Discourse Analysis for Mental Health Monitoring

Pedro Rafael Barriga Negra Ascensão

Thesis to obtain the Master of Science Degree in

Information Systems and Computer Engineering

Supervisors: Prof. David Manuel Martins de Matos

Prof. Alberto Abad Gareta

Examination Committee

Chairperson: Prof. Daniel Jorge Viegas Gonçalves

Supervisor: Prof. David Manuel Martins de Matos

Member of the Committee: Prof. Mário Jorge Costa Gaspar da Silva

October 2018

Dedicated to my family and friends

Acknowledgments

I would like to thank my dissertation supervisors Prof. David Martins de Matos, Prof. Alberto Abad, and also to Ph.D. student Anna Maria for their insight, support, and sharing of knowledge that made this Thesis possible.

Secondly, I would like to thank my family for always being there for me and without whom this project and course, in general, would not be possible.

Last but not least, to all my friends that with all the football games, beers and conversations kept reminding me that I am not the only one doing the thesis.

Resumo

A detecção de demência é importante não apenas para os doentes para os quais o diagnóstico é o primeiro passo para o apoio e tratamento adequado, mas também para os investigadores, que afirmam que o diagnóstico precoce é essencial para encontrar uma cura. Deste modo, esta tese aborda o diagnóstico da demência recorrendo ao processamento de linguagem natural para a análise do discurso de doentes com possíveis sintomas desta doença. Como o discurso humano envolve processos cognitivos relacionados a subsistemas linguísticos, a sua análise é uma componente fundamental para a avaliação de um potencial declínio da saúde mental do doente. Três abordagens usando diferentes técnicas foram experimentadas nesta tese para diagnosticar automaticamente demência recorrendo a transcrições de descrições de uma imagem. Usando uma abordagem *Bag-of-Words*, em combinação com um teste estatístico para a seleção de palavras, ao classificar entre doentes com Alzheimer e pessoas saudáveis, 86,1% das transcrições foram correctamente classificadas, o que corresponde a um valor superior em 4% relativamente ao estado da arte. A classificação entre indivíduos com demência, nas suas diferentes formas, e pessoas saudáveis é uma tarefa mais complexa do que a anterior. Mantendo esta abordagem, a percentagem de casos correctamente classificados foi semelhante ao estado da arte. Infelizmente, as outras duas abordagens, *embeddings* e *LSTM*, tiveram um pior desempenho quando comparadas com os resultados anteriores.

Palavras-chave: Demência, Processamento Natural de Linguagem, Classificação, Bag-of-Words

Abstract

Detection of dementia is important not only for patients for whom a diagnosis is the first step for appropriate support and treatment, but also for researchers, who claim that early diagnosis is an important step for finding a cure. Therefore, this thesis addresses dementia diagnosis by using natural language processing for the discourse analysis of patients with possible dementia disease. Since discourse production involves cognitive processes related to linguistic subsystems and knowledge of the world, its analysis is, therefore, a fundamental component for the assessment of potential mental health decline. Three approaches using different techniques were tried in this thesis to automatically detect dementia from transcripts of an image description task. Using a bag-of-words approach, in combination with a statistic test for the word selection, we reached an accuracy of 86.1%, which is 4% higher than the state-of-art, when classifying between Alzheimer's patients and healthy subjects. This same proposed approach, when classifying between different dementia diseases and healthy individuals, which is typically more challenging than considering only AD, performed, in terms of accuracy score, similarly to the state-of-art that classifies between Alzheimer's disease and control. Unfortunately, the other two approaches, an embeddings based strategy, and an LSTM Neural Network had poor performance when compared to previous work. Finally, the bag-of-words model was deployed on a Heroku server, and an Android client was built to send requests to the server and return to the user his probability to have dementia.

Keywords: Dementia, Natural Language Processing, Classification, Bag-of-words

Contents

Acknowledgments v
Resumo vii
Abstract ix
List of Figures xiii
List of Acronyms xv

1 Introduction **1**
1.1 Goal 3
1.2 Structure of the Document 3

2 Dementia Characterization **5**
2.1 Dementia and Discourse 5
2.1.1 Alzheimer's Disease 6
2.1.2 Mild Cognitive Impairment 8
2.2 Diagnostic 9
2.3 Dataset 9
2.4 Summary 10

3 Related Work **13**
3.1 Diagnosis as a Classification Problem 13
3.1.1 Alzheimer's vs Control 13
3.1.2 Mild Cognitive Impairment vs Control 15
3.2 Diagnosis as a Regression Problem 16
3.3 Discussion 17
3.4 Summary 18

4 Methodology **19**
4.1 Characterization of the DementiaBank dataset 19
4.1.1 Critique 21
4.2 Preprocessing 22
4.3 General method 24
4.4 Models 24

4.5 Evaluation . . . 25
4.6 Summary . . . 25

5 Diagnosis as Classification **27**
5.1 Bag-of-words approach . . . 27
5.1.1 Word selection . . . 28
5.1.2 Addition of Age and Mean Word Length as Features . . . 31
5.1.3 Alzheimer's vs Control . . . 33
5.2 Embeddings . . . 35
5.2.1 Alzheimer's vs Control . . . 39
5.3 Discussion . . . 41
5.4 Summary . . . 41

6 Diagnosis as Regression **43**
6.1 Bag-of-words Baseline . . . 43
6.2 LSTM Neural Networks . . . 45
6.3 Using MMSE Score for classification . . . 47
6.4 Discussion . . . 49
6.5 Summary . . . 50

7 Conclusions and Future Work **53**
7.1 Future Work . . . 54

Bibliography **57**

A Deployment and Client **65**
A.1 Deployment . . . 65
A.2 Android Client . . . 66

List of Figures

1.1 Population aged above 60 years by development group 2

2.1 Comparison of a normal aged brain and the brain with Alzheimer's 7
2.2 Image of the Cookie Theft test 10

3.1 Example of a co-occurrence network enriched with semantic information [48]. 15

4.1 Number of samples per disease type or control in DementiaBank 19
4.2 Number of recordings per subject in DementiaBank 20
4.3 Age Distribution 21
4.4 Average Length Word Distribution 22
4.5 Boxplot representing the MMSE between dementia disease types 23
4.6 Processing pipeline from clinical interview to evaluation 24

5.1 Results of the TF-IDF approach 28
5.2 χ^2 values for the top 30 words 29
5.3 Accuracy for different models as we vary the number of words included 30
5.4 Results for the Tf-IDF + χ^2 approach 30
5.5 Confusion Matrix for the TF-IDF + χ^2 approach 31
5.6 Accuracy for different combinations of features as we vary the number of words included . 32
5.7 Results for the TF-IDF + χ^2 approach using new features 33
5.8 Confusion Matrix for the TF-IDF + χ^2 approach using the age feature 33
5.9 Accuracy for different combinations of features as we vary the number of words included (AD) 34
5.10 Results for the Alzheimer's Disease classification 34
5.11 Confusion Matrix for the Alzheimer's Disease classification 35
5.12 Accuracy using the mean of the embeddings 37
5.13 Accuracy using the sum of the embeddings 37
5.14 Results for the mean embeddings approach 38
5.15 Confusion matrix for the mean of the embeddings approach 39
5.16 Results for the mean embeddings approach for classifying AD 40
5.17 Confusion Matrix for the mean embeddings approach for classifying AD 40

6.1 MAE for different models as we vary the number of words included 43
6.2 Minimum MAE for each model 44
6.3 Neural Network Model Architecture 46
6.4 MAE using LSTM network as we vary the number of the embeddings dimensions 47
6.5 Classification results for the regression models 48
6.6 Confusion matrix using the regression approach 49

A.1 Interaction between User, Android Client and the Server 65
A.2 Example of a request to the Flask localhost server and respective probability response. . 66

List of Acronyms

AD Alzheimer's Disease

AUC Area Under the ROC Curve

BoW Bag-of-Words

MAE Mean Absolute Error

MCI Mild Cognitive Impairment

MMSE Mini Mental State Examination

OCDE Organization for Economic Co-operation and Development

TF-IDF Term Frequency - Inverse Document Frequency

Chapter 1

Introduction

Mental health diseases are a broad category of brain diseases that cause a long-term and often gradual decrease in cognitive functions, such as the ability to recognize objects or the ability to understand and produce coherent language, and sometimes they are accompanied by a progressive impairment of memory [1]. These syndromes are often called dementia. In the first stages of dementia, the signs and symptoms may be subtle, which corresponds to the earliest stage of dementia, called Mild Cognitive Impairment (MCI). The patients with MCI may have some memory trouble, however, they solve everyday problems and handle their own life affairs well [2]. There are also other types of diseases belonging to the dementia group, Alzheimer's Disease (AD) being the most frequent one, affecting 50 million people worldwide (two-thirds of dementia cases) [3]. Nevertheless, there are other important mental health diseases, such as vascular dementia, Lewy body dementia, and frontotemporal dementia.

Furthermore, mental health diseases primarily affect senior citizens above the age of 60, where the prevalence ranges between 5%-7%, which increases with age [4]. This is worrisome, since the world's population is getting older (seen on Figure 1.1). Almost every developed country is dealing with the growth in the number and proportion of older people in their population [5].

More specifically, Portugal is the fourth country of the Organization for Economic Co-operation and Development (OCDE) with more subjects with dementia per 1000 inhabitants [6]. The mean of the OCDE countries in the OCDE is 14.8 subjects per 1000 inhabitants, while in Portugal is 19.9. Furthermore, according to this report, there are more than 205 thousand people in Portugal affected by dementia, a number that will increase to 322 thousand by 2037.

Economically, dementia diseases also burden health care systems. For instance, AD is projected to become a trillion dollar disease worldwide by 2018, which makes it one of the most costly diseases in developed countries [7].

Unfortunately, there is no cure yet for diseases that cause dementia, however researchers believe early detection will be the key to preventing, slowing and even stopping the disease [8]. Nevertheless, of the 47 million people that live with dementia nowadays, only 25% receive formal diagnosis [9]. There are some ways to diagnose dementia, such as, repeated medical follow-up with detailed cognitive assessment, neuroimaging and blood tests, which can be distressing for elderly people. Moreover, in

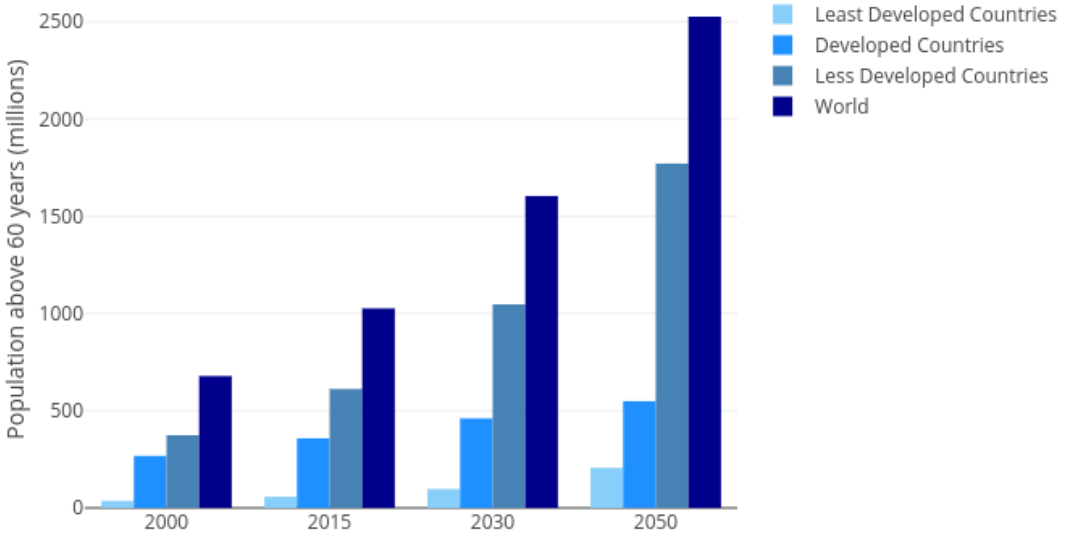

Figure 1.1: Population aged $\geq$ 60 years by development group. Data retrieved from [5].

developing countries, access to these type of resources is often not available, which is reflected in the high rates of undiagnosed dementia in those regions [9]. What society needs is a diagnostic tool which is non-invasive, inexpensive, and easy to administer, so patients in every region, developed or not, can receive medical care, as well as make lifestyle choices that can slow the progression of the disease [8].

One solution for this problem is to develop an automated tool that can diagnose patients by detecting changes on their language. Language, in general, is a capacity that allows the representation of concepts through an arbitrary set of symbols. In addition, in the case of humans, it can either represent concrete concepts as well as abstract propositions. Therefore, in order to produce and comprehend discourse, humans have to resort to cognitive processes [10]. Furthermore, while dementia is normally characterized by a decline in cognitive functions, such as attention, orientation, judgment, visuospatial abilities, and even memory in the case of AD [11], language impairments has been suggested as being more significant than other symptoms [12]. These language impairments include: difficulties in finding words, basic vocabulary, difficulty recalling names of everyday objects, tendency to repeat ideas, difficulty to produce sounds, syllables and words.

Therefore, given the importance of language impairments in detecting early signs of dementia, researchers are applying advances in machine learning (ML) and Natural Language Processing (NLP) to develop a tool that can help diagnose dementia based on a sample of a patient's speech. Such a tool would assist clinicians in making a diagnosis and hopefully would make obsolete the need for more invasive screening techniques. Furthermore, it could be easily distributed to developing countries via a mobile phone application, which would mitigate the problem of these regions not having access to more complex diagnose methods, such as, neuroimaging.

Finally, previous work in this area has shown positive preliminary results using language to distinguish between patients with and without dementia, as well as between subtypes of dementia. However, most of this previous work has been limited by small data sets and has focused only on one form of language production, namely spoken language [13, 14].

1.1 Goal

The main goal of this project is to develop a system that allows the detection of dementia disease through spontaneous speech. Therefore the problem at hand can be seen as a **classification problem** or a **regression problem**.

In the case of being a **classification problem**, given a transcript or a recording, the system will classify it as being produced by a dementia patient or by a healthy subject. In order to this to be possible, the system needs to know how the discourse of a dementia patient and a healthy subject differ.

If we see it as a **regression problem**, the system will give a score to each transcript, in which, for example, higher scores mean higher cognitive capabilities, therefore, there is no separation in classes between patients.

1.2 Structure of the Document

The rest of the thesis is structured as follows Chapter 2 explains the fundamental concepts involving medical background to understand the remainder of the thesis, more specifically, a medical overview of dementia with focus on AD and MCI. Chapter 3 discusses previous related work, which we divide in research studies focused in classification and regression of dementia from speech. Chapter 4 discusses the methodology common to each of the experiments, including preprocessing, evaluation and models used. We also provide an in-depth analysis of the dataset in this chapter. Chapter 5 and Chapter 6 presents all the approaches, including results and discussion, done by either treating the problem as a classification one or a regression one, respectively.

We conclude with Chapter 7 which summarizes the work and outlines some ideas for future research.

Chapter 2

Dementia Characterization

In this chapter, I present an overview of dementia focusing mainly on Alzheimer's Disease and Mild Cognitive Impairment since they will be the main focus of this thesis. The idea is not to provide an exhaustive review of the current state of medical knowledge of dementia, but rather to provide the relevant information to understand the work mentioned. I am going to focus on language impairments in AD and in MCI, since these are common dementia diseases. Finally, I will explain some diagnostic methods to screen dementia.

