptions generated for the two microphones (Tr/LR) is bigger than the difference between the manual transcription and the transcription used as the baseline for this manual correction (Tr/Fix).

6.2 Feature Sets Analysis

In this section, we provide an overview analysis of the feature sets developed to train our model, to better understand the results obtained and exhibited. This analysis was done through tables that display

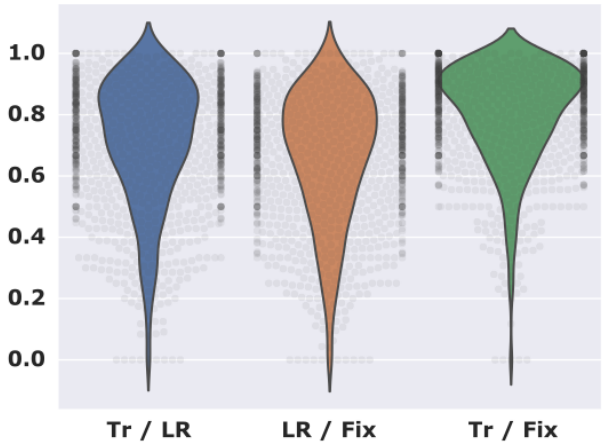


Figure 6.4: Violin and swarm plots display the similarity between the various transcriptions.

descriptive statistics of the various features extracted. Note that, to simplify this analysis, we only detail feature sets in their entirety, without any restriction in terms of features or subjects, using the manually fixed transcriptions. Multiple variations of these feature sets are developed to study the results for specific characteristics of feature sets.

The tables displayed in appendix E provide a statistical analysis of the features that make up the structure/coherence and content feature sets.

7

Results and their Discussion

Contents

The results obtained throughout the various studies are displayed in the form of tables to facilitate interpretation by the reader and comparison across studies. In section 7.1 we display and discuss the results obtained from the first study, a study in which we aim to test as many variations as possible to achieve a local maximum for the models developed. Section 7.2 focuses on the assessment classification studies, in which, our goal was to understand the confidence that can be established in the models developed and a metric for the importance of the various features used in the development of our models.

7.1 Exploration Classification Task

On this exploration classification task, we aimed at understanding the best hyperparameter variation for which to further develop and study our models. As discussed before, this stage also allowed us to understand the impact that the expansion of the corpus had on the results achieved. Lastly, it allowed us to discern whether the manual processing and correction of the automatic transcriptions had a direct effect on the results achieved.

7.1.1 Sound and Speech Feature Set

From figure 7.1, we observed a decrease in the results obtained with the increase in corpus size and complexity. The results displayed in this graph are for the best model, according to UAR, for each task and each possible target data group.

The fact that in every task, the achieved score decreases when comparing `V1 - Simple` and `V2 - Simple`, leads us to the conclusion, that the model developed from `V1 - Simple` overfitted to the small amount of data present since in the presence of more data the models were less successful. Nonetheless, the results obtained with this feature set for `V2 - Simple` are still within the acceptable range, classifying most of the subjects correctly.

We also verify a decrease in the results obtained when comparing `V2 - Simple` with `V2 - Complex`. This decrease can now be justified by two dependent factors. First, by adding more subjects to the corpus, it is to be anticipated that there would be a decrease in the scores obtained, as mentioned previously. Secondly, there is a factor of added complexity, by adding subjects with different diagnoses into the control group.

Note that an increase in the size or complexity of the corpus, used for the development of the models, does not immediately and inevitably cause a decrease in the results obtained. Feature extraction techniques and classification techniques, when capable, can adjust and maintain, and even improve, the results obtained. This decrease only reveals that the techniques used either reached their limits or

should be further explored, by varying parameters or other characteristics that might have an impact on the results.

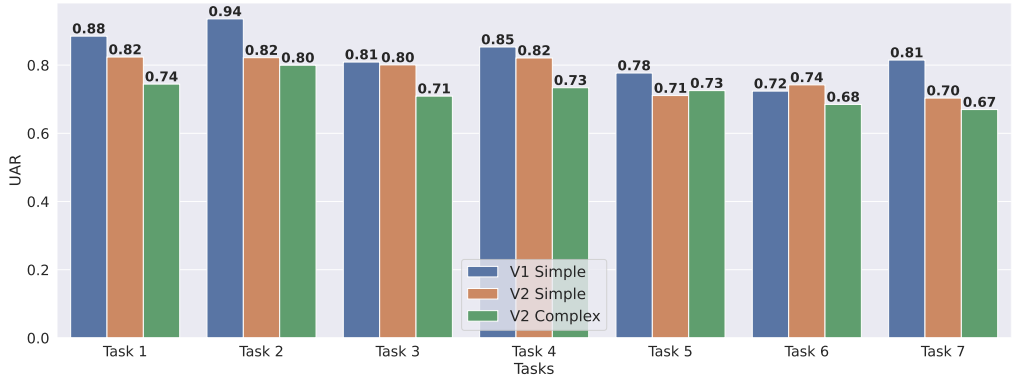


Figure 7.1: Bar graph with the results for the best models using sound + speech feature set, according to UAR, comparing across the various data variations per task, with manual transcriptions.

In the following table 7.1 we display the results for the best variation for each target data and target task across all registered metrics. The same table also displays the classification technique and hyperparameters that allowed us to reach those particular results.

Almost all of the best variations used RF as the classification technique. This technique provides robustness and flexibility that some of the remaining and more traditional techniques do not provide. Interestingly enough, two variations achieved their best score with DT. Still, both of these variations used `V1 - Simple` and for the remaining data groups for the same task RF was selected. The diminished complexity of `V1 - Simple` when compared with `V2 - Simple` or `V2 - Complex` might have allowed for the usage of less complex classification techniques.

Finally, figure 7.2 displays the best UAR scores with and without manual correction of the transcription, for each task. The impact of such correction is minimal, which is expected especially on the current feature set. Most of the features extracted rely on the audio files and not on the transcriptions generated for each subject-task pair. Nonetheless, the manual correction changed the obtained results even if only by a small amount. This impact signaled us, that even for features as simple as `Number of Words`, `Speaking Rate`, and `Articulation Rate` the process of manually fixing transcriptions might be beneficial as it means a more accurate corpus and consequently results.

7.1.2 Structure, Coherence, and Content Feature Set

In this subsection, we followed the same strategy for the analysis of structure, coherence, and content feature set, as the one followed in subsection 7.1.1. We started by analyzing the impact that the different target data groups had on the results obtained. Figure 7.3 displays the best variations, according to UAR, using transcriptions manually corrected for each target data group and task.

Table 7.1: Best models developed for the initial classification task, according to the UAR, with the speech and sound feature set, and manually fixed transcriptions. Hyper-parameters specified in the same order as defined for the classifiers in section 5.4.2.A. The best and worst score for each metric is signaled in bold.

Task	Data Variation	Classifier	Classifier Hyper-parameters	Accuracy	Recall	F1-Measure	UAR
1	V1 - Simple	RF	(10, gini, 1, None, 0)	0.900	0.824	0.862	0.885
	V2 - Simple	RF	(50, entropy, 16, auto, 0)	0.830	0.878	0.852	0.824
	V2 - Complex	RF	(150, gini, 16, None, 0)	0.747	0.707	0.720	0.744
2	V1 - Simple	RF	(10, gini, 16, None, 0)	**0.946**	0.889	**0.928**	**0.936**
	V2 - Simple	RF	(150, gini, 128, None, 0)	0.827	0.859	0.849	0.822
	V2 - Complex	RF	(50, gini, 64, auto, 0)	0.801	0.776	0.786	0.800
3	V1 - Simple	RF	(50, gini, 4, auto, 0)	0.824	0.743	0.765	0.809
	V2 - Simple	RF	(100, entropy, 8, auto, 0)	0.810	0.878	0.837	0.801
	V2 - Complex	RF	(50, gini, 16, log2, 0)	0.712	0.671	0.683	0.709
4	V1 - Simple	DT	(gini, 16, None, 0)	0.885	0.742	0.821	0.853
	V2 - Simple	RF	(50, gini, 128, log2, 0)	0.824	0.857	0.841	0.821
	V2 - Complex	RF	(100, gini, 64, auto, 0)	0.737	0.701	0.706	0.734
5	V1 - Simple	RF	(10, entropy, 2, None, 0)	0.783	0.750	0.730	0.777
	V2 - Simple	RF	(50, gini, 32, sqrt, 0)	0.730	0.867	0.783	0.711
	V2 - Complex	RF	(150, entropy, 8, None, 0)	0.726	0.711	0.707	0.725
6	V1 - Simple	RF	(10, entropy, 32, None, 0)	0.747	**0.611**	0.657	0.724
	V2 - Simple	RF	(50, gini, 16, sqrt, 0)	0.755	0.843	0.795	0.742
	V2 - Complex	RF	(150, entropy, 64, auto, 0)	0.685	0.675	0.667	0.685
7	V1 - Simple	DT	(entropy, 2, None, 0.1)	0.789	**0.944**	0.782	0.815
	V2 - Simple	RF	(100, gini, 32, log2, 0)	0.719	0.819	0.768	0.703
	V2 - Complex	RF	(50, gini, 64, log2, 0)	**0.672**	0.627	**0.642**	**0.670**

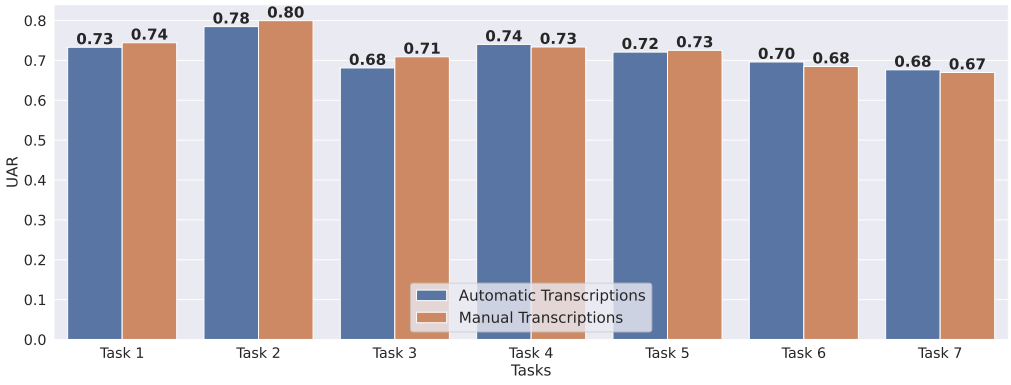


Figure 7.2: Bar graph with the results for the best models using sound + speech feature set, according to UAR, comparing automatic and manual transcriptions per task, with V2 - Complex data variation.

We verify that the increasing complexity of the data targets had a negative influence on the results obtained. With the added complexity, the feature extraction techniques seem to classify the various subjects less well. At least this is the general tendency of results obtained, *Task 1* seems to be the exception, where best results were achieved for `V2 - Simple` than for `V1 - Simple`. However, this improvement was only of 0.02 which can be considered negligible.

The reasoning behind this decrease is the same as for the sound and speech feature set. The decrease from `V1 - Simple` to `V2 - Simple` is justified by the increased complexity of having more subjects and consequently a more diverse corpus, even if it has the same groups (healthy controls and patients diagnosed with psychosis). The general decrease from `V2 - Simple` to `V2 - Complex`, is possibly justified by two different factors: the added complexity of having more subjects, and a corpus more diversified by including subjects with other diagnoses other than healthy or psychosis.

Note that, once again, techniques for feature extraction and classification, if powerful enough and properly applied, can surpass this escalation in complexity and adapt to bigger and more complex corpora.

Figure 7.3: Bar graph with the results for the best models using structure/coherence + content feature set, according to UAR, comparing across the various data variations per task, with manual transcriptions.

The best models, measured through UAR, for every task and data target, are displayed in table 7.2. This table presents the results for all recorded metrics, exclusively using the manually corrected transcriptions.

From this table, we understand that the RF classifier achieved the best scores across the various data and task targets. However, for this feature set, we verify that more of the variations deviate from this general tendency, with some reporting the best results with MLP, SVM, and even DT classifier. Interestingly, simpler classification techniques, such as SVM and DT reported better results than more complex techniques such as MLP and RF for more complex data variations such as `V2 - Complex` and `V2 - Simple`. This effect was verified on `Task 3` for both `V2 - Simple` and `V2 - Complex` and on `Task 5` for `V2 - Simple`.

Our team can only hypothesize possible justifications for this occurrence. We believe that the feature extraction techniques employed generated values that map more directly to the target groups, and that consequently less-complex classification models are required for the correct classification of subjects.

Table 7.2: Best models developed for the initial classification task, according to the UAR, with the structure, coherence, and content feature set, and manually fixed transcriptions. Hyper-parameters specified in the same order as defined for the classifiers on section 5.4.2.A.

Task	Data Variation	Classifier	Classifier Hyper-parameters	Accuracy	Recall	F1-Measure	UAR
1	V1 - Simple	RF	(10, entropy, 32, log2, 0)	0.789	0.588	0.678	0.749
	V2 - Simple	RF	(10, entropy, 64, None, 0)	0.776	0.805	0.800	0.772
	V2 - Complex	RF	(150, gini, 4, auto, 0)	0.691	0.659	0.663	0.689
2	V1 - Simple	RF	(50, gini, 64, log2, 0)	**0.880**	0.806	**0.841**	**0.867**
	V2 - Simple	RF	(150, gini, 8, None, 0)	0.780	0.812	0.807	0.775
	V2 - Complex	RF	(100, gini, 1, None, 0)	0.751	0.682	0.720	0.747
3	V1 - Simple	RF	(50, gini, 16, auto, 0)	0.802	0.714	0.735	0.786
	V2 - Simple	DT	(entropy, 2, log2, 0)	0.735	0.744	0.758	0.733
	V2 - Complex	SVM	(0.5, linear)	0.661	0.561	0.605	0.654
4	V1 - Simple	MLP	((50,), relu, 0.025, invscaling, 500)	0.747	0.710	0.667	0.739
	V2 - Simple	RF	(50, entropy, 2, None, 0)	0.690	0.714	0.714	0.688
	V2 - Complex	RF	(150, entropy, 32, auto, 0)	0.667	0.584	0.612	0.659
5	V1 - Simple	RF	(100, gini, 8, None, 0)	0.783	0.583	0.677	0.747
	V2 - Simple	DT	(entropy, 1, None, 0)	0.689	**0.855**	0.755	0.666
	V2 - Complex	RF	(50, entropy, 128, None, 0)	**0.637**	0.554	**0.586**	**0.631**
6	V1 - Simple	SVM	(2, sigmoid)	0.857	0.778	0.812	0.843
	V2 - Simple	RF	(50, entropy, 4, log2, 0)	0.728	0.807	0.770	0.716
	V2 - Complex	RF	(10, entropy, 4, log2, 0)	0.713	0.711	0.698	0.713
7	V1 - Simple	RF	(10, entropy, 32, log2, 0)	0.733	0.556	0.625	0.704
	V2 - Simple	RF	(50, gini, 4, None, 0)	0.671	0.807	0.736	0.650
	V2 - Complex	RF	(10, gini, 16, sqrt, 0)	0.689	**0.482**	0.593	0.677

Regarding the difference between the results achieved with and without manually corrected transcription for the structure/coherence and content feature set, we expected results to be impacted by this difference. As shown in figure 7.4, this preprocessing stage seems to have affected the results more on this feature set than on the speech and sound feature set. This was expected since this feature set relies heavily on the subjects' transcriptions. Interestingly, this variation was not consistent across the various tasks, with some reporting better results with the transcription corrections and others without. This variation appears to be more noticeable on tasks 5, 6, and 7. This might be because, in these tasks, subjects spoke more freely but in a structured manner. In tasks 1 and 2, subjects were forced to stay on topic and the discourse was unstructured since it relies on subjects enumerating words. In tasks 3 and 4, subjects had to structure their discourse but they were still forced to follow a specific topic and

task, not allowing for as much free expression.

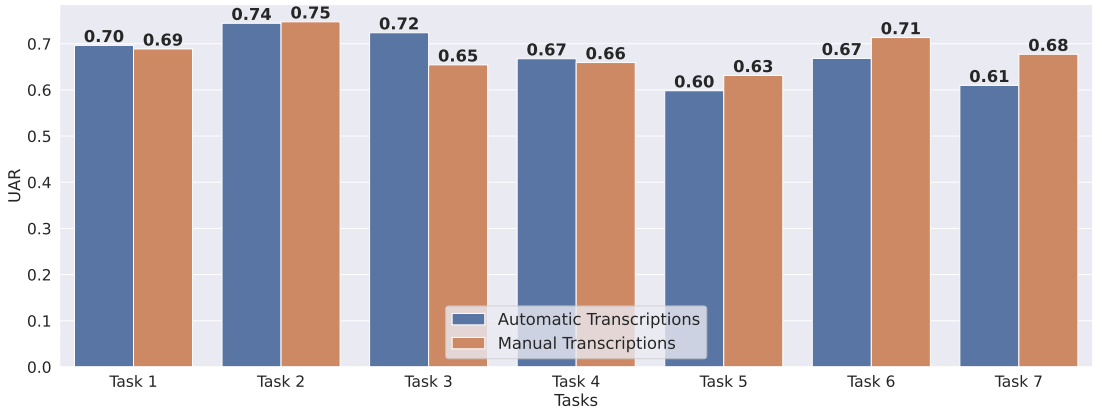


Figure 7.4: Bar graph with the results for the best models using structure/coherence + content feature set, according to UAR, comparing across automatic and manual transcriptions per task, with V2 - Complex data variation.

7.1.3 Discussion

Both feature sets seem to achieve lower scores with added data complexity and to be impacted by the manual correction of the various transcription.

Still, we believe it is important to directly compare the results obtained between feature sets, so as to better understand their capabilities.

Figure 7.5 directly compares the results obtained for both feature sets across the various tasks. This figure displays the UAR scores achieved with the manually corrected transcription with `V2 - Complex` as the data target. From this figure, we understand that *sound and speech features* achieved better results than *structure/coherence and content features*. Note, however, that this difference is small, on average the difference was of 0.0414 ($SD = 0.0422$). The biggest difference was felt on task 5, with *coherence/structure and content features* achieving results 0.10 under those of baseline, described in subsection 7.1.1. Interestingly, in tasks 6 and 7 coherence/structure and content feature sets achieved better results than those of baseline. This was unexpected since tasks 5, 6, and 7 are very similar in purpose.

Nonetheless, we can hypothesize a reason for this difference between task 5 and tasks 6 and 7. We believe that on tasks 6 and 7, content features might have aided the classification process by picking up speech apathy from patients diagnosed with psychosis, a symptom which has been extensively described in literature [10, 19, 59]. We gathered some confidence in this hypothesis by analyzing the results from the assessment classification task in section 7.2. This will be possible since in this section we will acquire and discuss scores that express feature importance on a task level.

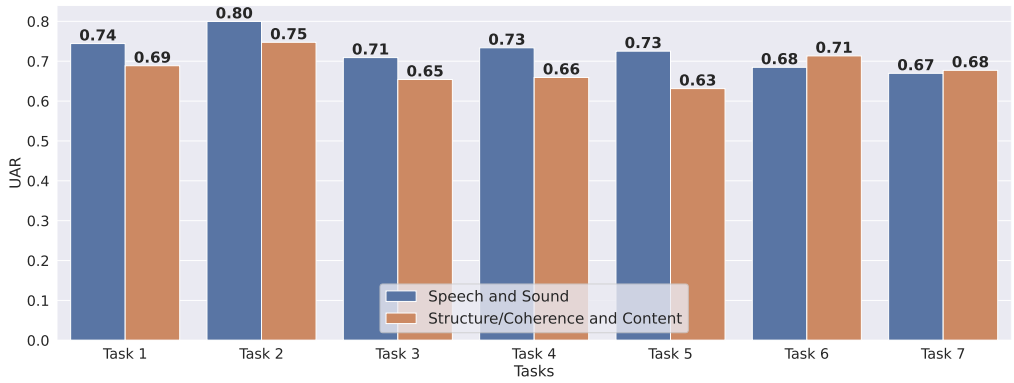


Figure 7.5: Bar graph with the results, according to UAR, comparing the various feature sets per task, with V2 - Complex data variation and manually corrected transcriptions.