2.1 Dementia and Discourse

Dementia is a general term for an abnormal decline in the cognitive/mental ability severe enough to interfere with daily life, in which the symptoms are often gradual and irreversible. To be classified as a neurocognitive disorder, the patient must show a "significant cognitive decline from a previous level of performance in one or more cognitive domains" [15]:

- Learning and memory;
- Language;
- Executive function;
- Complex Attention;
- Perceptual/Motor;
- Social cognition;

and:

a) The cognitive deficits interfere with the daily life;

b) The cognitive deficits do not only occur in occasions of a delirium;

c) The cognitive deficits are not explained by other mental disorder, such as, schizophrenia.

Furthermore, since discourse is related to cognitive processes, it is expected that dementia leads to language impairments [16]. This type of impairment is common to all dementia diseases, although they may manifest themselves differently depending of the underlying disease [17].

Alzheimer's Disease is the most common type of dementia that causes problems with memory, thinking and behavior. Usually, early clinical symptoms include difficulty remembering names and recent events, apathy and depression. In addition, impaired judgment, disorientation, confusion, trouble in speaking, swallowing and walking are later symptoms of this disease. Even tough we will cover this afterwards, regarding language impairments, AD patients have difficulties in finding the right word for objects and have an incoherent and loud speech. Although AD is the most common type of dementia, responsible for 60% to 80% of cases [3], there are also other important diseases.

Vascular dementia caused by a disease or injury to the brain that obstructs blood flow, is the second most common type, responsible for 10% of cases. The symptoms of this disease often overlap with those of AD, although without the impairments in memory. Concerning language impairments, these patients have difficulties in naming objects and in word comprehension. Their speech is also incomprehensible and with decreased complexity.

Lewy Bodies is another disorder related to dementia. This disease is characterized by a pattern of decline similar to AD, including problems with memory and judgment and behaviour changes. Furthermore, patients with Lewy Bodies have visual hallucinations, muscle rigidity and tremors. The language impairments are very similar to AD, however their severity may fluctuate daily.

Frontotemporal dementia which is caused by damage to brain cells, especially in the front and side regions of the brain, is rare disease. Main symptoms include changes in personality and behaviour. These patients have a slow and hesitant speech, with grammatical mistakes. They also have trouble in naming objects and in reading and writing text.

Parkinson's Disease is another important disease that can cause dementia. Dementia is a common symptom that develops in later stages of the disease. Patients with Parkinson's have non-articulated and slow speech, with a loss of verbal fluency.

2.1.1 Alzheimer's Disease

The cause of AD is poorly understood. Researchers believe that 70% of the risk is due to genetics, with other risk factors including an history of head injuries, depression or hypertension. Physical exercise and avoiding obesity may decrease the risk of the disease. However, these claims lack evidence to support them.

The most common early symptom of AD is a difficulty to remember past events. As the disease progresses, other symptoms include language impairments, disorientation, mood swings, loss of motivation and not managing self care. Eventually, usually after three to nine years, bodily functions are lost,

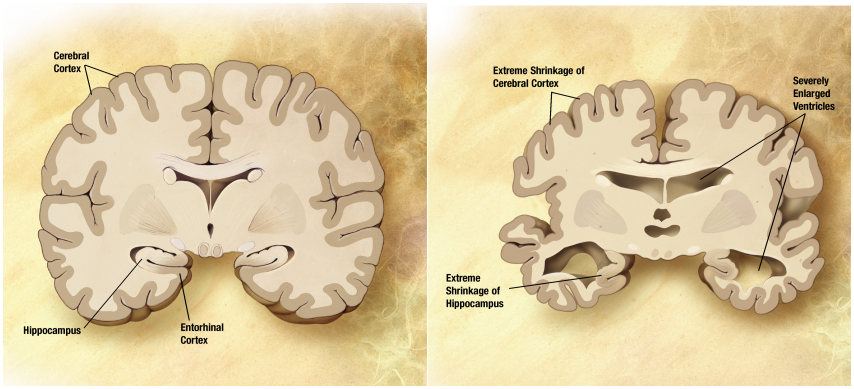

Figure 2.1: Comparison of a normal aged brain (left) and the brain of a person with Alzheimer's (right). [18]

causing death. One of the main problems is that at early stages these symptoms may be mistaken with the natural process of ageing. Figure 2.1 shows the difference between a normal brain from an aged person and a brain from a patient with AD.

Focusing on the symptom in study, in this case language impairment, researchers found impairments at the semantic level and also at the syntactic level.

Semantic studies indicate that changes in patients' speech are detectable in early stages of the disease. These changes include a decline in the complexity of the grammar and vocabulary used, word-finding difficulties, and semantic content deficiencies [19], which make the discourse of the patients seem "empty". Moreover, dementia patients tend to repeat ideas [20] and have difficulties in maintaining a conversation, since their speech is often incoherent [21]. In addition, other functions, such as understanding metaphor and sarcasm, are often impaired in patients with dementia [22]. Finally, studies also show that these patients have difficulties in maintaining a theme throughout the conversation [13].

Syntactically, the effects of dementia are controversial, since some studies have reported syntactical impairments in patients with dementia, while others claim that these apparent deficits are due to difficulties with memory and semantics, not to syntactic impairments [10]. On one hand, various studies have found evidence for a decrease in syntactic complexity. For instance, Bernard Croisile et al. [23] claimed that AD patients produce fewer subordinate clauses when compared to the control subjects; Seyed Sajjadi et al. [24] reported that patients with mild AD tend to produce fewer complex syntactic units in a picture description task when compared to an interview environment. On the other hand, Daniel Kempler et al. [25] observed that AD patients used the same range of syntactic constructions as control subject. Moreover, Guila Glosser and Toni Deser [26] interviewed four different type of subjects: patients with aphasia, patients with AD, patients with closed head injury and control subjects. The authors reached the conclusion that, syntactically, patients with AD and control subjects did not differ, i.e., both groups produced the same amount of syntactic errors.

In addition, some studies point out that the language decline in patients with dementia is heterogeneous, i.e., different patients with dementia would not have the same values for their language impairment. For instance, John Hodges and Karalyn Patterson [27] divided AD patients into three different categories, depending on their disease severity, and assessed their semantic impairment. The authors

reported a wide range of performance between the two lowest groups.

2.1.2 Mild Cognitive Impairment

Mild Cognitive Impairment is a neurological disorder affecting older adults that involves a decline in the cognitive function with minimal impact in day-to-day activities. Unlike other types of dementia, the patients with MCI have the ability to perform functional tasks, however slower and less efficient.

Moreover, some studies estimate that MCI affects 12-18% of the people over the age of 60 [28]. Annually, 8% to 15% of the people with MCI will progress to dementia while the rest either revert to normal cognition or stay in the same condition [28]. While MCI has no single cause, it can be due to neurodegenerative diseases (AD in most of the cases) or reversible causes, such as, psychiatric illness or metabolic disturbances (thyroid disease, vitamin B12 deficiency, etc.).

In addition, depending on the symptoms it presents, MCI can be classified in two different types. ”Memory MCI” or ”MCI Amnestic” happens when the predominant symptom is memory loss. Researchers have shown that these patients tend to progress to AD at a rate of 10% to 15% per year. When the patients have impairments in areas other than memory, it is classified as ”Multi-domain MCI” and these patients tend to be more likely to develop other types of dementia rather than AD.

Although there are less studies concerning MCI, some researchers have shown some results regarding language impairments in patients with MCI. Verbal fluency, which is individual's ability to retrieve and produce words that conform with a certain criteria within a specified amount of time, has been shown to discriminate accurately between MCI subjects and healthy controls. Cunje et al. [29] tested verbal fluency performance in patients with MCI using four categories (animals, towns, first names, fruits and vegetables) and a time window of 60 seconds, which showed good values for sensitivity and specificity for differences between the two groups. Similarly, Murphy et al. [30] found that patients with MCI-amnestic produced fewer items than healthy older adults.

Concerning word-finding difficulty, MCI patients also showed differences when compared with healthy elderly controls. Bennet et al. [31] found out that MCI patients had a lower baseline and declined more rapidly when compared to healthy controls in tests of semantic memory, however this impairment is more visible in patients with AD [32].

Syntactically, Ribeiro et al. [33] found significant impairments using a modified Portuguese version of the Token Test (the patient was requested to point to objects in the room or to carry out simple commands). A total of 33.7% of the patients with MCI participating in this study showed impairments in this task.

In conclusion, MCI has not been studied as extensively as AD and therefore there is less clinical data available, which makes more difficult to build a machine learning model to discriminate between groups.

2.2 Diagnostic

The process of getting diagnosed with almost any type of dementia is difficult and sometimes troublesome [34]. In the first place, if the person shows signs of cognitive impairments, the doctor administers a cognitive test. If the patient has a negative score, the doctor needs to rule out all other possible causes for these impairments, such as, strokes or nutritional deficiencies (lack of B12, for example). Then the doctor can order a Positron Emission Tomography scan or a Magnetic Resonance Imaging to diagnose Alzheimer's [35]. Although these tests are fairly accurate at later stages of the disease, the only way to be sure is to examine the brain tissue after the patient's death [34].

There are several cognitive tests. One of the most commonly used to screen for dementia is the Mini Mental State Examination (MMSE) [36]. MMSE is a 30-point questionnaire that is used to estimate the severity and progression of cognitive impairment. Scores in this test can be divided into four categories [36]:

- Normal cognition (24-30 points)
- Mild cognitive impairment (19-23 points)
- Moderate cognitive impairment (10-18 points)
- Severe cognitive impairment (0-9 points)

Regarding the advantages of MMSE, it does not require specialized equipment for administration, and has both validity and reliability for the diagnosis and longitudinal assessment of Alzheimer's disease [37]. However, there are disadvantages as well. First, the results are affected by demographic factors, such as age and education. Second, it lacks sensitivity for MCI and fails to adequately discriminate patients with mild Alzheimer's disease. Finally, it is insensitive to progressive changes occurring with severe Alzheimer's disease [36].

Furthermore, in our opinion, this process have problems that can be improved [34]. Firstly, it is expensive and invasive. Not all health insurances cover the exams needed to rule out dementia, and, some of these exams can be uncomfortable. Secondly, patients are only tested when they start to show signs of cognitive impairment, while research suggests that these diseases starts to damage the brain almost a decade before the first symptoms start to show up [38].

Finally, without a reliable, easy method to test dementia, it's difficult to find enough eligible patients for a clinical trial, and without these patients is hard to test new drugs in order to find the cure for these types of disease.

2.3 Dataset

Considering that the goal of the thesis was to build a classifier for the dementia disease, it is important to understand the datasets available for the task of classifying dementia disease through speech.

Despite the huge volume of research on language in health care, namely dementia, most studies are based on relatively small databases and have not originated in large collections of data [39]. In addition, studies that do collect and originate recordings and transcripts of older people are not available to other researchers [40].

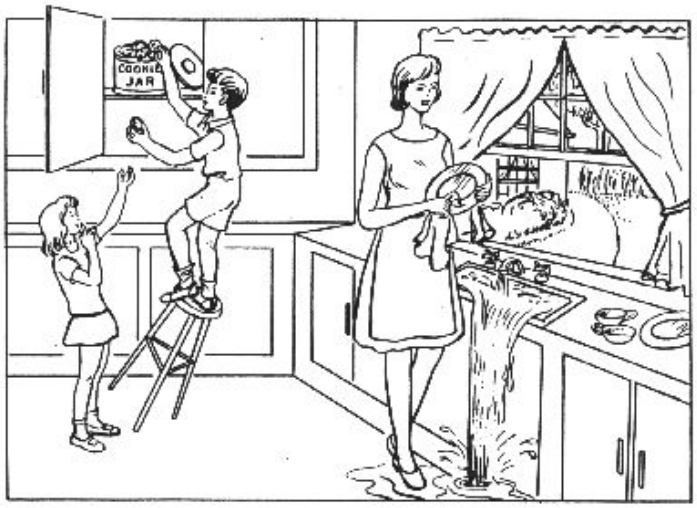

Figure 2.2: Image of the Cookie Theft test

Nevertheless, we used the publicly available DementiaBank dataset. It consists of transcripts and recording of English-speaking participants describing the ”Cookie Theft Picture”, which is an element of the Boston Diagnostic Aphasia Examination [41]. In this test, the examiner shows the picture represented in Figure 2.2 and asks the patient to tell everything happening in the picture. The examiner is permitted to encourage the subject to keep going if they do not produce a reasonable speech in terms of quantity. The answers are recorded and manually transcribed, including pauses, repetitions, errors, etc., which are then segmented into utterances. An example can be seen below:

> The scene is <in the>[/] in the kitchen. The mother is wiping dishes and the water is running on the floor. <A child is tryin(g) to get>[//] a boy is tryin(g) to get cookies outta [: out of] a jar and he's about to tip over on a stool. &uh the little girl is reacting to his falling. &uh it seems to be summer out. the window is open. The curtains are blowing. It must be a gentle breeze. There's grass outside in the garden. &uh mother's finished certain of the [/] the dishes. Kitchen's very tidy. The mother seems to have nothing in the house to eat except cookies in the cookie jar. &uh the children look to be almost about the same size. Perhaps they're twins. They're dressed for summer warm weather. The mother's in a short sleeve dress. I'll have to say it's warm.

2.4 Summary

In this chapter, I presented different types of diseases related to dementia, including their characteristics and the language impairments. More specifically, research shows that Alzheimer's Disease, which is the

most common dementia disease, impairs language and speech on semantic level and also on syntactic level. Semantically, these patients have an ”empty” discourse, incoherent, simple vocabulary and repeat ideas throughout their speech. Syntactically, although these are controversial claims, research reported that this type of patients produce fewer complex syntactic units and more syntactical errors. Furthermore, research also points to language impairments in patients with Mild Cognitive Impairment, which is a neurological disorder that involves a decline in the cognitive function with minimal impact to day-to-day activities. These impairments include difficulties in word-finding and in verbal fluency. The only procedure that is accurate in diagnosing AD is examining the brain tissue after the patient's death. Nevertheless, MMSE, which is a 30-point questionnaire that is used to estimate the severity and progression of cognitive impairment, is a cognitive test commonly used to screen dementia.

In the next Chapter, we will present and discuss previous work related to our goal, which is automatic diagnosis of dementia diseases by using spontaneous speech.

Chapter 3

Related Work

In this chapter, I describe related work on the automatic diagnosis of Alzheimer's and other types of dementia. With the recent advances in Machine Learning and Natural Language Processing, researchers have been more interested in developing models to automatically detect diseases related to dementia through discourse. This problem can be seen in two different ways, classification or regression. The review of the state of the art is therefore divided in these two categories.

3.1 Diagnosis as a Classification Problem

The common approach for detecting signs of dementia from speech is to treat it as a classification problem, usually binary classification between control subjects and patients with AD. In addition to AD, some research works have focused their attention towards MCI.