7.2 Assessment Classification Task

The assessment classification task was carried out with two main objectives: the acquisition of confidence scores for the models developed, and the analysis of the importance of the various features, that make up the structure, coherence, and content feature set, for each task.

7.2.1 Confidence Study

Starting with the confidence study, it is important to first understand that the provided analysis was done by repeating 10 times the development of the models that achieved the best scores for the structure/-coherence and content feature set with manual corrected analysis and data target of `V2 - Complex`, reported in table 7.2. The number of repetitions carried out was quite small, and in order to get a better estimate of the achieved results, further analysis should be done with more repetitions.

Interestingly, when comparing the results obtained in the previous study with the ones obtained for this study, we understand that the average UAR scores of the 10 repetitions are significantly lower than the score reported for the first study. We believe that this might have happened due to the big number of variations that were tested in the previous study, and out of which, we chose the best. By choosing the best variation to represent the competence of the feature set, it was to be expected that such value would provide an optimistic result.

Nonetheless, the confidence results obtained seem promising. The confidence interval computed for the 95% confidence level is very small independently of the task. As seen in table 7.3, the model that showed the highest variability and therefore provides less confidence in the results obtained was the one developed from task 7. Even for this task, the results obtained suggest that 95% of the time the results obtained for this model would be somewhere in between 0.538 and 0.557, giving us a margin of just 0.037. The confidence interval went as low as 0.004 in task 2.

The most important conclusion that can be extracted, is that the models developed seem to be consistent and robust and that although a significant difference exists between the results for the first study, reported in subsection 7.1.2, and this one.

Table 7.3: UAR confidence interval, represented by its lower, mean and upper bound, for each one of the various tasks. The number of repetitions used to calculate the confidence levels is presented as well as the confidence interval size for a facilitated analysis.

Task	#Repetitions	Confidence Interval			Interval Size
		Lower Bound	Mean	Upper Bound	
Task 1	10	0.639	0.659	0.649	0.020
Task 2	10	**0.738**	**0.741**	**0.740**	0.004
Task 3	10	0.647	0.667	0.657	0.021
Task 4	10	0.627	0.646	0.636	0.019
Task 5	10	0.575	0.595	0.585	0.020
Task 6	10	0.630	0.666	0.648	0.036
Task 7	10	0.538	0.575	0.557	0.037

Task 2 achieved, by far, the best results, with most of the remaining tasks being equally good between each other, with exception of task 5 and task 7, for which the results obtained seem to be less promising. Further conclusions will be drawn by analyzing the importance of the various features on these tasks.

7.2.2 Feature Importance

For an easier interpretation and analysis of the results obtained we provide, through figure 7.6, the relative importance of every feature for every task. This importance is relative since it is relative to the average established in subsection 7.2.1 for each task.

For the **first task**, it seems that almost no noticeable differences in features' importance are verified. The number of repeated edges in the word graph, which directly maps to the number of repetitions of identical sequences of words by the subject, seems to be the most relevant feature. Task 1 relies on the subject enumerating as many words as possible during a specified time interval. Consequently, we conclude that the model developed for task 1 fitted to the number of repetitions of sequences of words by the subjects during the enumeration. This conclusion reflects what is mentioned in literature [14, 18], that the discourse of patients diagnosed with psychosis is marked by repetitions, and we believe that this effect is worsened by the specific task being carried out.

In **task 2** there is a feature that clearly prevails over all the remaining features. The model developed for this particular task has fitted deeply to the number of nodes present in the word graph developed from subjects' transcriptions. The number of nodes in the word graph maps directly to the number of unique words part of the subjects' transcriptions. Remember that task 2 also relies on enumeration

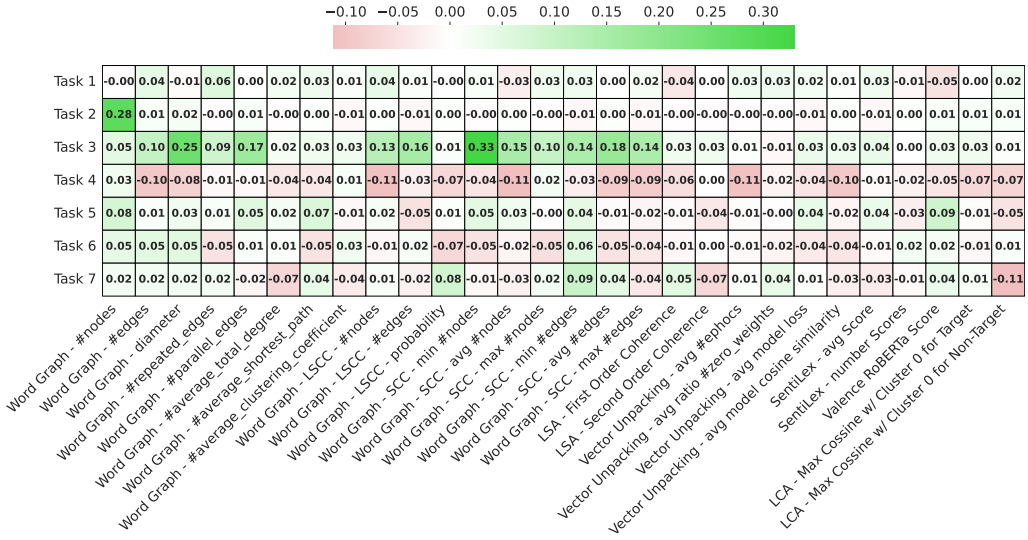


Figure 7.6: Relative feature importance per feature and task plotted into a heatmap. Positive/green values represent important features for which the results obtained decreased when the values were randomized. Negative/red values represent non-important features for which the results obtained increase when values are randomized.

by the subjects, and consequently, we are measuring the number of different words uttered by the subject. Patients diagnosed with psychosis are described as having a discourse with low complexity and a reduced vocabulary. Still, we do not understand why for task 2 this feature seems to be decisive, whereas for task 1 it was not.

The results obtained for the model developed on **task 3**, suggests that the model has fitted in general to a big number of word graph features, and therefore to the subject's discourse structure. This was surprising for our team since this task relies on the subject reading out loud a well-known story. We believe that since the subjects are reading a story, very little can be inferred from its transcriptions, and much less from differences in topic or sentiment. Consequently, the model fitted to the set of features that showed some dissimilarities. During the recordings, we observed that even with the story being provided, patients seemed to be inaccurate during their reading and that many times they repeated certain excerpts after making a mistake. This happened even if the mistake had happened a decent number of words ago.

Not much can be derived from the results for **task 4** since most of the features report almost the same importance. This was unexpected since in this task the subjects can speak more freely, we would have expected the models developed to have fitted to particular differences in subjects' discourse such as its structure. We believe that differences in subjects' discourse structure might have been suppressed by the fact that this task highly depended on the subject's knowledge of the story, which for some was natural and for others new. Furthermore, take as an example disorganized speech, a patient which is familiar with the story might behave fairly well in terms of structure, and a control subject who is just

reading the story for the first time might display disorganized speech due to reduced knowledge that it has from the story.

Our team noticed that content features for sentiment analysis seem to be more relevant on **tasks 5, 6, and 7**, than on the remaining tasks, especially the feature extracted from the fine-tuned RoBERTa model. This was expected since these three tasks rely a lot more on the personal analysis of the subjects and on their perception of the world. Patients diagnosed with psychosis are described in literature as displaying a lack of empathy and instead apathy, and sometimes displaying a more pessimistic outlook [10, 19, 59]. This hypothesis was verified by researching the profiling files for the models developed and checking that, for patients diagnosed with psychosis, low valence scores are more frequently associated with patients with psychosis than with the control group.

On **task 7**, maybe due to the nature of the image displayed to the subjects, more features seem to reveal being important. In this task, besides the features extracted from word graphs, features extracted through vector unpacking and LSA seem promising. Similarly, complex word graph features that seem almost irrelevant for the remaining tasks seem to provide important information regarding the subjects' execution of this particular task, such as `Word Graph - LSCC - probability`. We can only hypothesize that these features might have revealed themselves to be important due to some disruption that the image might have created on the subjects during the execution of the task.

Additionally, we also recognized that across all the techniques applied, features acquired from word graphs seem to be the best suited across all tasks for the acquisition of accurate and robust models. We believe that this effect comes as a consequence of two different factors: (i) first, this technique provides a wide range of features, that describe discourse, from the most basic (in which the number of words spoken is extracted) to the most complex (in which the probability of a specific sequence of words happening is measured); (ii) secondly, some of the remaining features might be undeveloped for the task at hand. As an example, more accurate and complex models for the assessment of the sentiment of a subject's discourse might have decreased the relevance of word graph features.

Finally, as an overall analysis of the heatmap displayed in figure 7.6, we can conclude that most of the features extracted are valuable for the classification task or at the very least neutral for this classification task. This is immediately noticeable through the heatmap since most of the cells are marked with white or a very pale shade of red, with the majority of the exceptions having a strong green association. The fact is that strong reds are almost nonexistent.

8

Conclusions and Future Work

Contents

Our results support our conclusion that high-level features that focus on speech structure, coherence and content are promising in terms of their capabilities on classifying subjects diagnosed with psychosis. The results obtained with our developed feature set seem to achieve results comparable to those of the established baseline. Furthermore, due to the confidence that they provide and the importance of the various features that make up this feature set, they seem promising and worthy of continuing research.