3.1.1 Alzheimer's vs Control

Kathleen Fraser et al. [13] considered 370 features of different types for capturing a wide range of linguistic phenomena for the DementiaBank dataset. The types of features range from *part-of-speech*, *syntactic complexity*, *vocabulary richness*, *information content* to *repetitiveness*. Using the top 35 features, a maximum average accuracy of 81.92% was achieved, with a recall of 81.9%. This represents the state of the art, since, to my knowledge, these are the best results obtained for this dataset. Finally, the authors argue that although speech impairment is a secondary symptom of AD, computational analysis of language may be the tool to monitor changes in cognitive status over the course of the disease, as well as responsiveness to treatments, and thus it can serve as a useful clinical tool for purposes well beyond diagnosis.

Thalia Field et al. [42] created a new set of features to detect dementia through speech corresponding to the spatial neglect. Spatial neglect is the phenomenon of reduced awareness on one side of the visual field which often occurs as a result of brain damage. Therefore, it is a common phenomenon on patients with Alzheimer and with other dementia diseases, as previous studies have shown [43, 44].

The authors divided the Cookie Theft Image, shown previously (Figure 2.2), into halves, strips and quadrants. For each division, they created a set of information units present on that division (for example, the girl is present on the left half), which then they used to calculate measures of spatial neglect. Adding these features to the ones retrieved by Kathleen Fraser et al. [13] increased the F-measure from 82.4% to 84.6%.

Another example is the work of Franz Rudzicz et al. [45], who gathered 200 features from two different datasets: DementiaBank [41] and Carolina Conversations Collections [40]. With Naïve Bayes and feature selection (top 15), the authors obtained an accuracy of 79.5% for the Carolina Conversations Collections dataset, while the best result for the DementiaBank dataset is an accuracy of 68.4%, using SVM with no feature selection. Although the authors did not analyze what was the cause of such a difference, we think this is due to overfitting when using the Carolina Conversations Collection dataset, since it is a small dataset (31 interviewees with Alzheimer and 41 controls).

Kairit Sirts et al. [46] used *idea density* (ID) to measure the rate of ideas expressed per word in a text, since it has been suggested that lower values of ID are associated with higher probability of developing AD [47]. On the DementiaBank, using these idea density measures, the authors obtained the following results: precision = 74.4%; recall = 75.0%; F-score = 72.7%. The interesting result is that the authors used a Bag-of-Words (BoW) approach as a baseline, which had results near the state of the art, in this case the work of Kathleen Fraser et al. [13], while being a much simpler approach. This observation suggests there is still room for studying simple feature sets for predicting AD.

Leandro dos Santos et al. [48] explored the use of vector-space structures for detecting dementia from speech transcripts. Besides using vector-space models, the authors also use *complex networks*. Each distinct word becomes a node and words that are adjacent in the text are connected by an edge. One problem that the authors faced was that, due to the small size of the transcripts, the networks were almost linear, providing little to none information relevant to the classification. Therefore, to address this challenge, the authors used word embeddings to enrich the networks with semantic information. The authors took advantage of this structure to build new edges between words. A new edge is added to the network if the vectors correspondent to the words have a cosine similarity (given by the proximity of the vectors) higher than a threshold. For a better understanding, Figure 3.1 presents an example of these networks. Since the authors are using networks, it is possible to have completely different features representing the topology of the network. Using these type of features and ensemble learning, the authors achieved an accuracy of 65%.

In addition, Maria Yancheva and Frank Rudzicz [19] produced another study based on the use of word embeddings for the detection of Alzheimer's through speech. The authors argue that previous studies of topic models for detecting Alzheimer's disease are not reliable, since they use a human written predefined list of annotated topics. These lists are subjective, as investigators use a different number of topics for the same picture [49–52] and are not exhaustive nor reflective of the content unit. Therefore, the authors present a generalizable method for generating automatically information content units (topics) for any given picture. The method is based on topic modelling using clusters of global word-vector representations. The word-vector model is first trained on a large normative general-purpose corpus,

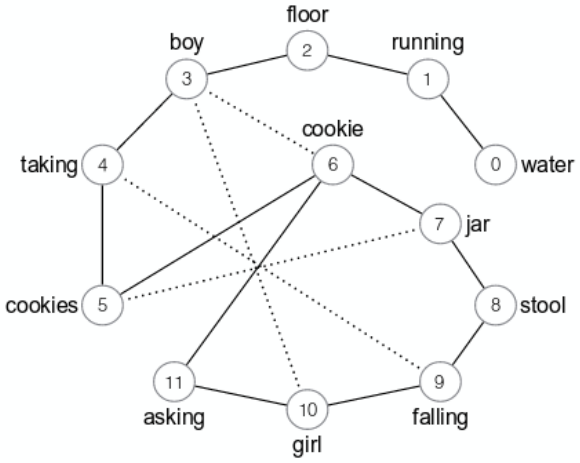

Figure 3.1: Example of a co-occurrence network enriched with semantic information [48].

which allows to avoid the sparsity of the available clinical datasets. This model is, then, augmented with the vector representations of words presented in the dataset in study, the DementiaBank. Using this technique, there was a significant overlap in the topics discussed between control and dementia groups. Thus, the authors investigated whether the overlapping topics are discussed in the same context. With this objective in mind, the word vectors were augmented with local context windows, i.e., each word vector was constructed using a linear combination of its original vector and the vectors of the surrounding context words. Using these methods, the authors reached some conclusions. Firstly, all of the control word clusters are used in the same contexts by both healthy participants and those with dementia. However, while the two groups discuss the same topics and use the same words in the same contexts, not all dementia participants identified all of the control topics or discussed them with the same frequency. The authors extracted the features from these clusters and used a Random Forest with 10-fold cross-validation to classify subjects, which matched the state-of-art results (F-score of 0.8).

3.1.2 Mild Cognitive Impairment vs Control

Although researchers have focused more on detecting AD, there are some that looked at automatically detecting MCI, a harder task than detecting AD, since there is less data available and MCI is less symptomatic than AD.

Roark et al. [53] showed the viability of this task by using transcripts and recordings of patients subjected to the Wechsler Logical Memory I/II test. In this test, the subject has to retell a story, once immediately after the story has been told and another after a delay of 30 minutes. The authors extracted a set of features that represented the complexity of the narrative, and another set that symbolize the duration of the speech (i.e. number of pauses, pause length, among others). Using the Area Under the ROC Curve (AUC) as measure of the performance, the authors obtained a value of 0.74 using SVMs as the classifier, which showed that NLP techniques can help to automatically distinguish between healthy

controls and MCI patients.

In addition, Vaden Masrani et al. [54] tackled this task using the DementiaBank as dataset. In this dataset, only 43 patients were classified as MCI and 256 as possible/probable AD. Therefore, the authors used the data from Alzheimer's patients to augment the MCI dataset. The authors implemented two simple methods for domain adaptation; AUGMENT [55] which has been shown to be effective on a wide range of datasets, and CORAL [56], which works by aligning the covariances of the source and target features. Furthermore, the authors argue that there is a measure of coherence that has been neglected in previous studies, which is *discourse analysis*, which corresponds to features that represents the coherence between sentences. These features were added to the features retrieved by Kathleen Fraser et al. in [13]. The main positive result achieved by the authors is that domain adaptation does help with the task of detecting MCI. Overall, the best approach is the AUGMENT adaptation without the discourse features (F-measure = 0.712). Surprisingly, using Alzheimer-only data for the training (F-measure = 0.681) outperforms the use of MCI-only data for the training (F-measure = 0.640). In addition, the discourse features did not improve the performance of any approach. Concluding this work, the authors mentioned that the lack of data is a major obstacle to develop a tool to diagnose dementia, more especially MCI, from speech.

Finally, other authors have had success in this task in other languages besides English. Tóth et al. [57] instead of using the manual transcripts, used the recordings combined to an automatic speech recognition to automatically identify acoustic features. For Hungarian patients, although manual transcripts had better results, the authors reach a F-measure of 85.3%. Similar results have been found on French and Greek patients [58, 59]

3.2 Diagnosis as a Regression Problem

Neurocognitive diseases that provoke dementia are usually complex and vary in their symptoms from person to person. For that reason, it might be more useful to predict scores than trying to predict a diagnosis (dement vs healthy). However, it seems that there has been very little work on prediction of clinical scores using speech when compared to the work made to classify patients.

Maria Yancheva et al. [60] used a set of 477 features in order to predict MMSE scores. The data used by the authors is from the DementiaBank dataset. In order to model the longitudinal progression of the MMSE scores and the features, the authors constructed a dynamic Bayes Network with continuous nodes. A bayesian network is a probabilistic model that represents a set of random variables and their conditional dependencies via a directed acyclic graph model, which relates variables to each other over adjacent time steps [61]. In their research, each time slice (Q_t, Y_T) represents an annual visit for a subject. Each node Q_t represents the MMSE score for that visit, while Y_t is the vector of features. Performance was measured as the Mean Absolute Error (MAE) between the actual and predicted MMSE scores. Since not all subjects have the same number of samples, the error is measured at the first and last nodes and averaged. The lowest MAE is achieved when selecting the top 40 features (MAE of 3.83). However, when focusing on individuals with more samples, the MAE improved to 2.91. This

study, among other things, reinforces the importance of longitudinal data collection.

Nicklas Linz et al. [62] used regression models to predict MMSE scores based on person's performance on the Semantic Verbal Fluency (SVF) task. In this task, the participants are asked to name as many members of a semantic category as possible in a given time period. The data used for this study was collected during the Dem@Care [63] and ELEMENT [64] projects. Furthermore, from the performance of the SVF task, the authors extracted several features, such as word count and semantic distance, among others. Over a set of 179 patients with different degrees of dementia, the authors achieved a MAE of 2.2.

Nevertheless, this type of approach has received more attention in the image processing community and multiple authors are using brain imaging features (average regional grey matter, tissue volume of MRI, etc.) to predict clinical scores [65, 66]. Lei Huang et al. [67] based on imaging data extracted from MRI, used a Random Forest Regressor to predict clinical scores, including MMSE. This leads to the best mean absolute error of 1.68 for MMSE.

3.3 Discussion

First, there are few studies focused on the detection and classification of Alzheimer's and related dementia diseases using mainly audio features and perhaps it can still be an area of study with possible developments. However, it does not seem the best approach for the goal in mind. On one hand, it demands a good recording protocol, so the retrieval of features can occur easily. On the other hand, different studies showed that audio features, when compared with the ones retrieved from text transcripts, do not provide any information gain [13] [60].

Furthermore, we saw that studies focusing on text features produce the best results. Without a doubt, the work of Kathleen Fraser et al. [13] is the state of the art for detecting Alzheimer's. Nevertheless it is really a complex approach to the problem. As Kairit Sirts et al. [46] mentioned, really simple approaches, such as bag-of-words, produced similar results. In addition, we believe that better results can be achieved using different representations of the data available. For example, the classifier from the work of Maria Yancheva et al. [19], that used a vector space representation, reached almost the same performance as the classifier of Kathleen Fraser et al. [13].

Regarding the studies that used vector-space representations, it is possible to conclude that is an area understudied for the diagnosis of Alzheimer's and dementia in general, since we only found three different studies using this approach. As shown by Fraser et al. [68] and Maria Yancheva et al. [19] these models can add new interesting features that can improve the performance of a classifier.

Concerning the problem of having small datasets, we saw two different studies tackling the issue with different approaches. Foremost, Maria Yancheva et al. [19] used a large general purpose corpus to train the word-vector model and then, augmented it with the vector representations of the words from the DementiaBank dataset. Vaden Masrani et al. [54] has another approach to the issue, which is domain adaptation. Both approaches showed an improvement in the results.

In addition, with some exceptions, most of the work relies in manual annotations or on manual work.

This means that if an application was built based on these models, it would have to rely on clinicians to annotate the transcripts. Therefore, the application could not provide a feedback to the clinician in real-time and would give the clinician extra work, instead of helping him. Thus, a model that could automatically classify the patient without any help of annotations or contexts, would be better in the real world, in our opinion, even if it had worse results than the state-of-art.

Finally, the datasets used are either small or not co-aligned in terms of age or education, which makes the results less reliable. Therefore, there is the need for more longitudinal studies gathering more recordings that can be used as datasets for future work.

3.4 Summary

We presented several studies that try to automatically diagnose patients with dementia. Fraser et al. [13] is the state of the art when classifying between AD patients and control subjects, using the DementiaBank dataset. The authors achieved an average accuracy of 82% by using a top 35 linguistic features. Thalia Field et al. [42] added a set of spatial features into the set of features used by Fraser et al. [13], which increased the performance of the model by 2%.

Less research has been done regarding the classification between MCI patients and healthy subjects. Vaden Masrani et al. [54] achieved a F-measure of 71.2% by augmenting the data of the MCI patients in the DementiaBank dataset with the data of the subjects with AD.

In addition, other researchers have treated this problem as a regression one. Maria Yancheva et al. [60] used a set of 477 features in order to predict MMSE scores of the patients in the DementiaBank dataset. By using a dynamic Bayes Network with continuous nodes, the authors reached a MAE of 2.91.

Finally, we argued that most of the work rely in manual annotation or manual work, which rely on clinicians to annotate future corpus. Furthermore, the datasets used are small, which makes the results less reliable and increases the probability of overfitting. Unfortunately, these are the only datasets publicly available.

In the next chapter, we discuss the methodology common to all experiments, including dataset used, preprocessing, models used and how the evaluation is performed.

Chapter 4

Methodology

This chapter describes the methodology that is common to all experiments; the dataset and its preprocessing, the general processing pipeline, the models used and the method of evaluation for the models. Any differences from the general procedure explained here is discussed within each chapter.

4.1 Characterization of the DementiaBank dataset

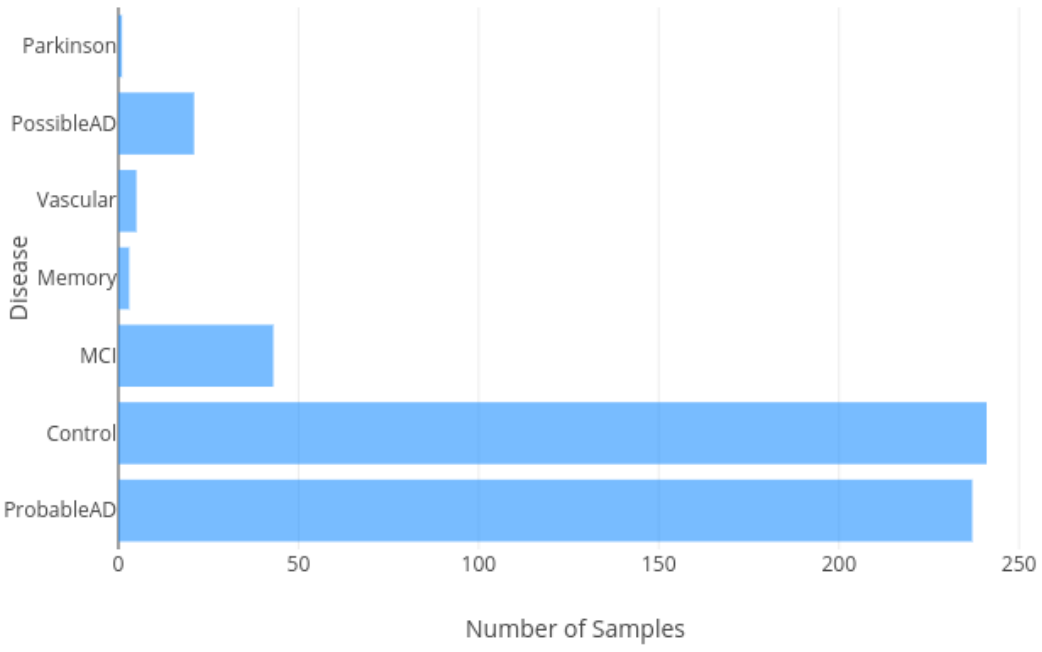

Figure 4.1: Number of samples per disease type or control in DementiaBank

DementiaBank consists of 241 samples from 98 healthy controls and 310 samples from 195 patients with dementia. Furthermore, of the 310 interviews with patients, 43 samples were classified as mild cognitive impairment and 258 samples as possible/probable Alzheimer's, as shown on the Figure 4.1. According to the study, a sample is classified as "possible AD" if the patient shows signs of Alzheimer's

plus something else. In addition, the label 'Memory' corresponds to samples of patients that show signs of MCI plus memory impairments.

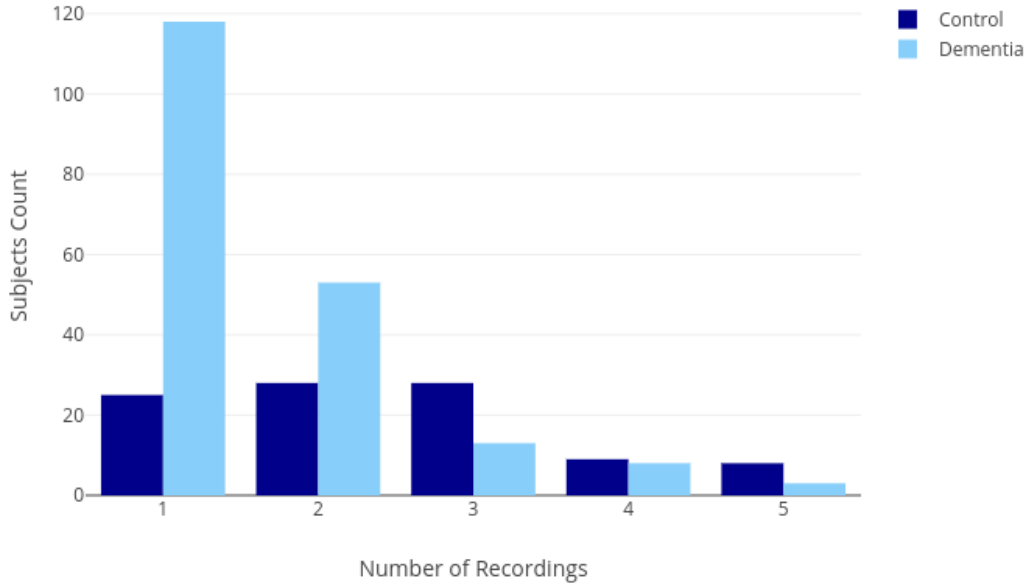

Figure 4.2: Number of recordings per subject in DementiaBank. For example, the bar corresponding to the value "2" means the amount of subjects that have 2 interviews.

Regarding the distribution of samples per patients/controls, it is possible to observe in the Figure 4.2 that although we have less controls, they have in average more interviews than dementia patients. In addition, regarding the distribution of age between the two groups, it is possible to see in Figure 4.3 that there is a misalignment between them. This can influence results analysis, since the differences found in the discourse can be due to the difference of age and not due to the patient having dementia or not. However, it can also be seen as a feature that can help machine learning models to discern between the two groups.

Similarly, it is possible to see in Figure 4.4 that there is also a difference in the average word length between the two groups. Those with dementia are more likely to use shorter words. Therefore, this characteristic seems to be a good feature to use to classify patients.

Other characteristics present in the descriptions were studied, such as, number of words in the transcript, number of stopwords, number of nouns, number of verbs, etc. These features were based on the 370 features used on the work of Kathleen Fraser et al. [13]. However, their distributions were similar between the two groups, which suggests that they are not good characteristics to separate the two groups.

Figure 4.5 represents the distribution of MMSE score among the different types of patients. Obviously, control and MCI subjects have higher scores in a cognitive test when compared to AD patients. The dotted line in black represents the threshold for considering the patient as having some mental health decline according to the test.

According to this image, even if we implement a Regressor that perfectly predicts the MMSE score, if

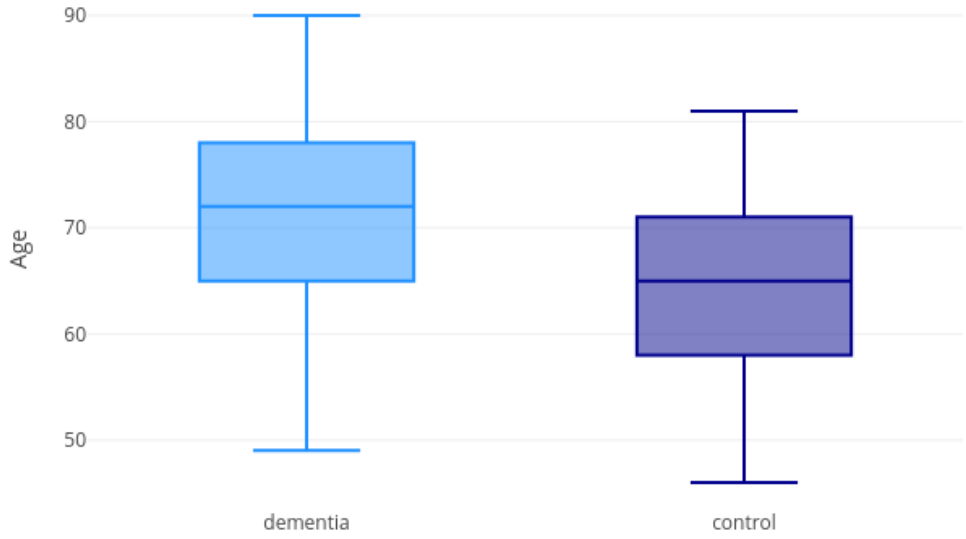

Figure 4.3: Boxplot representing the age distribution between the two groups

we try using the MMSE scores to classify subjects, we will error in some cases. According to the MMSE test, if the score is between 24 and 30, the subject presents normal cognition. If we consider the value 24 as our threshold (dotted line in Figure 4.5) for deciding if a patient is dement or not, all patients with MCI and Memory (MCI with memory problems) are going to be classified as being healthy. In fact, some authors have already presented that the MMSE test is unable to distinguish between healthy controls and MCI subjects [69]. Furthermore, if we try this approach, 25% of Probable AD and 25% Possible AD will be misclassified, which is a problem again. Nevertheless, we are still going to try this approach, since most patients with dementia are below the MMSE score of 24 if we remove the MCI patients.

4.1.1 Critique

The DementiaBank dataset presented is far from perfect. The biggest problem comes from the fact that the dataset is, without a doubt, small. First of all, its size makes overfitting almost impossible to avoid, and makes outliers and noise have a large impact on the data. Furthermore, and more importantly, its size is not compatible with the use of some machine learning algorithms. For instance, deep learning methods, such as neural networks, are probably not possible to use for the classifier, at least not directly. This means that the possible approaches are limited to the conventional methods of machine learning and statistics. In addition, it is not possible to merge the dataset with others in order to create a bigger one, since, unfortunately, the transcripts produced by the patients in this dataset are very specific to one task that is not used in other datasets.

Furthermore, when gathering the data, the researchers did not take into account the characteristics of the patients. For example, regarding age, Figure 4.3 shows the mismatch in age between the two groups. In addition, the two groups are, also, not matched in terms of education, for example.

Finally, having just this dataset available really emphasizes the need for more longitudinal studies

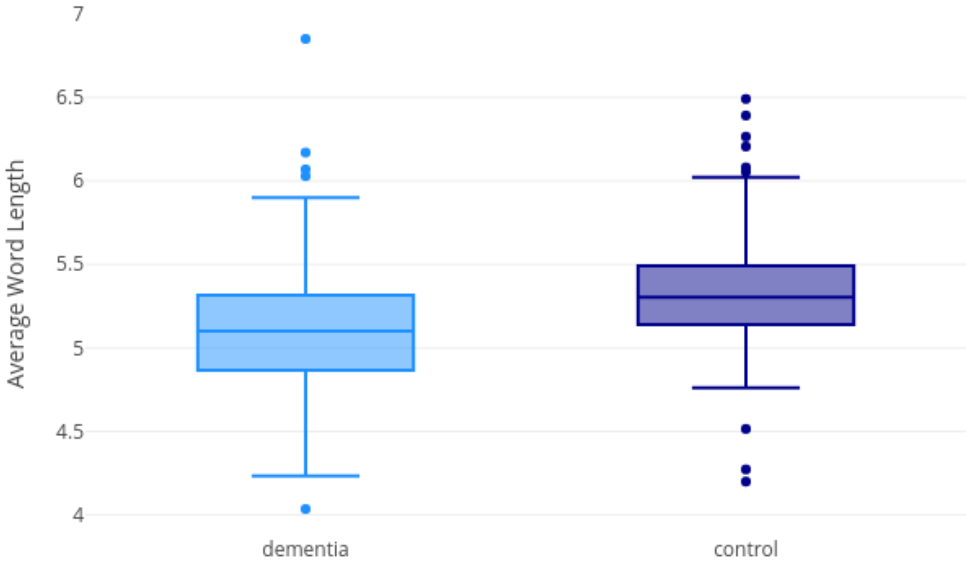

Figure 4.4: Boxplot representing the average word length distribution between the two groups

with the purpose of gathering recordings and information of patients with dementia. Fortunately, at the moment, studies are being conducted, such as, the Talk2Me[1] project from the University of Toronto. Unfortunately, it will take some time to be complete, which means that the datasets gathered in these studies were not be available in time to be used in this thesis.

4.2 Preprocessing

As with any dataset, before start building models, we need to preprocess the information in order to obtain it within the format that we want.

Most of the preivous work relied on the human annotations of the transcripts for their models. Since someone has annotated the transcripts before feeding it, the authors could not implement a system that retrieved the discourse of one person and give the result in real-time. In fact, the only application we found that automatically classifies discourse transcripts is made by WinterLights Labs[2], which relies on the work of Kathleen Fraser et al. [13]. However, their application is not commercially available, therefore we could not test it. Although it is true that the transcripts used for training are mannually transcribed, we believe that we could obtain a transcript of the same quality by using an automatic speech to text system.

For that reason, all the approaches mentioned in this thesis were built under the assumption that the system should be capable to process all transcripts automatically and give the result in real-time. This means that the model can not rely in human annotations. Hence, the transcripts in the dataset are preprocessed to obtain the transcripts in their raw format. We remove all human annotations except for the completion of words (”tryin(g)” $\rightarrow$ ”trying”), since an automatic speech-to-text it will always yield

[1] https://www.cs.toronto.edu/talk2me/about/
[2] http://www.winterlightlabs.com/index.html

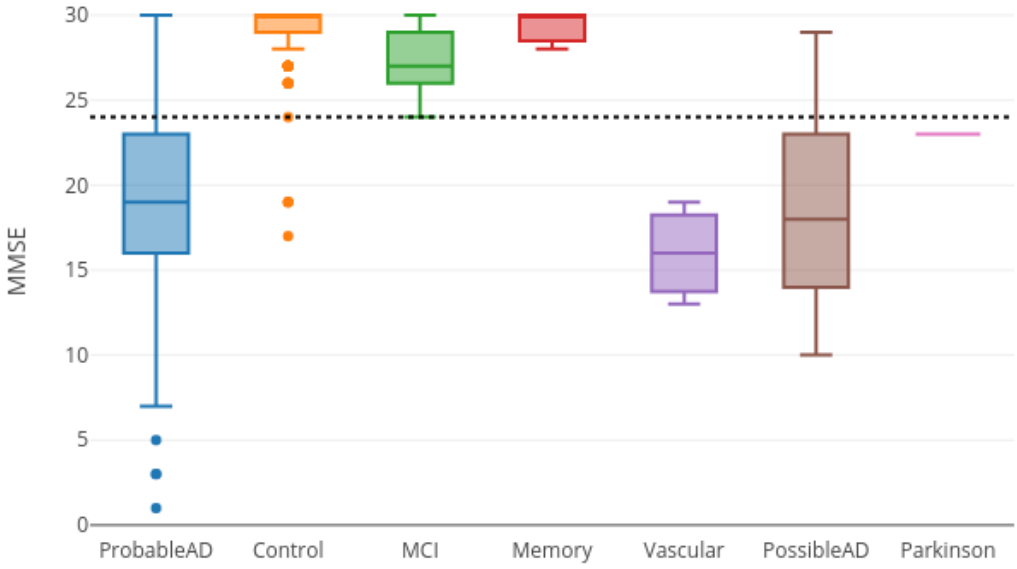

Figure 4.5: Boxplot representing the MMSE between dementia disease types

”trying” instead of ”tryin”. All text is converted to lower case. As result, the transcript quoted on Section 4.1 becomes the following text:

> the scene is in the in the kitchen. the mother is wiping dishes and the water is running on the floor. a child is trying to get a boy is trying to get cookies outta a jar and he’s about to tip over on a stool. uh the little girl is reacting to his falling. uh it seems to be summer out. the window is open. the curtains are blowing. it must be a gentle breeze. there’s grass outside in the garden. uh mother’s finished certain of the the dishes. kitchen’s very tidy. the mother seems to have nothing in the house to eat except cookies in the cookie jar. uh the children look to be almost about the same size. perhaps they’re twins. they’re dressed for summer warm weather. the mother’s in a short sleeve dress. i’ll have to say it’s warm.

We could still process the transcripts, for example, by stemming the words or by applying lemmatization to them. Both processes share the goal of getting the root of the word, or more formally, reduce inflectional forms of a word to a common base form [70], but they act differently. Whereas stemming follows an heuristic approach that drops the suffix of words in order to get closer to the common base form, lemmatization uses a dictionary and morphological analysis of words to return the base form of words, known as lemma. If shown the word ”saw”, stemming would tend to return only s, while lemmatization would take into account if the word was the verb or the noun, and correspondingly, return see or saw as the base form of the word. These could reduce the dimension of bag-of-words approaches, however this will stay as future work.

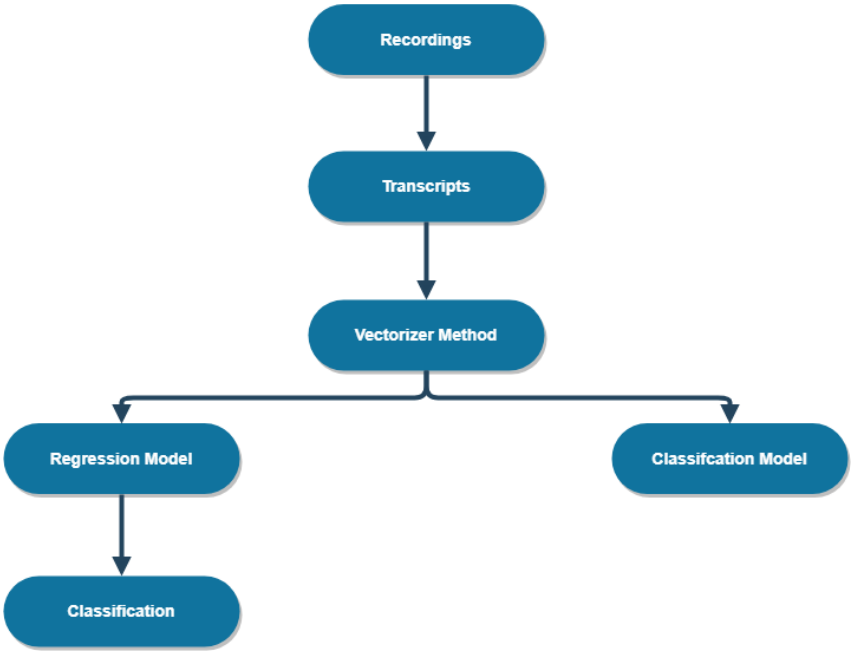

Figure 4.6: Processing pipeline from clinical interview to evaluation

4.3 General method

The general method for any approach is shown in the Figure 4.6. First, from each recording of the interview, we get the respective transcript (this has already been done). Afterwards, before feeding this transcript to the model used, it is needed to vectorize it. The idea here is to get a matrix, where each row is a transcript, and each column is a feature, thus each cell is the weight of the feature on the transcript of that row.