We understood that with an increase in the size or complexity of the corpus, worse results were achieved. This was expected but verified our hypothesis that initial models had overfitted to the data provided, both to the small size and the lack of variability in the control group. Although this caused a decrease in the metrics recorded, we acknowledge that this complexity should be pursued since it allows for a more accurate representation of the real world and consequently more meaningful results.

Similarly, we also recognized that the manual correction of the automatically generated transcriptions is important for an accurate representation of the subjects' task executions, even if the impact generated on the results achieved was negligible. Interestingly, this manual correction of the transcriptions impacted the results obtained for both feature sets.

By directly comparing the results obtained for each one of the feature sets, we concluded that the results achieved for our focus feature set (structure, coherence, and content features) are comparable to those of the established baseline. However, for most tasks, structure, coherence, and content feature sets, achieved slightly worse results, except for tasks 6 and 7. On average, by using feature extraction techniques that focus on discourse's structure, coherence, and content, our results decreased 4.14% $(SD = 4.22\%)$. Particularly, on tasks 6 and 7, the targeted feature set achieved on average an increase of 2.00% $(SD = 1.00\%)$

Our study on the confidence of the models developed led us to the conclusion that our initial study might have been too optimistic about the results obtained since we focused on the variations that originated the best results and not on an average from the various variations. Nonetheless, it verified, even if for a small set of repetitions that our models are robust, providing a very small confidence interval ($Mean = 2.24\%$, $SD = 1.04\%$) for a confidence level of 95%.

Last but not least, our study of the features' importance revealed to be crucial for a good and complete understanding of the most meaningful features for the models developed for each task. We understood that word graph features seem to be the most reliable and essential across the various models developed for each task. Sentiment-based techniques appear to have more significance on tasks 5, 6, and 7, in which we, purposely, try to evoke sentiments and personal perceptions of the world from subjects. However, this importance is still fairly weak, possibly due to the simplicity of the models for sentiment analysis that were applied. Finally, from an overall perspective of the results obtained we understood that there are features that are particularly useful for certain tasks. However, no feature suggested that its removal would cause a significant increased in the results obtained.

In summary, our results serve as support for future developments in the domain of computerized solutions to aid mental health diagnosis. We believe that we achieved our goals, by obtaining promising results with structure, coherence, and content features, in which we analyzed the impact of the expansion of the corpus, the manual correction of the transcriptions, and the importance of the various features. Additionally, we also succeeded by submitting a paper, which was accepted for publication, with part of our work, to a well-known conference that focuses on speech and language technologies, *IberSpeech*.

8.1 Limitations and Future Work

During the execution of our work, our team faced various limitations, mostly due to underdeveloped NLP domains for European Portuguese and time constraints. The existence of better models for '*sentence segmentation*', '*lemmatization*', '*transcription*', '*valence*' and '*sentiment analysis*' of text would have impacted greatly our work. Their existence would have likely improved the results obtained and saved time that was allocated to developing some of our models. With this time we could have explored more feature extraction techniques that we did not have the opportunity to explore, such as LCB, SRL, and dependency parsing, described in section 2.2. If we had more time we could have improved on current studies and possibly carried out further studies. As for the current studies, we could have studied in more detail the confidence of the models developed by carrying out more repetitions.

Regarding future work, our studies provided an analysis that allowed for the understanding of the impact that adding patients with bipolar disorder had on the results. However, it did not provide any insights into how well the models differentiated patients diagnosed with psychosis from patients diagnosed with bipolar disorder. It would have been interesting to know exactly how many patients diagnosed with bipolar disorder were misidentified as diagnosed with psychosis. Furthermore, an in-depth analysis of these subjects specifically would provide an understanding of the downfalls of the models developed.

A study focused on the merging of sound and speech features with structure, coherence and content features could also reveal interesting results. We concluded that these feature sets by themselves achieve good results, by joining both, at the very least we expect results to be more consistent and models to be more robust. Possibly the results would improve by this merge.

Finally, improvements in the techniques applied for structure, coherence, and especially content feature extraction could reveal major improvements in the results obtained. For instance, we followed the standard technique of achieving embeddings through LSA for the computation of coherence values for the subjects on the various tasks. However, other, more recent, and complex techniques, could be applied for the acquisition of word or sentence embeddings, such as *GLoVE* embeddings [10, 30]. Ideally, we also would have liked to have explored more techniques for the content analysis of speech. For example the application of complex transformers, such as GPT-3, for the acquisition of sentiment valence scores, general sentiment analysis, and LCB.

Bibliography

[1] S. L. James, D. Abate, K. H. Abate, S. M. Abay, C. Abbafati, N. Abbasi, H. Abbastabar, F. Abd-Allah, J. Abdela, A. Abdelalim *et al.*, "Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017," *The Lancet*, vol. 392, no. 10159, pp. 1789–1858, 2018.

[2] R. Coentre and P. Levy, "Early intervention in psychosis in portugal: Where are we?" *Acta medica portuguesa*, vol. 33, pp. 540–542, 9 2020.

[3] T. R. Insel, "Rethinking schizophrenia," *Nature*, vol. 468, pp. 187–193, 2010. [Online]. Available: https://doi.org/10.1038/nature09552

[4] E. J. Clemmer, "Psycholinguistic aspects of pauses and temporal patterns in schizophrenic speech," *Journal of Psycholinguistic Research*, vol. 9, pp. 161–185, 1980. [Online]. Available: https://doi.org/10.1007/BF01067469

[5] M. Forjó, H. S. Pinto, and A. Abad, "Contributions towards the possible identification of psychosis through speech processing in portuguese," Master's thesis, Instituto Superior Técnico - Universidade de Lisboa, 2021.

[6] B. Elvevåg, P. W. Foltz, M. Rosenstein, and L. E. Delisi, "An automated method to analyze language use in patients with schizophrenia and their first-degree relatives," *Journal of neurolinguistics*, vol. 23, pp. 270–284, 5 2010. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/20383310https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2850213/

[7] D. Iter, J. Yoon, and D. Jurafsky, "Automatic detection of incoherent speech for diagnosing schizophrenia," in *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, 2018, pp. 136–146.

[8] N. B. Mota, N. A. P. Vasconcelos, N. Lemos, A. C. Pieretti, O. Kinouchi, G. A. Cecchi, M. Copelli, and S. Ribeiro, "Speech graphs provide a quantitative measure of thought

disorder in psychosis," *PloS one*, vol. 7, pp. e34 928–e34 928, 2012. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/22506057https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3322168/

[9] K. McManus, E. K. Mallory, R. L. Goldfeder, W. A. Haynes, and J. D. Tatum, "Mining twitter data to improve detection of schizophrenia," *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, vol. 2015, pp. 122–126, 3 2015. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/26306253https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4525233/

[10] E. S. Kayi, M. Diab, L. Pauselli, M. Compton, and G. Coppersmith, "Predictive linguistic features of schizophrenia," *arXiv preprint arXiv:1810.09377*, 2018.

[11] G. Bedi, F. Carrillo, G. A. Cecchi, D. F. Slezak, M. Sigman, N. B. Mota, S. Ribeiro, D. C. Javitt, M. Copelli, and C. M. Corcoran, "Automated analysis of free speech predicts psychosis onset in high-risk youths," *npj Schizophrenia*, vol. 1, no. 1, pp. 1–7, 2015.

[12] N. B. Mota, M. Copelli, and S. Ribeiro, "Thought disorder measured as random speech structure classifies negative symptoms and schizophrenia diagnosis 6 months in advance," *npj Schizophrenia*, vol. 3, p. 18, 2017. [Online]. Available: https://doi.org/10.1038/s41537-017-0019-3

[13] N. Rezaii, E. Walker, and P. Wolff, "A machine learning approach to predicting psychosis using semantic density and latent content analysis," *npj Schizophrenia*, vol. 5, p. 9, 2019. [Online]. Available: https://doi.org/10.1038/s41537-019-0077-9

[14] G. R. Kuperberg, P. K. McGuire, and A. S. David, "Reduced sensitivity to linguistic context in schizophrenic thought disorder: evidence from on-line monitoring for words in linguistically anomalous sentences." *Journal of abnormal psychology*, vol. 107, p. 423, 1998.

[15] C. M. Corcoran, F. Carrillo, D. Fernández-Slezak, G. Bedi, C. Klim, D. C. Javitt, C. E. Bearden, and G. A. Cecchi, "Prediction of psychosis across protocols and risk cohorts using automated language analysis," *World psychiatry : official journal of the World Psychiatric Association (WPA)*, vol. 17, pp. 67–75, 2 2018. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/29352548https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5775133/

[16] S. Just, E. Haegert, N. Kořánová, A.-L. Bröcker, I. Nenchev, J. Funcke, C. Montag, and M. Stede, "Coherence models in schizophrenia," in *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, 2019, pp. 126–136.

[17] T. J. Spencer, B. Thompson, D. Oliver, K. Diederen, A. Demjaha, S. Weinstein, S. E. Morgan, F. Day, L. Valmaggia, G. Rutigliano, A. D. Micheli, N. B. Mota, P. Fusar-Poli, and P. McGuire, "Lower speech connectedness linked to incidence of psychosis in people at

clinical high risk," *Schizophrenia Research*, vol. 228, pp. 493–501, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0920996420304588

[18] M. Alpert, A. Kotsaftis, and E. R. Pouget, "Speech fluency and schizophrenic negative signs," *Schizophrenia Bulletin*, vol. 23, pp. 171–177, 1997.

[19] G. Gosztolya, A. Bagi, S. Szalóki, I. Szendi, and I. Hoffmann, "Making a Distinction Between Schizophrenia and Bipolar Disorder Based on Temporal Parameters in Spontaneous Speech," in *Proc. Interspeech 2020*, 2020, pp. 4566–4570. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-0049

[20] N. B. Mota, R. Furtado, P. P. C. Maia, M. Copelli, and S. Ribeiro, "Graph analysis of dream reports is especially informative about psychosis," *Scientific Reports*, vol. 4, p. 3691, 2014. [Online]. Available: https://doi.org/10.1038/srep03691

[21] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "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, 2015.

[22] A. Pompili, R. Solera-Ureña, 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.

[23] J. H. Barnett, U. Werners, S. M. Secher, K. E. Hill, R. Brazil, K. Masson, D. E. Pernet, J. B. Kirkbride, G. K. Murray, E. T. Bullmore *et al.*, "Substance use in a population-based clinic sample of people with first-episode psychosis," *The British Journal of Psychiatry*, vol. 190, no. 6, pp. 515–520, 2007.

[24] P. D. Harvey, J. Lombardi, M. Leibman, M. Parrella, L. White, P. Powchik, R. C. Mohs, M. Davidson, and K. L. Davis, "Age-related differences in formal thought disorder in chronically hospitalized schizophrenic patients: a cross-sectional study across nine decades," *American Journal of Psychiatry*, vol. 154, pp. 205–210, 1997.

[25] E. Chaika, "At issue: Thought disorder or speech disorder in schizophrenia?" *Schizophrenia Bulletin*, vol. 8, pp. 587–591, 1982.

[26] V. A. Curtis, E. T. Bullmore, M. J. Brammer, I. C. Wright, S. C. R. Williams, R. G. Morris, T. S. Sharma, R. M. Murray, and P. K. McGuire, "Attenuated frontal activation during a verbal fluency task in patients with schizophrenia," *American Journal of Psychiatry*, vol. 155, pp. 1056–1063, 8 1998, doi: 10.1176/ajp.155.8.1056. [Online]. Available: https://doi.org/10.1176/ajp.155.8.1056

[27] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 1st ed. USA: Prentice Hall PTR, 2000.

[28] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis." *Discourse Processes*, vol. 25, pp. 259–284, 1998.

[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[30] R. Chandra and A. Krishna, "Covid-19 sentiment analysis via deep learning during the rise of novel cases," *PLOS ONE*, vol. 16, pp. e0 255 615–, 8 2021. [Online]. Available: https://doi.org/10.1371/journal.pone.0255615

[31] E. Vaaras, S. Ahlqvist-Björkroth, K. Drossos, and O. Räsänen, "Automatic analysis of the emotional content of speech in daylong child-centered recordings from a neonatal intensive care unit," *arXiv preprint arXiv:2106.09539*, 2021.

[32] P. Carvalho and M. J. Silva, "Sentilex-pt: Principais características e potencialidades," *Oslo Studies in Language*, vol. 7, no. 1, 2015.

[33] A. F. Anees, A. Shaikh, A. Shaikh, and S. Shaikh, "Survey paper on sentiment analysis: Techniques and challenges," *EasyChair2516-2314*, 2020.

[34] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown, "An introduction to decision tree modeling," *Journal of Chemometrics: A Journal of the Chemometrics Society*, vol. 18, no. 6, pp. 275–285, 2004.

[35] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their applications*, vol. 13, no. 4, pp. 18–28, 1998.

[36] I. Rish *et al.*, "An empirical study of the naive bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, no. 22, 2001, pp. 41–46.

[37] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[38] M.-C. Popescu, V. E. Balas, L. Perescu-Popescu, and N. Mastorakis, "Multilayer perceptron and neural networks," *WSEAS Transactions on Circuits and Systems*, vol. 8, no. 7, pp. 579–588, 2009.

[39] D. Svozil, V. Kvasnicka, and J. Pospichal, "Introduction to multi-layer feed-forward neural networks," *Chemometrics and intelligent laboratory systems*, vol. 39, no. 1, pp. 43–62, 1997.

[40] O. Irsoy and C. Cardie, “Deep recursive neural networks for compositionality in language,” in *Advances in neural information processing systems*, 2014, pp. 2096–2104.

[41] K. Hitczenko, H. Cowan, V. Mittal, and M. Goldrick, “Automated coherence measures fail to index thought disorder in individuals at risk for psychosis,” in *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*. Association for Computational Linguistics, 6 2021, pp. 129–150. [Online]. Available: https://aclanthology.org/2021.clpsych-1.16

[42] C. Carvalho and A. Abad, “TRIBUS: An end-to-end automatic speech recognition system for European Portuguese,” in *Proc. IberSPEECH 2021*, 2021, pp. 185–189.

[43] John Bent, “Data-Driven Batch Scheduling,” Ph.D. dissertation, University of Wisconsin, Madison, may 2005.

[44] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009.

[45] T. Ylonen, “Ssh–secure login connections over the internet,” in *Proceedings of the 6th USENIX Security Symposium*, vol. 37, 1996, pp. 40–52.

[46] P. A. Rocha and D. Santos, “Cetempúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa,” *quot; In Maria das Graças Volpe Nunes (ed) V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000)(Atibaia SP 19-22 de Novembro de 2000) São Paulo: ICMC/USP*, 2000.

[47] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. ” O’Reilly Media, Inc.”, 2009.

[48] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, “Stanza: A Python natural language processing toolkit for many human languages,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020. [Online]. Available: https://nlp.stanford.edu/pubs/qi2020stanza.pdf

[49] G. A. Miller and J. A. Selfridge, “Verbal context and the recall of meaningful material,” *The American journal of psychology*, vol. 63, no. 2, pp. 176–185, 1950.

[50] C. E. Shannon, “Prediction and entropy of printed english,” *Bell system technical journal*, vol. 30, no. 1, pp. 50–64, 1951.

[51] R. Rehurek and P. Sojka, “Gensim–python framework for vector space modelling,” *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, vol. 3, no. 2, 2011.

[52] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 2020, pp. 38–45.

[53] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," *arXiv preprint arXiv:1911.02116*, 2019.

[54] audEERING GmbH. (2021) openSMILE Python. [Online]. Available: https://audeering.github.io/opensmile-python/s

[55] A. Hagberg, P. Swart, and D. S Chult, "Exploring network structure, dynamics, and function using networkx," Los Alamos National Lab.(LANL), Los Alamos, NM (United States), Tech. Rep., 2008.

[56] C. Molnar, *Interpretable machine learning*. Lulu. com, 2020.

[57] A. Fisher, C. Rudin, and F. Dominici, "All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously." *J. Mach. Learn. Res.*, vol. 20, no. 177, pp. 1–81, 2019.

[58] E. S. Ristad and P. N. Yianilos, "Learning string-edit distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 5, pp. 522–532, 1998.

[59] M. A. Covington, S. A. Lunden, S. L. Cristofaro, C. R. Wan, C. T. Bailey, B. Broussard, R. Fogarty, S. Johnson, S. Zhang, and M. T. Compton, "Phonetic measures of reduced tongue movement correlate with negative symptom severity in hospitalized patients with first-episode schizophrenia-spectrum disorders," *Schizophrenia research*, vol. 142, no. 1-3, pp. 93–95, 2012.

A

Study's Protocol

Protocolo de recolha das amostras

Material Necessário
- Cronómetro;
- Gravador de áudio.

Duração do procedimento
Cerca de 20 minutos.

Procedimento
1 - Leia o documento "Consentimento Informado" ao participante e esclareça qualquer eventual dúvida sobre o estudo.

2 - Preencha as suas informações no documento "Consentimento Informado", e assine e date que explicou e esclareceu o participante.

3 - Dê ao participante uns minutos para interiorizar a informação e para assinar o documento "Consentimento Informado".

4 - Prepare o cronómetro para 1 minuto.

5 - Ligue o gravador, e prepare-se para gravar o áudio.

6 - Peça ao participante que enumere em 1 minuto palavras que comecem com "P", e após terminar de explicar a tarefa, comece o cronómetro e a gravar o áudio.

7 - Quando terminar o minuto, pare o gravador e diga "Obrigada, terminou o tempo".

8 - Prepare o cronómetro para 1 minuto e o gravador.

9 - Peça ao participante que enumere em 1 minuto diferentes animais, e após terminar de explicar a tarefa, comece o cronómetro e a gravar o áudio.

10 - Quando terminar o minuto, pare o gravador e diga "Obrigada, terminou o tempo".

11 - Prepare o gravador.

12 - Peça ao participante para ler a história "O Vento Norte e o Sol" (presente no anexo I), e após terminar de explicar a tarefa, comece a gravar o áudio.

13 - Quando o participante terminar a tarefa, pare o gravador.

15 - Prepare o gravador.

16 - Peça ao participante para contar a história "Os Três Porquinhos", caso este não saiba a história, leia-a (presente no anexo II). Depois de terminar de ler a história, peça-lhe para a contar de volta e comece a gravar o áudio.

17 - Quando o participante terminar a tarefa, desligue o gravador.

18 - Prepare o gravador.

19 - Mostre ao participante a imagem no anexo III, e peça para que construa uma história ou diga o que lhe vem à cabeça, após terminar de explicar a tarefa, comece a gravar o áudio.

20 - Quando o participante terminar a tarefa, desligue o gravador.

21 - Prepare o gravador.

22 - Mostre ao participante a imagem no anexo IV, e peça para que construa uma história ou diga o que lhe vem à cabeça, após terminar de explicar a tarefa, comece a gravar o áudio.

23 - Quando o participante terminar a tarefa, desligue o gravador.

24 - Prepare o gravador.

25 - Mostre ao participante a imagem no anexo V, e peça para que construa uma história ou diga o que lhe vem à cabeça, após terminar de explicar a tarefa, comece a gravar o áudio.

26 - Quando o participante terminar a tarefa, desligue o gravador.