The matrix is fed to the models to train them. The models used differ depending if we want to treat the problem as a classification or a regression problem. Finally, in the case of using a Regression Model, the scores obtained by the model can be used to classify the subjects. Usually, we will try to predict the MMSE score using a regression model, and use this score to classify the patient as healthy or not.

4.4 Models

We used several models from the scikit-learn package[3] either for classification or regression. Usually, only traditional models were considered due to the size of the dataset. The reader is presumed to have some familiarity with traditional Machine Learning models (Logistic Regression, Random Forests, Support Vector Machines (SVMS), etc) and therefore the details of each model will not be discussed[4]. The hyperparameters of these models are obtained using a Random Search algorithm.

Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be the fitness function that must be maximized, in this case, for example, the accuracy function. Let $x \in \mathbb{R}^n$ designate a candidate solution in the search-space, in this case a set of

[3]http://scikit-learn.org/stable/

[4]Those unfamiliar with these kind of models, I recommend the blogpost: https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/

specific parameters in all the search-space of the parameters. The Random Search algorithm can be described as follow:

1. Initialize x with a random position in the search-space
2. Until a termination criterion is met (number of iterations chosen), repeat the following:
   (a) Sample a new position y from the hypersphere of a given radius surrounding the current position x
   (b) If $f(y) > f(x)$, then move to the new position by setting $x = y$
3. Return the best setting x

4.5 Evaluation

A 10-fold cross validation procedure was used to evaluate different models. However, since patients have more than one interview, we guaranteed that these interviews were in the same train or test fold.

For classification models, we used accuracy, F-measure and AUC to report the results for each approach. For regression models, we used MAE, since it is easier to interpret.

4.6 Summary

In this chapter, I presented the methodology common to all experiments. Foremost, I have used the DementiaBank for all experiments. In this dataset, although there are less control subjects, they have in average more interviews than dementia patients. Furthermore, there is a mismatch between the age and average word length between the two groups, which can influence the results. The MMSE of the subjects in this dataset is also a concern, since there are control subjets with a score below 24 and dementia patients with scores above 24, which is the threshold to classify a patient as having a cognitive impairment.

Regarding the preprocessing of the transcripts, I removed all annotations, in order to build a model that does not need any human work to properly function. This will allow building, in the future, applications around these models that can predict in real-time the probability of a patient having dementia.

The general method for the approaches is simple. We transform each transcript into a vector and feed it into either a classification or regression model. The models used are present in the scikit-learn package. Finally, I have used a cross-validation strategy with different metrics to evaluate the performance of these models.

In the next chapter, I will present two different approaches that automatically classify transcripts of spontaneous speech as being produced by patients with dementia or not.

Chapter 5

Diagnosis as Classification

In this chapter, I propose and evaluate two approaches to identify patients with dementia using speech: a bag-of-words approach using Term Frequency–Inverse Document Frequency (TF-IDF) and other using word embeddings, described in sections 5.1 and 5.2, respectively.

5.1 Bag-of-words approach

Following the work of Kairit Sirts et al. [46], where the authors showed that a simple bag-of-words approach had similar results to the state of art, we decided to build a simple model using TF-IDF as our baseline. TF-IDF is a numerical value that reflects how important a word is to the document [71]. This value is calculated in our experiment as follows:

$$tfidf_{t,d} = \log_2(1 + tf_{t,d}) * \log_2(\frac{N}{df_t}) \tag{5.1}$$

where t and d are the term and document for which we are computing a feature, $tf_{t,d}$ is the term frequency of term t in document d, N is the total number of documents we have, while df_t is the number of documents that contain t.

By applying TF-IDF to our transcripts, we obtain a sparse matrix (most of the elements are zero) with 551 rows corresponding to our transcripts, and 2000 columns corresponding to the unique words in the transcripts. Each element $m_{i,j}$ in the matrix m is the TF-IDF value for the transcript in the row i for the word in the column j.

Figure 5.1 represents the Accuracy, Area Under Curve (AUC) and F-measure across a range of models, using the TF-IDF technique explained earlier. Logistic Regresison and Support Vector Machines had similar results (Accuracy = 0.76, AUC = 0.76, F1 = 0.74). Although Random Forests had been reported as having worse results when dealing with high dimensionality matrices, they were close to the best performance in terms of results using this approach. The KNN model had clearly the worst results by a long margin.

Although we did not perform any kind of word selection, we still got promising results when compared to the state-of-art, which made us think that it would be worth to improve this approach.

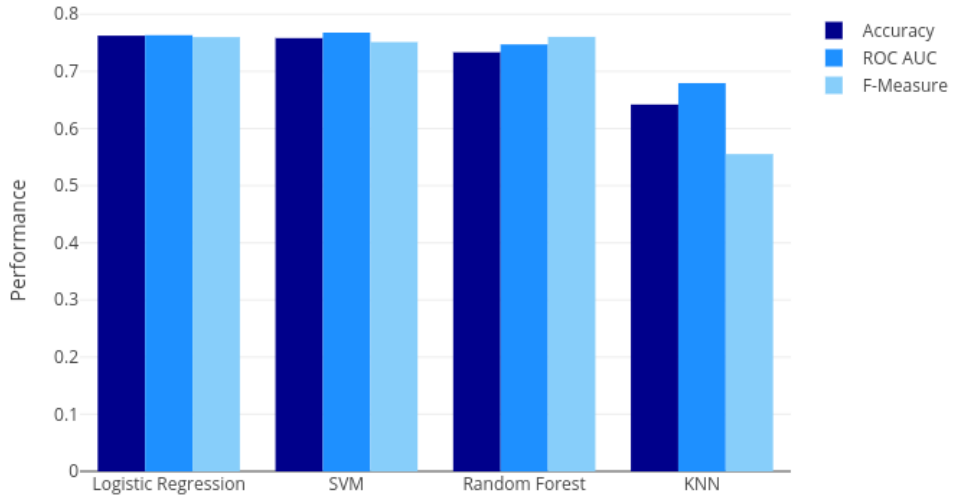

Figure 5.1: F-measure, Accuracy, and AUC for each model using TF-IDF approach

5.1.1 Word selection

As discussed previously, applying TF-IDF to our matrix gives us a matrix with several features with a weight value of 0. For example, we expect stop words like ”the”, ”and”, etc. to always appear in the transcripts no matter the label. Therefore, these kind of words will almost always have a value of 0 in all transcripts. Furthermore, since the transcripts correspond to an image description task, several words (e.g. ”girl”, ”boy”, ”cookie”) will appear frequently in several transcripts, which means that training the models using all words is not optimal, considering that we are fitting our models to words that have zero meaning to them.

For this reason, it would be better to train our models using only the words that appear less often and appear or in the control transcripts or in the dementia transcripts. To achieve this, we used the χ^2 formula.

The χ^2 formula measures how much expected counts and observed counts of variables/distributions deviate from each other. It can be used to test for Independence between two variables, like defined by the equation below, where $O_{x_1x_2}$ is the observation of the conjunction of variables and $E_{x_1x_2}$ the corresponding expected value, this is, the expected value given our hypothesis that the variables are independent.

$$\chi^2 = \sum \frac{(O_{x_1x_2} - E_{x_1x_2})^2}{E_{x_1x_2}} \tag{5.2}$$

In this case, for feature selection, we want to test the independence of our features from the class labels. In our particular case, we define $x_1 = w$ as our word and $x_2 = c$ as our class label (dementia or control). A small chi-squared value will mean that the term is closer to independence from the class, and a bigger value means that it is dependent on the class, which is exactly what we want.

Figure 5.2 represents the top 30 words with higher χ^2 scores. Clearly, the word ”action” has the highest χ^2 value. This word appears in 68 control transcripts, however only appears in 17 dementia transcripts. Normally, this word is used in the beginning of the description of the image, as a repetition

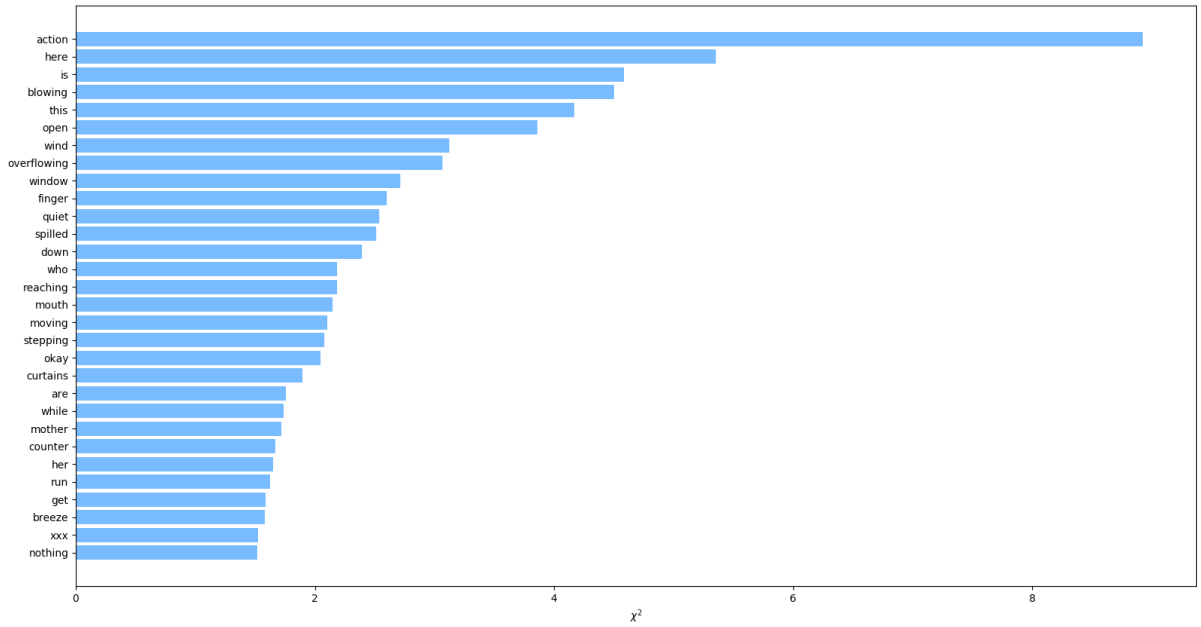

Figure 5.2: χ^2 values for the top 30 words

of the task they need to do (e.g. ”all of the action you see going on. okay”). While control patients usually use ”overflowing” (101 transcripts of control, 53 transcripts of dementia) to describe the water overflowing in the sink, dementia patients normally use ”spilled” to describe this event (1 transcript of control, 22 transcript of dementia). Imperceptible words, represented as ”xxx” in the transcripts happens more in dementia transcripts than in control transcripts, as one would expect. Finally, the word ”here” appears in 40 dementia transcripts an 10 control transcripts. At first glance, this seemed strange, since ”here” does not describe anything in the picture. By analyzing the transcripts that contained this word, it seems that dementia patient use ”here” like a pointer to the image, and tend to repeat the word throughout the recording, for example:

> ”well I see this woman **here**. (...) and **here** she’s standing. and she’s got some kind of a stick I would say **here**. (...) should I still stay on this side **here** or go over **here**? (...) did did we mention this other person **here**? (...) then **here**’s a girl. and she’s got some sticks **here**. and **here**’s a boy evidently.”

Thus, these results suggest that the use of the χ^2 can be helpful to obtain the words that can help discern between the classes. Therefore, we used our model and trained and evaluated them using only the top k words with higher χ^2 values.

The results of this approach are represented in the Figure 5.3. By analyzing these results, it is possible to see that the χ^2 approach helps us get better results. For almost every model, the accuracy peak is reached way before the model reaches the maximum of features.

In fact, if we analyze the Figure 5.4, which represents the accuracy, AUC and F1 values for each model at the optimum number of features, it is possible to see that by using this approach we increase our accuracy results by 8.2 on average. However, this major increase is mainly due to the KNN model

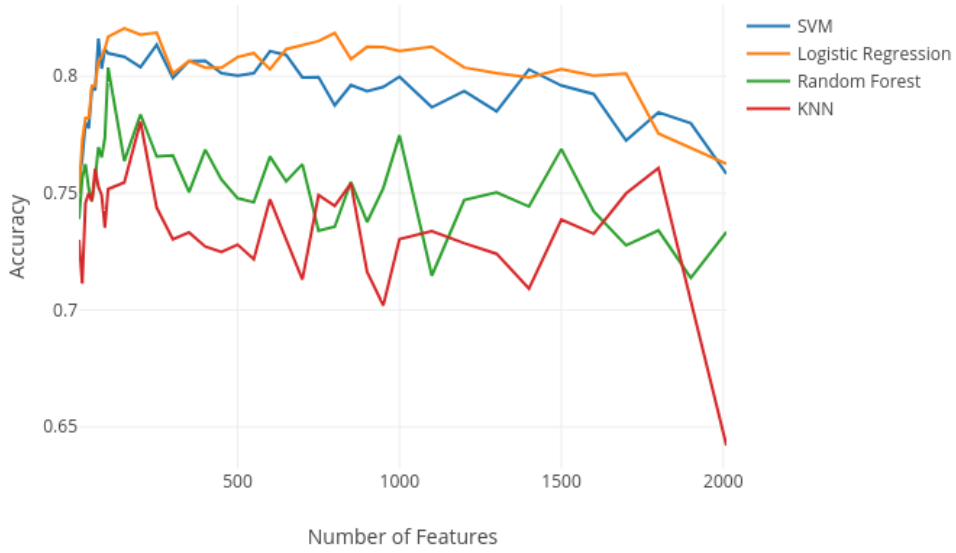

Figure 5.3: Accuracy for different models as we vary the number of words included

went from a accuracy score of 64% to 78%. Nevertheless, the other models also had a increase in performance. The best model was once again the Logistic Regression Classifier (Accuracy = 0.820, AUC = 0.817, F1 = 0.813), with SVM being a close second. Also, once again, Random Forest had results similar to the previous one. These results are similar to the results of the state-of-art.

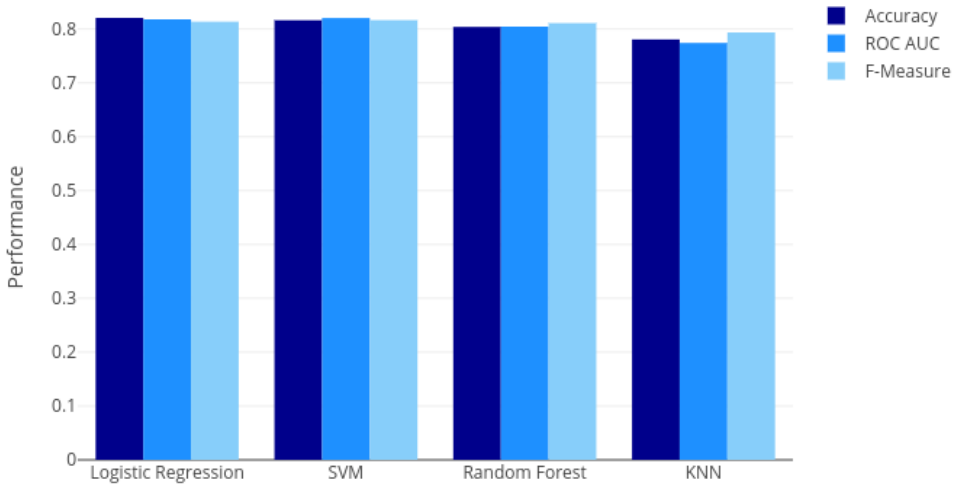

Figure 5.4: Accuracy, AUC and F-measure for each model at their optimum number of words (i.e. the peak performance in Figure 5.3)

In addition, Figure 5.5 represents the confusion matrix for the best model, in this case, the Logistic Regression Classifier. If we consider an scenario in which a tool as the one proposed is used as a preliminary screening tool, it is preferable to have an higher false positive rate than a false negative one, since misclassified healthy patients will be further examined using other more accurate diagnostic methods. Unfortunately, we obtain a higher rate of false negatives than false positives. While 16% of the control transcripts were classified as being a transcript produced by patients with dementia, 19% of the transcripts produced by dement subjects were classified as being a control transcript. Nevertheless,

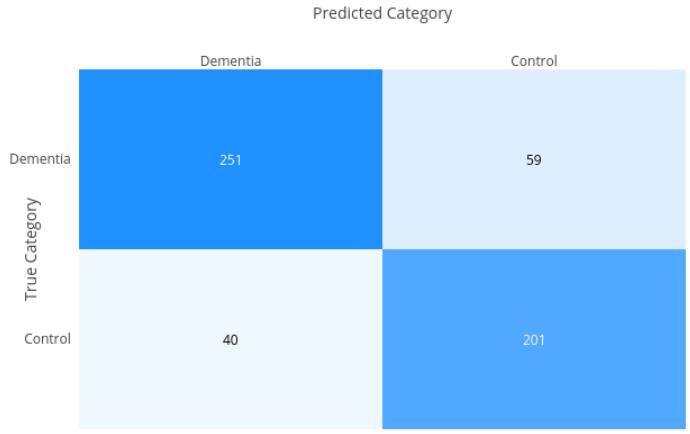

Figure 5.5: Confusion matrix for the Logistic Regression model

since no other related work reported the confusion matrix, we cannot tell if this approach is a beneficial or not in this regard.

Concluding, by using a simpler approach, we obtained results similar to the state-of-art. Furthermore, previous work only focused on ”Alzheimer’s vs Control”, while in this approach we focused on ”Dementia vs Control”, which is a harder task, since language impairments are not same for all types of dementia.

5.1.2 Addition of Age and Mean Word Length as Features

The previous results were obtained using only the TF-IDF matrix. However, other features could be combined in an attempt to improve our results. Previously, in Chapter 4, we concluded that both the age and the mean average word length were good features to distinguish between control subjects and patients with dementia.

Scikit-learn provides an object called FeatureUnion that allows the user to parallelize the workflows and have several transformations applied in parallel to the dataset. This allows us to transform our transcripts into a matrix using Tf-IDF, while at the same time standardize our other two features and concatenate the results in order to then feed it to our classifier.

Both features are standardized by removing the mean and scaling to unit variance, as shown on the following equation, in this case for the age standardization:

$$age_i = \frac{age_i - mean(age)}{stdev(age)} \tag{5.3}$$

where age_i is the age of the subject i and age is a list with all the ages of the subjects. This process is similar for the standardization of the average word length.

Figure 5.6 represents the accuracy performance of the Logistic Regression model with or without the additional features, while varying the number of words used for the Tf-IDF matrix. We only show the Logistic Regression model, since we showed previously that it was the model with the best accuracy performance. After analysing the figure, we can conclude that adding these features provides little gain to our models, since the peak performance of our models are very similar. Second, when the number

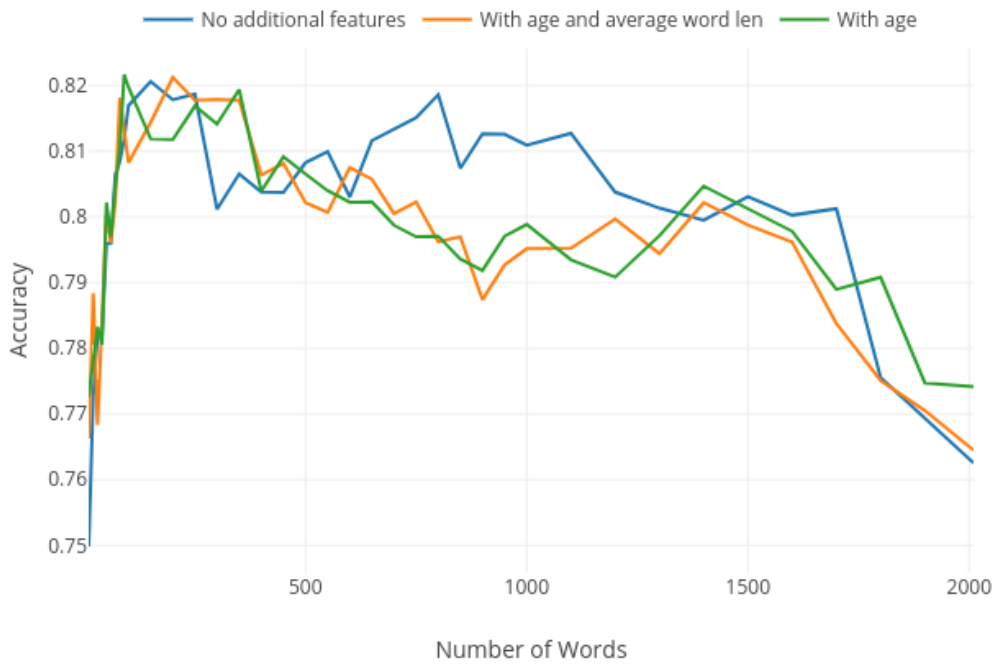

Figure 5.6: Accuracy for different combinations of features as we vary the number of words included

of words used for the Tf-IDF matrix is low, the model that use additional features outperforms the model studied previously. However, when the number of words increases, the previous model starts to have better performance when compared to the other two. Finally, it is also possible to conclude that all models have a significant decrease in performance when the number of words used reach the maximum value (around 2000 words), which is similar to what we saw and discussed previously.

Figure 5.7 presents the peak performance for the three models mentioned in the previous paragraph. First, this confirms that the three models have almost equal peak performance. The model that uses the Tf-IDF matrix in combination with the age of the subject had the best performance with an accuracy of 82,16%, +0.04% more than the model that uses both features and +0.1% than the model that does not use any of the new features. Although, at first glance, these two features seemed to be able to distinguish the two types of subjects, it is possible to conclude that these features are not meaningful to our models, since the increase is near 0 in terms of accuracy performance.

Figure 5.8 shows the confusion matrix for the Logistic Regression classifier using both the Tf-IDF matrix and the age feature. When compared to the Figure 5.5, it's possible to conclude that the increase in accuracy is due to 5 dement transcripts that previously were classified as being healthy, were now classified as being transcripts produced by patients with dementia.

However, this model has also a decrease in the control subjects classified correctly, since 3 control subjects that were previously correctly identified as such, are now classified by the new model as having dementia.

In conclusion, although it appeared that this approach would increase our performance, in truth, it did neither improve nor deteriorate our models. Nevertheless, with more time, other features could be studied that might improve the results.

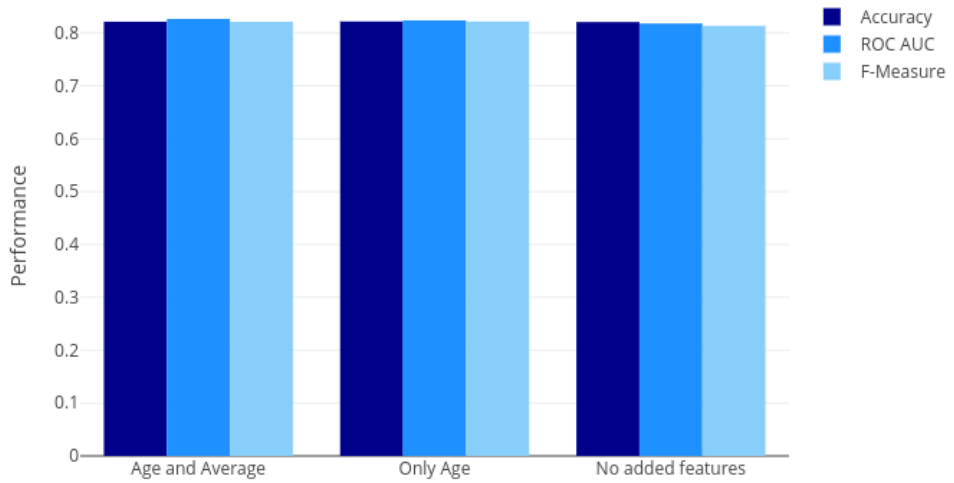

Figure 5.7: Accuracy, AUC and F-measure for each feature combination at their optimum number of words (i.e. the peak performance in Figure 5.6)

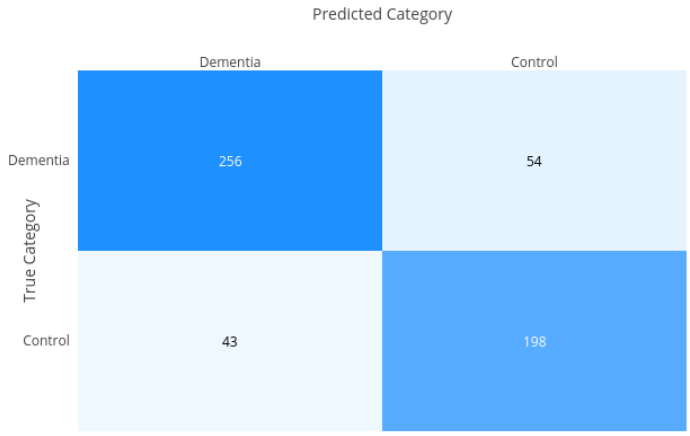

Figure 5.8: Confusion matrix for the Logistic Regression classifier using the Age feature

5.1.3 Alzheimer's vs Control

Until now, we have only focused on Dementia vs Control in our classification. However, since language impairments and their degree are different between the different dementia diseases, one can argue that it makes more sense to classify between a disease that causes dementia (like AD) and control. Therefore, we used the previous approach to do classification between AD patients, since this is the most predominant disease in our dataset, and Control. In our dataset, in total, we have 258 transcripts of patients with AD and 241 transcripts of control subjects.

Figure 5.9 represents the Logistic Regression classifier using the TF-IDF approach varying the number of words used using the χ^2 test, with or without the additional features (age and average word length). We only show Logistic Regression classifier in the Figure for two reasons:

1. It is the best classifier for this approach, as seen previously;

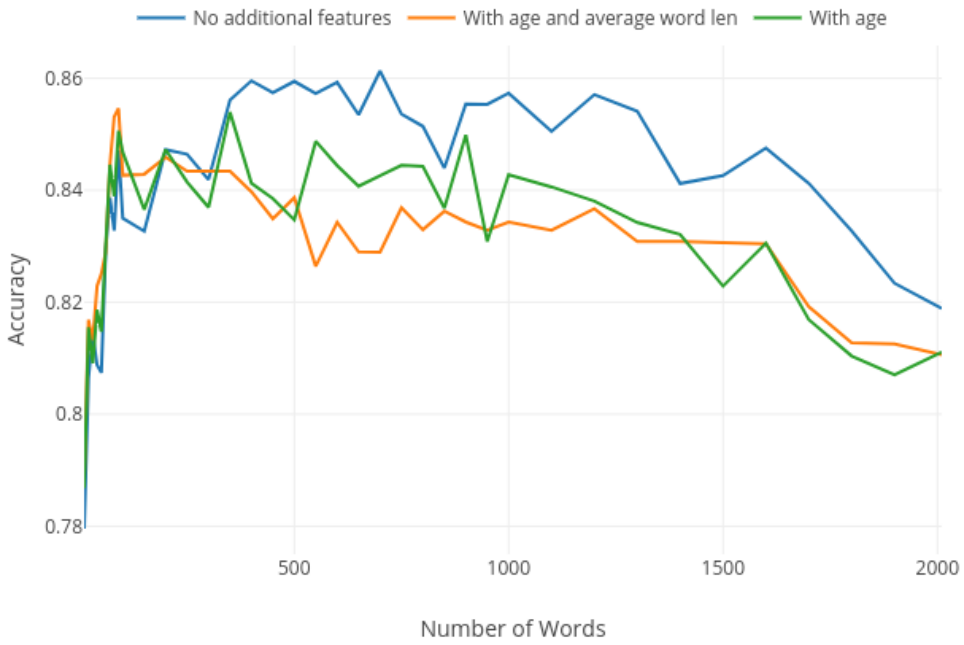

Figure 5.9: Accuracy for different combinations of features as we vary the number of words included (AD)

2. Adding the others models would make the picture more cluttered and harder to understand.

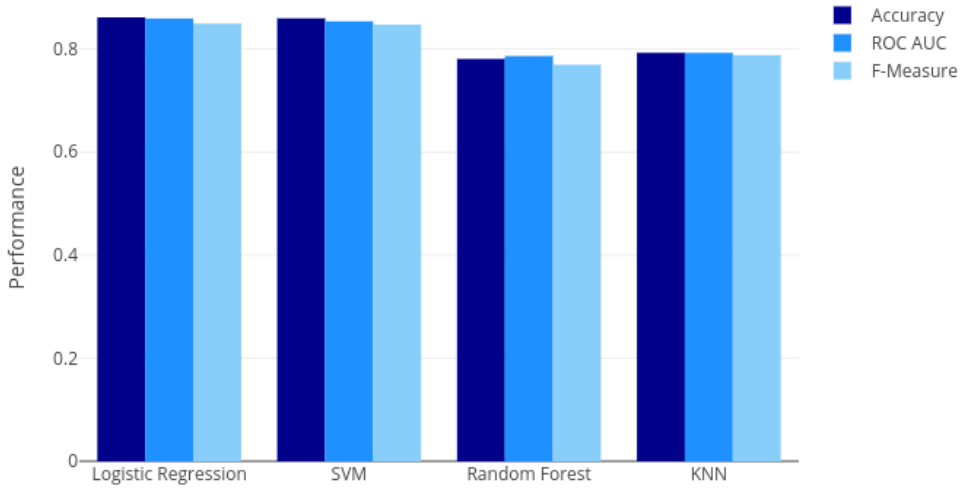

Figure 5.10: Accuracy, AUC and F-measure for each model at their optimum number of words (i.e. the peak performance in Figure 5.9)

Figure 5.9 allows us to draw some conclusions already. First, when the number of words used in Tf-IDF matrix are low, the model that uses both additional features outperforms the others, with the model that adds only the age feature being a close second, maybe due to the fact it has more information to discern between the two groups. When the number of words used in the Tf-IDF matrix increases, the model that does not use the additional features clearly outperforms the others. Second, the peak performance from the model without the additional features is higher than the others models, which

means that, once again, the addition of these features did not help our models.