Anexo I

"O Vento Norte e o Sol"

O vento norte e o sol discutiam qual dos dois era o mais forte, quando sucedeu passar um viajante envolto numa capa. Ao vê-lo, põem-se de acordo de que aquele que primeiro conseguisse obrigar o viajante a tirar a capa seria considerado o mais forte. O vento norte começou a soprar com muita fúria, mas quanto mais soprava, mais o viajante se aconchegava à sua capa, até que o vento norte desistiu. O sol brilhou então com todo o esplendor, e imediatamente o viajante tirou a capa. O vento norte teve assim de reconhecer a superioridade do sol.

Anexo II

"Os Três Porquinhos"

Era uma vez três irmãos porquinhos que viviam com a mãe, muito felizes. Dois porquinhos eram preguiçosos e não ajudavam nada em casa, enquanto o terceiro porquinho era trabalhador.

Um belo dia, a mãe, achando que os filhos já tinham maturidade, chamou-os e disse:
– Meus meninos, chegou a altura de saírem de casa pois já são grandes o suficiente para viverem sozinhos. Tenham juízo e muito cuidado com o Lobo Mau.

Dito isto, a mãe deu um farnel a cada um dos três porquinhos, assim como algumas economias para que comprassem material para construírem as suas casas.

E lá partiram os três porquinhos.

O primeiro porquinho, que era preguiçoso, decidiu construir uma casa que não desse trabalho nenhum. Apesar dos irmãos o avisarem que não era seguro, construiu uma casa de palha num só dia! Após terminar a sua casa de palha, foi tocar flauta e dançar.

O segundo porquinho, que era menos preguiçoso que o primeiro, resolveu construir a sua casa em madeira. Apesar de ser mais segura do que a casa de palha, uma casa de madeira não era, contudo, resistente o suficiente para impedir a entrada do Lobo Mau, advertiu o terceiro porquinho. No entanto, o nosso porquinho ignorou o conselho do irmão. E assim, em apenas dois dias construiu a sua casa de madeira!

Dito isto, pegou no violino e foi tocar e dançar juntamente com o primeiro irmão.

Por sua vez, o terceiro porquinho, que como vimos era trabalhador e precavido, decidiu construir a sua casa com tijolos.

– Os tijolos são um material muito resistente. Assim, o Lobo Mau não conseguirá destruir a minha casa – disse para si.

Dito isto, pôs-se, pacientemente, a trabalhar. A construção foi dura e avançou lentamente. Enquanto isso, os seus dois irmãos tocavam e dançavam.

Finalmente, após algumas semanas terminou a construção de uma sólida casa em tijolos.

Passado algum tempo, surgiu na floresta o Lobo Mau. Percebendo a presença dos porquinhos, pensou para si mesmo:
– Mas que bela refeição tenho à minha espera: três porquinhos bem gordinhos!

Dito isto, foi bater à casa de palha, que era a do primeiro porquinho. Vendo que era o Lobo Mau, o porquinho respondeu:
– Vai-te embora porque não te irei abrir a porta, Lobo Mau!

O Lobo Mau respondeu:
– Então vou soprar e soprar até levar esta casa pelo ar!

Dito isto, o lobo pôs-se a soprar e a casa de palha foi toda pelo ar. O primeiro porquinho, em pânico, conseguiu refugiar-se na casa de madeira que pertencia ao segundo irmão. O Lobo Mau dirigiu-se então para a casa de madeira, bateu à porta e pediu para entrar. Disseram então os dois porquinhos:
– Vai-te embora pois nunca te iremos abrir a porta, ó Lobo Mau!

O Lobo Mau respondeu:
– Então vou soprar e soprar até levar esta casa pelo ar!

E mais uma vez, o lobo pôs-se a soprar e a soprar e a casa de madeira acabou por ir toda pelo ar. O primeiro e o segundo porquinho, em pânico, fugiram e refugiaram-se na casa de tijolos que pertencia ao terceiro irmão.

Então, o Lobo Mau dirigiu-se para a casa de tijolos, e tal como das outras vezes, bateu à porta e pediu para entrar. Responderam então os três porquinhos:
– Vai-te embora já que não te vamos abrir a porta, Lobo Mau!

O Lobo Mau riu-se e respondeu:
– Então vou soprar e soprar até levar esta casa pelo ar!

E dito isto, confiante, começou a soprar e a soprar, e a soprar…a soprar…até que ficou sem ar. A casa, não se tinha mexido nem sequer um polegar!

O Lobo Mau subiu, então, ao telhado e tentou entrar na casa pela chaminé. No entanto, como o terceiro porquinho era muito precavido, tinha deixado um caldeirão de água a ferver debaixo da chaminé. Mal desceu pela chaminé abaixo, o lobo caiu no caldeirão e apanhou um enorme escaldão.

Depois fugiu e nunca, mas nunca mais voltou. Os três porquinhos ficaram a viver na casa de tijolos em segurança e muito felizes.

Anexo III

Anexo IV

Anexo V

B

Controls' Consent Form

Consentimento Informado

Qual é o objetivo deste estudo?

Este estudo, desenvolvido no âmbito de uma tese de mestrado em Engenharia Informática e de Computadores do Instituto Superior Técnico, pretende recolher amostras de fala de indivíduos, de modo a criar um classificador na área da saúde mental.

O que é que pedimos aos indivíduos que participam no estudo?

Será pedido aos participantes que numa sessão única de cerca de 20 minutos realizem cinco tarefas de produção de discurso, de modo que o áudio das suas respostas seja gravado e posteriormente estudado.

Que dados sobre o participante serão recolhidos?

Será registado o género, a faixa etária e a escolaridade dos participantes, para que seja possível fazer uma caracterização demográfica das amostras.

Como é que se garante a privacidade da informação recolhida?

Não será guardada nenhuma informação que permita relacionar os dados com a identidade pessoal, nem qualquer forma de contacto com o participante.
Toda a informação colhida é absolutamente confidencial, e será garantido o anonimato. A cada amostra será atribuído um identificador, sendo que todos os dados serão guardados dentro do servidor do INESC-ID, protegidos com senha de acesso e apenas acessíveis, tratados e manipulados dentro do INESC-ID por um membro da equipa de investigação ou autorizado pela mesma.

Como serão usados os dados colhidos?

Os dados colhidos serão tratados através dos seus identificadores e serão analisados pela equipa de investigação, assegurando a sua confidencialidade. Os dados poderão também ser utilizados para apresentação ou exibição de resultados, devidamente anonimizados, em publicações científicas, conferências ou eventos semelhantes.

Quais os riscos do estudo?

Não existem quaisquer riscos associados à realização das tarefas compreendidas no estudo.

Quais os benefícios do estudo?

Não existem benefícios diretos para os participantes do estudo.

Consentimento Informado

Quais os seus direitos enquanto participante no estudo?

É inteiramente livre de participar ou não no estudo, e pode a qualquer momento interromper a sua colaboração no mesmo, sem que com isso seja prejudicado no seu acompanhamento clínico, na instituição.
Como participante tem direito a solicitar ao responsável da Proteção de Dados o acesso aos dados pessoais que lhe digam respeito, fornecendo o identificador da amostra que lhe será atribuído. Tem também os direitos de retificação, remoção, limitação e oposição do tratamento, incluindo o direito de retirar consentimento em qualquer altura, sem prejuízo da licitude do tratamento eventual e previamente consentido. Para além disto, tem também o direito de apresentar reclamação à Comissão Nacional de Proteção de Dados.

Quais os custos ou incómodos para os participantes do estudo?

Não existem quaisquer custos ou incómodos inerentes à participação no estudo.

Quem contactar para colocar questões ou problemas?

Se tiver alguma questão relacionada com a sua participação no estudo, poderá contactar a investigadora principal Helena Sofia Pinto através do email sofia.pinto@tecnico.ulisboa.pt, ou o investigador Alberto Abad através do email alberto.abad@inesc-id.pt.

Encarregado de Proteção de Dados do INESC-ID:
dpo@inesc-id.pt

Informações CHLO:

Email do Encarregado de Proteção de Dados:
dpo@chlo.min-saude.pt

A investigadora principal,

Helena Sofia Andrade Nunes Pereira Pinto
(Professora Auxiliar)
Departamento de Engenharia Informática
Instituto Superior Técnico
Universidade de Lisboa

Consentimento Informado

Confirmo que expliquei à pessoa abaixo indicada, de forma adequada e inteligível, os procedimentos necessários a este estudo. Respondi a todas as questões que me foram colocadas e assegurei-me de que houve um período de reflexão suficiente para a tomada de decisão.

Nome do Investigador: _____

Data: ___/____/_____
Assinatura: _____

À pessoa/representante/acompanhante:

Por favor, leia com atenção todo o conteúdo deste documento. Não hesite em solicitar mais informações caso não esteja completamente esclarecido/a. Verifique se todas as informações estão corretas. Se tudo estiver conforme, então assine este documento e preencha com as suas informações.

Declaro ter compreendido os objetivos do estudo me foi proposto participar e explicado pelo profissional que assina este documento, ter-me sido dada oportunidade de fazer todas as perguntas sobre o assunto e para todas elas ter obtido resposta esclarecedora, ter-me sido garantido que não haverá prejuízo para os meus direitos assistenciais se eu recusar esta solicitação e os meus direitos como participante, e ter-me sido dado tempo suficiente para refletir sobre esta proposta.

Autorizo/Não autorizo ***(riscar o que não interessa)*** *a minha participação no estudo indicado e a gravação áudio das minhas respostas às tarefas de produção de fala.*

Nome: _____
Data: ___/____/_____
Assinatura: _____

Identificador da amostra (preenchido pelo investigador): _____

Nota: Este documento é feito em duas vias – uma para o processo e outra para ficar na posse de quem consente.

Consentimento Informado

Confirmo que expliquei à pessoa abaixo indicada, de forma adequada e inteligível, os procedimentos necessários a este estudo. Respondi a todas as questões que me foram colocadas e assegurei-me de que houve um período de reflexão suficiente para a tomada de decisão.