Since the addition of features did not help any classifier, Figure 5.10 represents the performance of the four models, without using the additional features discussed previously, in terms of accuracy, AUC and F1. Once again, Logistic Regression classifier has the best results (Accuracy = 0.861, AUC = 0.858, F1 = 0.849), while, once more, SVM classifier is a close second. Although, almost every model had improvements in performance when compared to the previous approach of classifying between Dementia and Control, surprisingly, Random Forest had a decrease in performance.

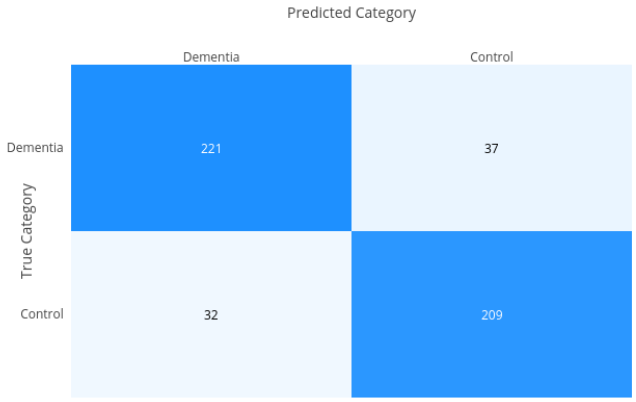

Figure 5.11: Confusion matrix for the Logistic Regression model

In addition, Figure 5.11 represents the confusion matrix for the best model, in this case, the Logistic Regression Classifier. While 13% of the control transcripts were classified as being a transcript produced by patients with dementia, 14% of the transcripts produced by dement subjects were classified as being a control transcript, which means that once again we have a higher rate of false negatives than false positives. Nevertheless, in this classification between Alzheimer's patients and Control subjects, both rates are lower.

In conclusion, we obtain promising results using Tf-IDF and χ^2 approach to classify between Alzheimer's and Control, results that are higher in terms of accuracy by 3% when compared to the state-of-art, while being simpler and easier to implement on a medical system.

5.2 Embeddings

Word embedding is the collective name for a set of language modeling and feature learning techniques in NLP where words from the vocabulary are mapped to vectors of real numbers, in such a way that closely related words/phrases have similar vectors. The general idea of vector space models is simple: when constructing two semantic spaces from two different corpora, it is expected the differences between the spaces to be related to the differences between the corpora. As an example, if we generate word vectors from a corpus of text about cars, and another set of word vectors about wildlife, we expect that the word *jaguar* will have two very different representations in these two spaces [68].

This methodology has become an increasing popular area of research in computational linguistics, with notable successes on tasks such as, query expansion for information retrieval [70], synonym iden-

tification [72], sentiment analysis [73], machine translation [74], and many others. Therefore, in the last years, the use of this model for detecting dementia through speech has been growing in the research environment.

Although, these embeddings are usually used as input for neural networks, in our case we will use these vectors in order to transform our transcripts into vectors, so we can train our models. There are several ways to construct a transcript vector representation from these vectors. In this case, we only used the mean and the sum of all the word vectors present in that transcript. We can represent this as follow:

$$mean_vector(t_i) = \frac{\sum_{j=0}^{k} vector(w_{i,j})}{k} \tag{5.4}$$

$$sum_vector(t_i) = \sum_{j=0}^{k} vector(w_{i,j}) \tag{5.5}$$

where t_i is the transcript i, $w_{i,j}$ is the word in the position j in the transcript i, $vector(w_{i,j})$ is the vector representation of the word $w_{i,j}$ and k is the total number of words in the transcript i.

This method instead of producing a sparse matrix like the TF-IDF method, produces a dense one (most entries are nonzero), which usually makes the classifier perform better. Applying this method to all the transcripts, we get a matrix $N \times M$ where N is the number of transcripts and M is the number of dimensions chosen for our vectors. Since our dataset is small, which makes it harder to train the models to create the mapping, we use a pre-trained model to obtain our word vectors. In this case, we use a model from GloVe 1.2 [75] that uses a combined corpus of Wikipedia 2014[1] plus Gigaword 5[2] for training. This corpus has a vocabulary of 400k words and a vector dimension of 50, 100, 200 and 300.

For the words present in the transcripts that are not in the GloVe model, we create a random vector for them with the same dimension using a uniform distribution. In a uniform distribution, samples are uniformly distributed over the interval $[low, high]$, in this case $[-1, 1]$, since the values for GloVe vectors are inside this interval. In other words, any value within the given interval is equally likely to be drawn, therefore the probability density function of the uniform distribution is:

$$p(x) = \frac{1}{high - low} \tag{5.6}$$

anywhere within the interval $[low, high]$, and zero elsewhere.

[1] https://dumps.wikimedia.org/enwiki/20140102/
[2] https://catalog.ldc.upenn.edu/LDC2011T07

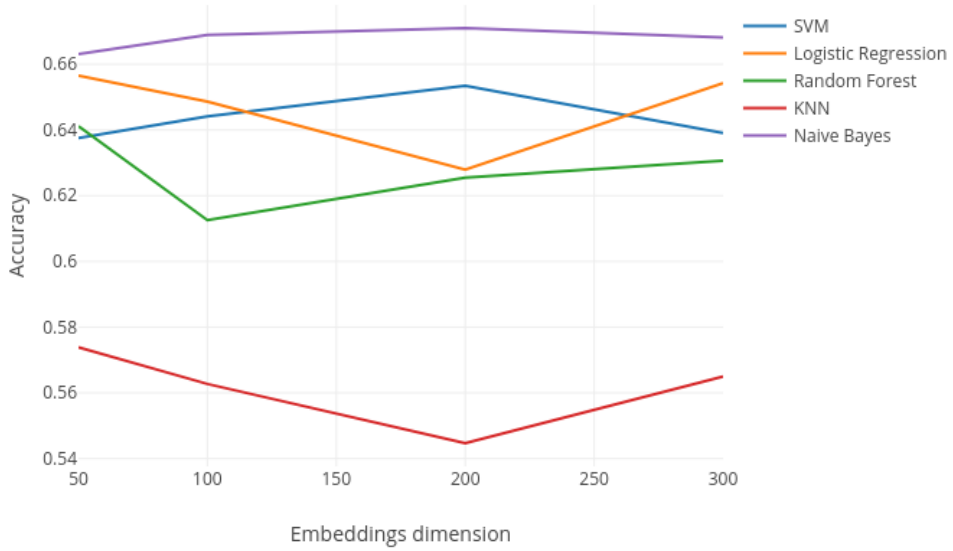

Figure 5.12: Accuracy for each model using the mean approach as we vary the number of dimensions

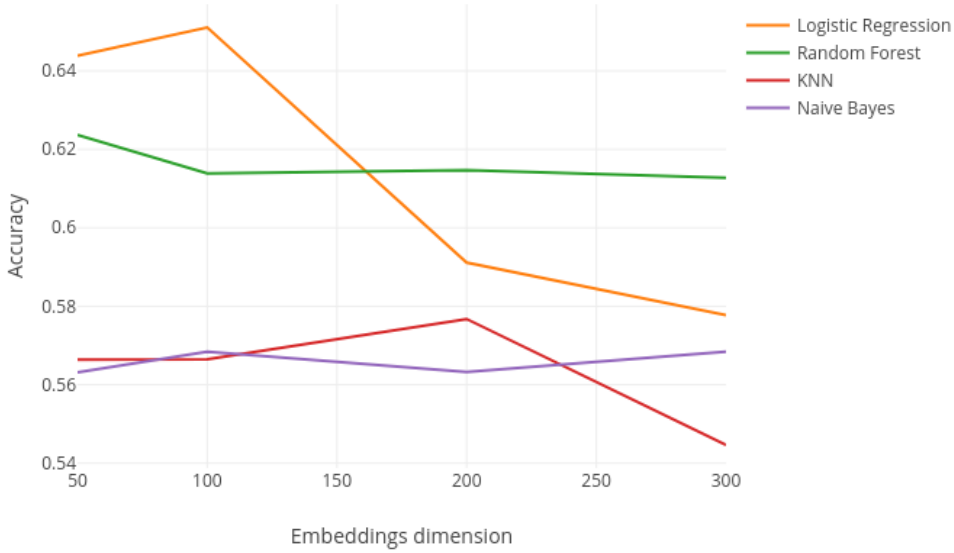

Figure 5.13: Accuracy for each model using the sum approach as we vary the number of dimensions

Figure 5.12 and 5.13 represents the Accuracy across a range of regressors using the GloVe embeddings model with different dimensions using the mean and the sum respectively. Support Vector Machines are not present in Figure 5.13, since, for some reason, when using the sum approach, it took around 10 hours to do one fold, and since, we are testing for 38 different k values, and for each k value we evaluate it using 10 folds, it would take around 158 days to have all values for the SVM classifier. Nevertheless, it is possible to draw some conclusions out of these two pictures. First, the Naive Bayes is the best classifier when using the mean approach, however, when using the sum approach it is one of the worst with a decrease in accuracy around 10%. In fact, every model has a better performance using the mean approach, although their decrease is not that significant when compared to the Naive

Bayes model. Second, it seems that this approach has a poor performance when compared to the Tf-IDF approach. For example, the best embedding model that uses the Naive Bayes classifier has a peak accuracy performance of 67.10%, while the simpler TF-IDF approach that does not use word selection nor additional features has a performance of 76.25%, which is 9.15% higher. Moreover, the models or increase in performance when the embeddings dimensions are increasing in low values and then decrease with higher embeddings dimension, or the opposite happens, the performance decreases with low values for the embeddings dimension, and increases with high values for the embeddings dimension. Basically, no model follows a linear function, where it either increases in performance with the increase of the dimension, or decreases with the increase of the dimension.

That being said, Figure 5.14 represents the best performance of all models using the mean approach, i.e., the peak performance in Figure 5.12. As previously concluded, Naive Bayes has the best performance across two performance metrics: Accuracy and ROC AUC. Logistic Regression, SVM and Random Forest have similar performance. However, Random Forests have the best F-measure performance (F-measure = 68.04%). KNN is clearly the worst model, with performances values lower than 60% in all metrics. Since Naive Bayes is the best classifier when considering Accuracy and ROC AUC performance, we will use this model to produce and evaluate the confusion matrix.

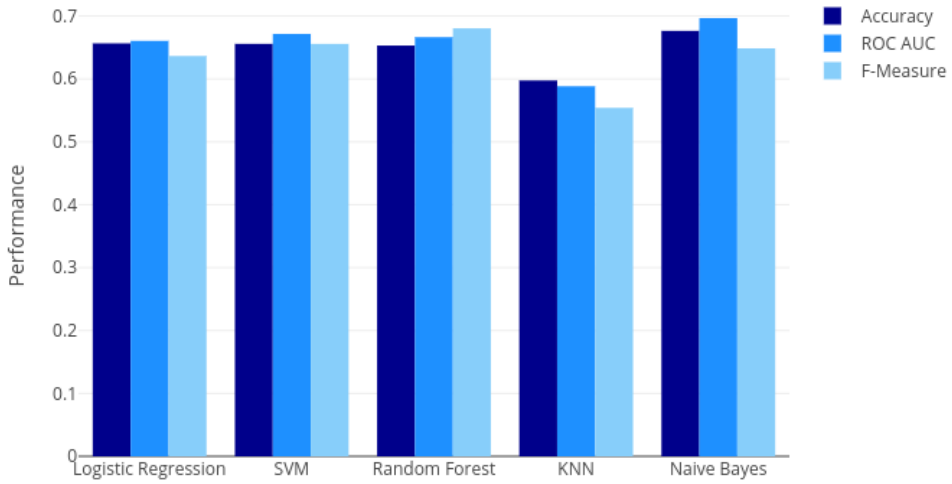

Figure 5.14: Accuracy, AUC and F-measure for each model at their optimum number of dimensions (i.e. the peak performance in Figure 5.12)

Figure 5.15 depicts the confusion matrix for the Naive Bayes classifier using the mean embeddings approach. When compared to Figure 5.8, which shows the confusion matrix for the Tf-IDF approach using additional features, we can conclude that the decrease in accuracy is due to a decrease in the number of dementia patients classified as having dementia. While in the TF-IDF approach we had 82.6% patients correctly classified, in this case, using the embeddings approach, we have only 58% correctly classified patients with dementia. There is also a decrease in the percentage of well classified control patients (from 82.2% to 73.9%), although less significant. This also means, we have a much higher false negative rate, which, as we already discussed, is harmful to our models.

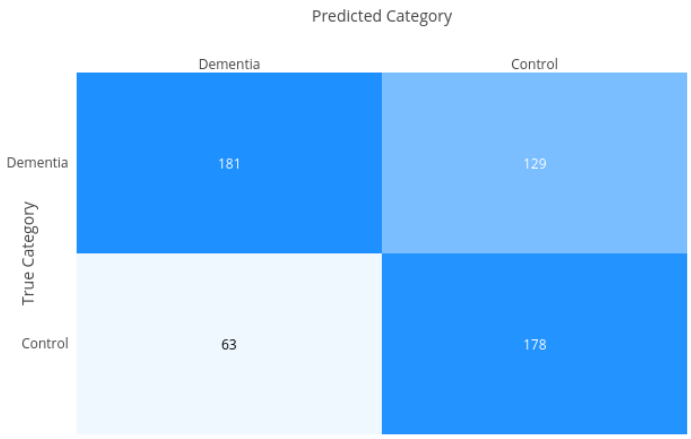

Figure 5.15: Confusion matrix for the Naive Bayes classifier

5.2.1 Alzheimer's vs Control

Once again, up until now, we only focused on Dementia vs Control in our embeddings approach for the classification of the patients. As already discussed, since language impairments and their degree are different between the different dementia diseases, one can argue that it makes more sense to classify between a disease that causes dementia and control. Therefore, once again, we will use solely control subjects and Alzheimer patients, since these are the two types most predominant in the dataset.

Therefore, we repeated the approach mentioned previously, however this time, only for patients with Alzheimer's and control subjects. Figures 5.16 and 5.17 show the results of this repetition. Figure 5.16 depicts the peak performance in terms of accuracy, ROC AUC and F-measure of the 5 chosen classifiers. Regarding the best model, in this case, SVM has the best performance, with an accuracy of $\approx$74,28%, which is almost 12% lower than the Logistic Regression model that uses the TF-IDF approach to classify between these two groups. In addition, the Logistic Regression model that uses embeddings is very close to the SVM one, with only a drop in accuracy of less than 3%. The other 3 classifiers have a significant lower performance not reaching an accuracy performance of 70%. Furthermore, once again, the KNN classifier has a meaningful worse performance when compared to all others.