Nome do Investigador: _____

Data: ___/____/_____
Assinatura: _____

À pessoa/representante/acompanhante:

Por favor, leia com atenção todo o conteúdo deste documento. Não hesite em solicitar mais informações caso não esteja completamente esclarecido/a. Verifique se todas as informações estão corretas. Se tudo estiver conforme, então assine este documento e preencha com as suas informações.

Declaro ter compreendido os objetivos do estudo me foi proposto participar e explicado pelo profissional que assina este documento, ter-me sido dada oportunidade de fazer todas as perguntas sobre o assunto e para todas elas ter obtido resposta esclarecedora, ter-me sido garantido que não haverá prejuízo para os meus direitos assistenciais se eu recusar esta solicitação e os meus direitos como participante, e ter-me sido dado tempo suficiente para refletir sobre esta proposta.

Autorizo/Não autorizo ***(riscar o que não interessa)*** *a minha participação no estudo indicado e a gravação áudio das minhas respostas às tarefas de produção de fala.*

Nome: _____
Data: ___/____/_____
Assinatura: _____

Identificador da amostra (preenchido pelo investigador): _____

Nota: Este documento é feito em duas vias – uma para o processo e outra para ficar na posse de quem consente.

Consentimento Informado

Identificador da amostra (preenchido pelo investigador): _____

Por favor faça uma cruz de acordo com as suas informações:

1. **Género:**
   Feminino ___
   Masculino ___
   Não quer dizer ___

2. **Faixa etária:**
   18 - 29 anos ___
   30 - 39 anos ___
   40 - 49 anos ___
   50 - 59 anos ___
   60 - 69 anos ___
   70 - 79 anos ___
   > 80 anos ___

3. **Escolaridade:**
   4ª Ano ___
   6º Ano ___
   9º Ano ___
   12º Ano ___
   Ensino Superior/Universidade ___

Nota: Esta página será guardada pelo investigador para que seja feita a caracterização demográfica da amostra.

C

Patients' Consent Form

Consentimento Informado

Qual é o objetivo deste estudo?

Este estudo, desenvolvido no âmbito de uma tese de mestrado em Engenharia Informática e de Computadores do Instituto Superior Técnico, pretende recolher amostras de fala de indivíduos, de modo a criar um classificador na área da saúde mental.

O que é que pedimos aos indivíduos que participam no estudo?

Será pedido aos participantes que numa sessão única de cerca de 20 minutos realizem cinco tarefas de produção de discurso, de modo que o áudio das suas respostas seja gravado e posteriormente estudado.

Que dados sobre o participante serão recolhidos?

Será registado o género, a faixa etária e a escolaridade dos participantes, para que seja possível fazer uma caracterização demográfica das amostras.

Como é que se garante a privacidade da informação recolhida?

Não será guardada nenhuma informação que permita relacionar os dados com a identidade pessoal, nem qualquer forma de contacto com o participante.
Toda a informação colhida é absolutamente confidencial, e será garantido o anonimato. A cada amostra será atribuído um identificador, sendo que todos os dados serão guardados dentro do servidor do INESC-ID, protegidos com senha de acesso e apenas acessíveis, tratados e manipulados dentro do INESC-ID por um membro da equipa de investigação ou autorizado pela mesma.

Como serão usados os dados colhidos?

Os dados colhidos serão tratados através dos seus identificadores e serão analisados pela equipa de investigação, assegurando a sua confidencialidade. Os dados poderão também ser utilizados para apresentação ou exibição de resultados, devidamente anonimizados, em publicações científicas, conferências ou eventos semelhantes.

Quais os riscos do estudo?

Não existem quaisquer riscos associados à realização das tarefas compreendidas no estudo.

Quais os benefícios do estudo?

Não existem benefícios diretos para os participantes do estudo.

Consentimento Informado

Quais os seus direitos enquanto participante no estudo?

É inteiramente livre de participar ou não no estudo, e pode a qualquer momento interromper a sua colaboração no mesmo, sem que com isso seja prejudicado no seu acompanhamento clínico, na instituição.
Como participante tem direito a solicitar ao responsável da Proteção de Dados o acesso aos dados pessoais que lhe digam respeito, fornecendo o identificador da amostra que lhe será atribuído. Tem também os direitos de retificação, remoção, limitação e oposição do tratamento, incluindo o direito de retirar consentimento em qualquer altura, sem prejuízo da licitude do tratamento eventual e previamente consentido. Para além disto, tem também o direito de apresentar reclamação à Comissão Nacional de Proteção de Dados.

Quais os custos ou incómodos para os participantes do estudo?

Não existem quaisquer custos ou incómodos inerentes à participação no estudo.

Quem contactar para colocar questões ou problemas?

Se tiver alguma questão relacionada com a sua participação no estudo, poderá contactar a investigadora principal Helena Sofia Pinto através do email sofia.pinto@tecnico.ulisboa.pt, ou o investigador Alberto Abad através do email alberto.abad@inesc-id.pt.

Encarregado de Proteção de Dados do INESC-ID:
dpo@inesc-id.pt

Informações CHLO:

Email do Encarregado de Proteção de Dados:
dpo@chlo.min-saude.pt

A investigadora principal,

Helena Sofia Andrade Nunes Pereira Pinto
(Professora Auxiliar)
Departamento de Engenharia Informática
Instituto Superior Técnico
Universidade de Lisboa

Consentimento Informado

Confirmo que expliquei à pessoa abaixo indicada, de forma adequada e inteligível, os procedimentos necessários a este estudo. Respondi a todas as questões que me foram colocadas e assegurei-me de que houve um período de reflexão suficiente para a tomada de decisão.

Nome do Investigador: _____

Data: ___/____/_____
Assinatura: _____

À pessoa/representante/acompanhante:

Por favor, leia com atenção todo o conteúdo deste documento. Não hesite em solicitar mais informações caso não esteja completamente esclarecido/a. Verifique se todas as informações estão corretas. Se tudo estiver conforme, então assine este documento e preencha com as suas informações.

Declaro ter compreendido os objetivos do estudo me foi proposto participar e explicado pelo profissional que assina este documento, ter-me sido dada oportunidade de fazer todas as perguntas sobre o assunto e para todas elas ter obtido resposta esclarecedora, ter-me sido garantido que não haverá prejuízo para os meus direitos assistenciais se eu recusar esta solicitação e os meus direitos como participante, e ter-me sido dado tempo suficiente para refletir sobre esta proposta.

Autorizo/Não autorizo ***(riscar o que não interessa)*** *a minha participação no estudo indicado e a gravação áudio das minhas respostas às tarefas de produção de fala.*

Nome: _____
Data: ___/____/_____
Assinatura: _____

Identificador da amostra (preenchido pelo investigador): _____

Nota: Este documento é feito em duas vias – uma para o processo e outra para ficar na posse de quem consente.

Consentimento Informado

Confirmo que expliquei à pessoa abaixo indicada, de forma adequada e inteligível, os procedimentos necessários a este estudo. Respondi a todas as questões que me foram colocadas e assegurei-me de que houve um período de reflexão suficiente para a tomada de decisão.

Nome do Investigador: _____

Data: ___/____/_____
Assinatura: _____

À pessoa/representante/acompanhante:

Por favor, leia com atenção todo o conteúdo deste documento. Não hesite em solicitar mais informações caso não esteja completamente esclarecido/a. Verifique se todas as informações estão corretas. Se tudo estiver conforme, então assine este documento e preencha com as suas informações.

Declaro ter compreendido os objetivos do estudo me foi proposto participar e explicado pelo profissional que assina este documento, ter-me sido dada oportunidade de fazer todas as perguntas sobre o assunto e para todas elas ter obtido resposta esclarecedora, ter-me sido garantido que não haverá prejuízo para os meus direitos assistenciais se eu recusar esta solicitação e os meus direitos como participante, e ter-me sido dado tempo suficiente para refletir sobre esta proposta.

Autorizo/Não autorizo ***(riscar o que não interessa)*** *a minha participação no estudo indicado e a gravação áudio das minhas respostas às tarefas de produção de fala.*

Nome: _____
Data: ___/____/_____
Assinatura: _____

Identificador da amostra (preenchido pelo investigador): _____

Nota: Este documento é feito em duas vias – uma para o processo e outra para ficar na posse de quem consente.

Consentimento Informado

Identificador da amostra (preenchido pelo investigador): _____

Informações do participante (a preencher pelo investigador):
(Por favor faça uma cruz de acordo com as informações)

1. **Doença e o seu tipo:**
   Psicose Afetiva ___
   Psicose Funcional ___
   Distúrbio bipolar I ___
   Distúrbio bipolar II ___
   Desordem ciclotímica ___
   Outros Transtornos Bipolares ___

2. **Valor total na escala BPRS** (*Brief Psychiatric Rating Scale*): _____

3. **Tempo de Doença** (em anos): ____

4. **Género:**
   Feminino ___
   Masculino ___
   Não quer dizer ___

5. **Faixa etária:**
   18 - 29 anos ___
   30 - 39 anos ___
   40 - 49 anos ___
   50 - 59 anos ___
   60 - 69 anos ___
   70 - 79 anos ___
   > 80 anos ___

6. **Escolaridade:**
   4ª Ano ___
   6º Ano ___
   9º Ano ___
   12º Ano ___
   Ensino Superior/Universidade ___

Nota: Esta página será guardada pelo investigador para que seja feita a caracterização demográfica da amostra.

D

Features Extracted

Table D.1: Speech features extracted from recordings and/or transcriptions, and their respective description.

Feature	Feature Description
Number Words	The number of words spoken.
Number Syllables	The number of syllables spoken.
Audio Duration (seconds)	The duration of the audio segment.
Speaking Rate (words / second)	The number of words spoken per second.
Articulation Rate (syllables / second)	The number of syllables spoken spoken per second.

Table D.2: Sound features extracted, using eGeMAPS, from recordings. Features displayed do not include their variations such as *mean*, *standard deviation*, among others.

Features		
F0semitoneFrom27.5Hz_sma3nz	spectralFlux_sma3	loudness_sma3
mfcc1_sma3	jitterLocal_sma3nz	shimmerLocaldB_sma3nz
HNRdBACF_sma3nz	logRelF0-H1-H2_sma3nz	F1frequency_sma3nz
F1amplitudeLogRelF0_sma3nz	F2frequency_sma3nz	F2amplitudeLogRelF0_sma3nz
F3frequency_sma3nz	F3amplitudeLogRelF0_sma3nz	alphaRatioV_sma3nz
hammarbergIndexV_sma3nz	slopeV0-500_sma3nz	slopeV500-1500_sma3nz
spectralFluxV_sma3nz	mfcc1V_sma3nz	alphaRatioUV_sma3nz
hammarbergIndexUV_sma3nz	slopeUV0-500_sma3nz	slopeUV500-1500_sma3nz
spectralFluxUV_sma3nz	loudnessPeaksPerSec	VoicedSegmentsPerSec
MeanVoicedSegmentLengthSec	StddevVoicedSegmentLengthSec	MeanUnvoicedSegmentLength
StddevUnvoicedSegmentLength	equivalentSoundLevel_dBp	

Table D.3: Content features extracted from transcriptions and their respective description.

Technique	Feature	Feature Description
LCA	Max Cossine w/ Cluster ⟨#⟩ for Target	Maximum cossine between sets of words and each (#) cluster developed for target group.
	Max Cossine w/ Cluster ⟨#⟩ for Non-Target	Maximum cossine between sets of words and each (#) cluster developed for non-target group.
Sentiment Analysis	Sentilex - avg Score	Number of words mapped to Sentilex lexicon
	Sentilex - number Scores	Average score from words mapped to Sentilex lexicon.
	Valence RoBERTa - Score	Score achieved when feeding subject's transcriptions through fine-tuned RoBERTa model for valence score.

Table D.4: Structure/Coherence features extracted from transcriptions and their respective description.

Technique	Feature	Feature Description
Word Graph	#nodes	Number of nodes in word graph.
	#edges	Number of edges in word graph.
	diameter	Diameter of the word graph.
	#repeated_edges	Number of repeated edges in word graph.
	#parallel_edges	Number of parallel edges in word graph.
	#average_total_degree	Average total degree of nodes in word graph.
	#average_shortest_path	Average shortest path in word graph.
	#average_clustering_coefficient	Average clustering coefficient from the word graph.
	LSCC - #nodes	Number of nodes in largest strongly connected component from word graph.
	LSCC - #edges	Number of edges in largest strongly connected component from word graph.
	LSCC - probability	Probability of largest strongly connected component happening by random shuffling subject's transcription.
	SCCs - min #nodes	Minimum number of nodes from all of the strongly connected components in word graph.
	SCCs - avg #nodes	Average number of nodes from all of the strongly connected components in word graph.
	SCCs - max #nodes	Maximum number of nodes from all of the strongly connected components in word graph.
	SCCs - min #edges	Minimum number of edges from all of the strongly connected components in word graph.
	SCCs - avg #edges	Average number of edges from all of the strongly connected components in word graph.
	SCCs - max #edges	Maximum number of edges from all of the strongly connected components in word graph.
LSA	First Order Coherence	Average of the cosine difference between each set of words and the subsequent set of words. Typically each each set of words represents a sentence.
	Second Order Coherence	Average of the cosine difference between each set of words and the set of words two positions ahead. Typically, each set of words represents a sentence.
Vector Unpacking	avg #ephocs	Average number of ephocs required to fully develop the neural networks used for vector unpacking.
	avg model cosine similarity	Average of the neural networks' cosine similarity score achieved.
	avg model loss	Average of the neural networks' loss score achieved.
	avg ratio #zero_weights	Average ratio of weights, in the various neural networks, set as 0 when model has achieved its best performance.

E

Feature Set Analysis

Table E.1: Overview analysis and profiling of the structure/coherence feature set extracted from recordings and transcriptions.

Feature	count	mean	std	min	25%	50%	75%	max
Word Graph - #nodes	1241	38.36	32.51	0.00	16.00	29.00	57.00	271.00
Word Graph - #edges	1241	64.64	81.93	0.00	17.00	35.00	93.00	919.00
Word Graph - diameter	1241	9.25	4.36	-1.00	7.00	8.00	11.00	32.00
Word Graph - #repeated edges	1241	7.47	17.03	0.00	0.00	1.00	9.00	237.00
Word Graph - #parallel edges	1241	9.28	19.12	0.00	0.00	2.00	12.00	263.00
Word Graph - avg degree	1241	2.30	0.92	-1.00	2.00	2.00	3.00	6.00
Word Graph - avg shortest path	1241	3.91	1.82	-1.00	2.00	4.00	5.00	9.00
Word Graph - avg clust. coefficient	1241	0.02	0.09	-1.00	0.00	0.02	0.05	0.30
Word Graph - LSCC - #nodes	1241	33.09	35.12	-1.00	6.00	22.00	57.00	267.00
Word Graph - LSCC - #edges	1241	59.26	84.26	-1.00	7.00	30.00	93.00	915.00
Word Graph - LSCC - probability	1241	0.22	0.36	-1.00	0.00	0.00	0.30	1.00
Word Graph - SCCs - min #nodes	1241	13.22	27.48	-1.00	1.00	1.00	1.00	217.00
Word Graph - SCCs - avg #nodes	1241	19.80	27.68	-1.00	1.63	5.80	29.00	217.00
Word Graph - SCCs - max #nodes	1241	33.09	35.12	-1.00	6.00	22.00	57.00	267.00
Word Graph - SCCs - min #edges	1241	22.39	56.47	-1.00	0.00	0.00	0.00	543.00
Word Graph - SCCs - avg #edges	1241	35.21	59.92	-1.00	0.86	6.67	50.00	543.00
Word Graph - SCCs - max #edges	1241	59.27	84.25	-1.00	7.00	30.00	93.00	915.00
LSA - First Order Coherence	1241	0.74	0.28	-0.16	0.70	0.86	0.92	1.00
LSA - Second Order Coherence	1241	0.64	0.37	-0.32	0.41	0.85	0.91	1.00
Vector Unp. - avg #ephocs	1241	1404.63	908.80	11.00	673.00	1435.00	2062.00	5000.00
Vector Unp. - avg cosine similarity	1241	0.10	0.13	0.03	0.04	0.06	0.10	1.00
Vector Unp. - avg loss	1241	0.07	0.09	1.10e-05	0.02	0.02	0.11	0.48
Vector Unp. - avg ratio zero weights	1241	-0.78	0.12	-1.00	-0.84	-0.76	-0.70	-0.36

Table E.2: Overview analysis and profiling of the content feature set extracted from recordings and transcriptions.

Feature	count	mean	std	min	25%	50%	75%	max
SentiLex - avg Score	1241.00	-0.03	0.65	-1.00	-0.50	0.00	0.50	1.00
SentiLex - number Scores	1241.00	3.63	5.14	0.00	1.00	2.00	4.00	57.00
Valence RoBERTa Score	1241.00	0.59	0.07	0.32	0.55	0.59	0.65	0.73
LCA - Max Cossine w/ Cluster 0 for Target	1241.00	-0.02	0.14	-0.42	-0.14	-0.03	0.09	0.44
LCA - Max Cossine w/ Cluster 1 for Target	1241.00	0.24	0.09	-0.19	0.17	0.25	0.30	0.57
LCA - Max Cossine w/ Cluster 2 for Target	1241.00	-0.05	0.09	-0.33	-0.11	-0.07	-0.01	0.44
LCA - Max Cossine w/ Cluster 3 for Target	1241.00	0.38	0.15	-0.16	0.28	0.40	0.48	0.70
LCA - Max Cossine w/ Cluster 4 for Target	1241.00	-0.02	0.11	-0.35	-0.11	-0.03	0.05	0.38
LCA - Max Cossine w/ Cluster 5 for Target	1241.00	-0.01	0.14	-0.40	-0.12	0.00	0.08	0.34
LCA - Max Cossine w/ Cluster 6 for Target	1241.00	0.39	0.12	-0.31	0.34	0.42	0.48	0.70
LCA - Max Cossine w/ Cluster 7 for Target	1241.00	0.04	0.10	-0.27	-0.02	0.03	0.11	0.39
LCA - Max Cossine w/ Cluster 8 for Target	1241.00	-0.03	0.08	-0.24	-0.08	-0.03	0.03	0.43
LCA - Max Cossine w/ Cluster 9 for Target	1241.00	0.27	0.11	-0.13	0.18	0.26	0.34	0.72
LCA - Max Cossine w/ Cluster 0 for Non-Target	1241.00	0.08	0.08	-0.24	0.02	0.08	0.14	0.39
LCA - Max Cossine w/ Cluster 1 for Non-Target	1241.00	0.29	0.10	-0.25	0.21	0.29	0.35	0.58
LCA - Max Cossine w/ Cluster 2 for Non-Target	1241.00	0.17	0.09	-0.11	0.12	0.16	0.23	0.58
LCA - Max Cossine w/ Cluster 3 for Non-Target	1241.00	0.15	0.14	-0.33	0.06	0.14	0.24	0.52
LCA - Max Cossine w/ Cluster 4 for Non-Target	1241.00	0.21	0.10	-0.20	0.14	0.22	0.29	0.48
LCA - Max Cossine w/ Cluster 5 for Non-Target	1241.00	0.36	0.17	-0.15	0.25	0.35	0.45	0.79
LCA - Max Cossine w/ Cluster 6 for Non-Target	1241.00	-0.02	0.09	-0.36	-0.08	-0.03	0.03	0.37
LCA - Max Cossine w/ Cluster 7 for Non-Target	1241.00	-0.11	0.11	-0.41	-0.20	-0.12	-0.02	0.31
LCA - Max Cossine w/ Cluster 8 for Non-Target	1241.00	0.15	0.09	-0.34	0.09	0.14	0.20	0.48