That being said, we used the best performing model, SVM in this case, to analyse its confusion matrix, represented in the Figure 5.17. First, we can conclude that there is a vast number of patients with dementia being classified as control subjects, in fact, almost 30% of the dementia patients are being classified as being healthy. If we compare this percentage to the one obtained in the Section 5.1.3 where we used a bag-of-words approach to classify between these two groups, which was 14% of misclassification of dementia subjects, the percentage obtained in this approach was more than double. Moreover, this type of misclassification is, arguably, the worst case of missed categorization for this problem, as discussed previously. Second, there was also an increase of misclassification of control subjects. In the Section 5.1.3, the percentage of healthy subjects classified as having dementia was 13%, however in this approach we increased the false positive rate to 22%. Although this increase is smaller than the increase in the false negative rate, it is still significant and an indicator that this approach might not be appropriate for this problem.

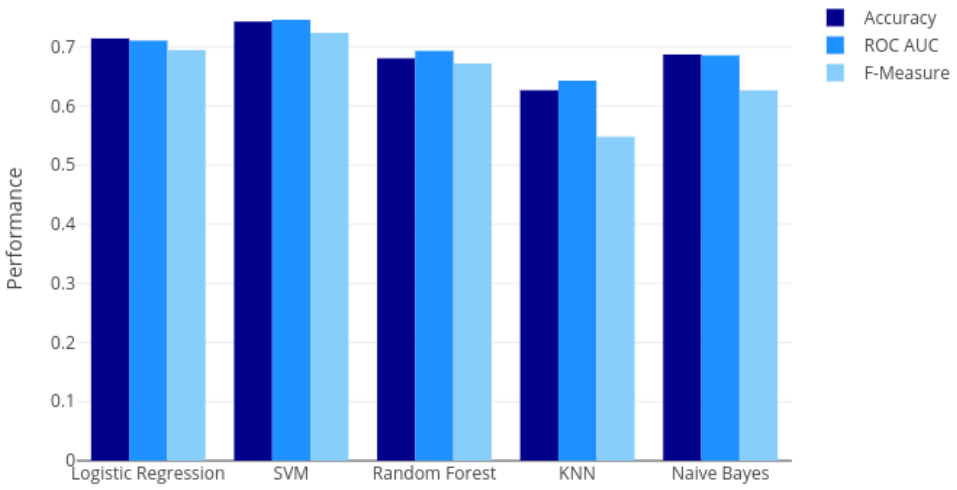

Figure 5.16: Accuracy, AUC and F-measure for each model at their optimum number of dimensions

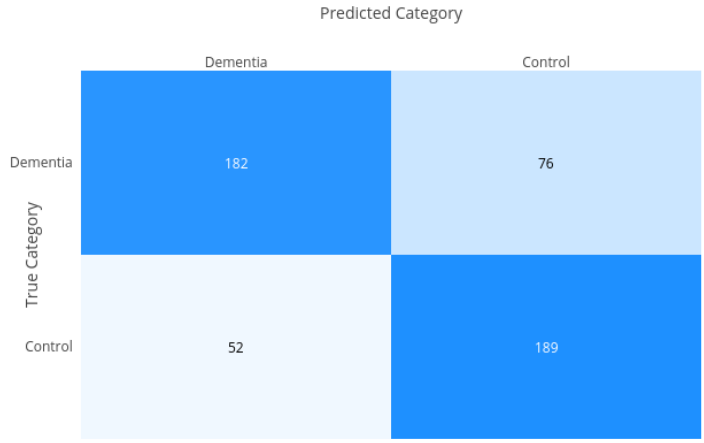

Figure 5.17: Confusion matrix for the SVM classifier

Overall, this approach has worse performance than the TF-IDF approach discussed previously, in all aspects. Although before implementing this approach, we thought that maybe it was a good solution to the problem, in hindsight, there are problems in this approach. Since we are using random vectors for the words not present in the GloVe model and this embedding model was trained on written corpus, while we are using a speech corpus, there will be a significant amount of words present in our corpus that will have a randomized vector representation. This means that, when doing the average for each dimension of the embeddings, there will be a significant random weight, which can be harmful to our models. Due to time constraints it was impossible to us to try other approaches as a solution for the unknown words, however, the next subsection will discuss some of them, among other things. Nevertheless, the main conclusion for this strategy is that it is not suitable for our problem, since when compared to the state-of-art or the previously mentioned BoW approach, it is clearly behind them in terms of performance.

Finally, since these results are worse than the ones of the previous approach, we thought that adding new features would only be time consuming.

5.3 Discussion

There are some conclusions that can be drawn after using these two different approaches. The main one is that the Tf-IDF in combination with the χ^2 test, when compared to the state-of-art, is, without a doubt, a very remarkable strategy to our problem. However, more concepts related to this approach could have been studied with more time. First, we only used unigrams for the Tf-IDF matrix, yet we could have studied the impact of bigrams or even trigrams. Although this would increase substantially the dimensionality of this matrix, by using the χ^2 test we could choose the most impactful ones, and even conclude if there are bigrams or trigrams mainly present on one of the groups. Second, we also could have studied the impact of more features like Age or the Average Word Length. For example, we could have added the 30 most impactful features of the work of Kathleen Fraser [13]. Nevertheless, it proved to be an excellent approach, surpassing the state-of-art.

Regarding the embeddings approach, it proved to be a poor strategy for the problem at hand, since it has worse performance on all fronts when compared to the BoW approach and most of the research previously done by other authors. As already discussed, this poor performance may be due to using a written text embeddings model in combination with assigning random vector representations to the words not present in the model. With more time, we could have studied alternatives to this problem. We could have removed the unknown words, however this means a loss of information, which could be harmful to our models. More importantly, we could have tried another model, a model that is trained on spoken text. As explained previously, there is a significant amount of tokens in spoken text that do not exist when writing (e.g. ”uhm”, ”ah”), therefore it is only normal that by using a written text embeddings model, plenty of words present in our transcripts will not appear on the embeddings model and therefore will have a random vector representation. By using a embeddings model that is trained on spoken text, we would reduce the amount of words not present in the model, and therefore have less words with a random representation. For example, we could have tried training the moded using the CALLHOME[3] dataset, which consists of 120 unscripted 30-minute telephone conversations between native speakers of English.

Nevertheless, by treating our problem as a classification one, we reached an approach with high performance (better than the state-of-art), while being simpler and easier to deploy.

5.4 Summary

In this chapter, we presented two different approaches to identify patients with dementia using speech. This includes a bag-of-words approach using TF-IDF and other using word embeddings.

The TF-IDF approach in combination with the χ^2 test for word selection surpasses the performance of the state of the art, with an accuracy of 86% when classifying between AD patients and control subjects. This performance drops to 82% in terms of accuracy when classifying between dementia patients and healthy subjects. In addition, we studied the impact of adding the age and the average word length as

[3] https://catalog.ldc.upenn.edu/LDC97S42

features for our models. However, this led to a decrease in performance.

However, the embeddings approach had a performance much lower than near the state of the art. When classifying between dementia patients and control subjects, it reached an accuracy of 67.10%, almost 15% lower than the state of the art. Furthermore, when classifying between AD patients and control subjects, this approach recorded a performance of 74.3%, which is almost 12% lower than the TF-IDF approach.

In the next chapter, we will present two different approaches that automatically predict MMSE scores based on the patient's transcripts.

Chapter 6

Diagnosis as Regression

In this chapter, I propose and evaluate different approaches to predict MMSE scores based on the patient's transcripts. I started by using a bag-of-words approach baseline, since this approach had a great performance in the classification task. Then, I tried to use Long Short-Term Memory (LSTM) Neural Networks for this task. In addition, I used these predicted MMSE scores to classify patients and compare its results with the results obtained in the previous chapter. Finally, I discuss both approaches and their results.

6.1 Bag-of-words Baseline

Since TF-IDF plus the χ^2 test as feature selection had a great performance in the classification task, we used the same method as baseline for predicting the MMSE scores. Therefore, this method is the same as described in the Section 5.1 except for the models used, regression models, in this case.

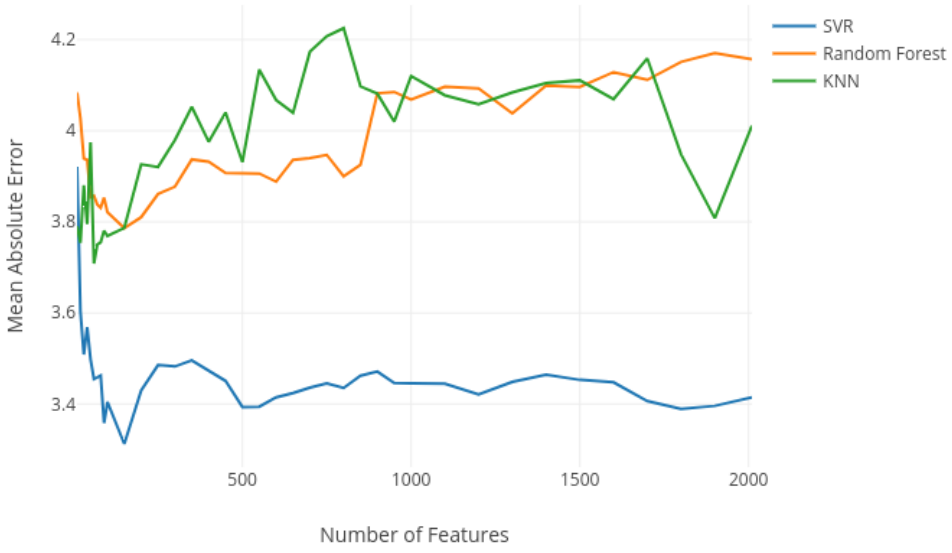

Figure 6.1: MAE for different models as we vary the number of words included

Figure 6.1 represents the MAE as I vary the number of words used based on the χ^2 test. Before analyzing the results, I just want to mention, as a disclaimer, that we also used Linear Regression as the regressor. However as the number of words increased, the error was also increasing, reaching MAE values of 20, which made difficult to interpret the rest of differences between the other three regressors. Therefore, this model was not included in the Figure 6.1. In addition, while in the classification task I was interested in increasing the accuracy, i.e I wanted the number of words that maximizes the accuracy function, in this case we are interested in the number of words that minimizes the error function.

Although the word selection decrease the error, the difference between the error while using all words and the minimum error is small. For example, for the Support Vector Regressor (SVR) the error is 3.41 while using all words and it is 3.31 for the optimal number of words, which means that the difference between using the χ^2 test or not in this case is a decrease of 0.1 in the MAE. In fact, with exception of the case of the Linear Regressor which I already mentioned, using the χ^2 test decreases our error on average by 0.21. Since the scale of the MMSE goes from 0 to 30, a decrease of that degree in our absolute error is not really significant.

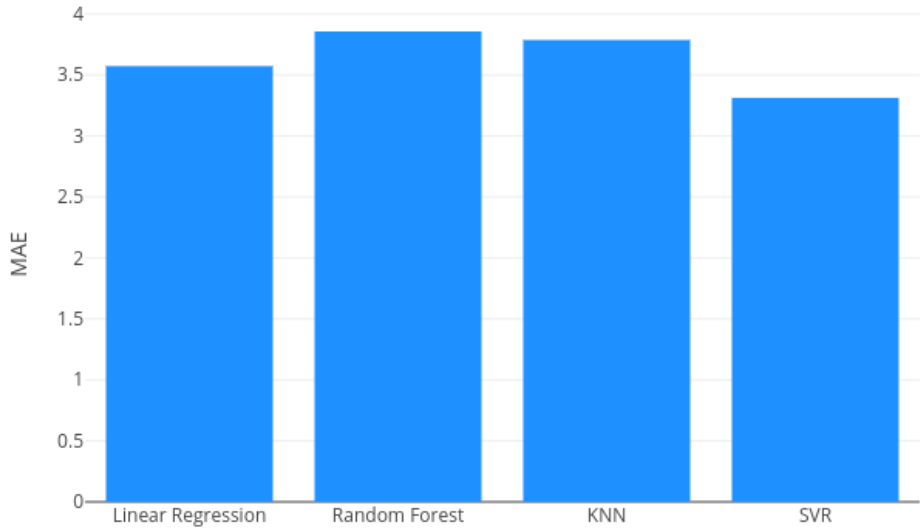

Figure 6.2: Minimum MAE for each model

Secondly, as we can see on Figure 6.2, the errors minimized for each model do not differ that much except for SVR model, since most of the errors are between 3.5 and 4, which once again on a scale from 0 to 30, is not that significant. Clearly, SVR outperforms the other regressors with a MAE of 3.31. When compared to the work of Maria Yancheva et al. [60] where they used all transcripts, we reached a lower error by 0.48. If compared to when the authors only used the individuals with more transcripts, we obtained a higher error by 0.40. Nevertheless, both of the differences are a bit of irrelevant, in our opinion, taking into account the scale of the MMSE test. However, once again, while the authors approach relies on 477 features that need to be precomputed, our approach is simpler and only depends on the words said, which requires little preprocessing before the model can return the prediction.

Anyway, our results are still far from the ones obtained by Lei Huang et al. [67], who obtained a

MAE of 1.68 using MRI images, which represents a decrease of the error of 1.63 when compared to our approach. This difference in error is relevant when accounted for the scale of the MMSE test, that goes from 0 to 30. Nevertheless, we still think this a good result, since when solely comparing to the research predicting the MMSE score using spontenaous speech, we have similar or better results while using a simpler approach.

6.2 LSTM Neural Networks

Long Short-Term Memory networks are a special kind of a Recurrent Neural Network (RNN). Unlike other traditional neural networks, RNN has an internal memory, which allows them to remember important things about the input they received, which enables them to be very precise in predicting what's coming next. In a RNN, the information cycles through a loop. This means that when it makes a decision, it takes into consideration the current input and also what it has learned from the inputs it received previously. Therefore a Recurrent Neural Network has two inputs, the present and the recent past. The relevance of this type of network is due to a lot of sequencial data contain crucial information about what is coming next. This is the reason why they are the preferred network in research for sequential data like time series, speech, text, financial data, audio, video, etc, and because they can form a much deeper understanding of a sequence and its context, compared to other types of neural networks.

LSTM networks are an extension for recurrent neural networks, with the purpose of extending their memory. The memory of a LSTM can be seen as a gated cell that decides whether or not to store or delete information (e.g if it opens the gates or not), based on the importance it assigns to the information. The assigning of importance happens through weights, which are also learned by the network, which means that it learns over time which information is important and which not. In an LSTM there are three gates: input, forget and output gate. These gates determine whether or not to let new input in (input gate), delete the information because it is not important (forget gate) or to let it impact the output at the current time step (output gate).

As mentioned previously in this document, such approaches need large amounts of training data in order to fully take advantage of them. Consequently, our hypothesis is that in the task at hand these deep structures will not provide any advantage compared to previously tested traditional machine learning schemes. In this section, we try to use an RNN architecture to validate (or disregard) this hypothesis.

Since I had no previous knowledge of Neural Networks and how to build them, I decided to use a more high-level and friendly-user API called Keras, instead of using a more low-level libraries like TensorFlow or PyTorch. Keras is an open source neural network library written in Python. It is capable of running on top of TensorFlow, Microsoft Cognitive Toolkit or Theano. Designed to enable fast experimentation with deep neural networks, it focuses on being user-friendly, modular, and extensible.

The architecture of the network built is represented on Figure 6.3. The input layer receives the transcripts as sequences of words and outputs these as an array of embeddings using the GloVe pre-trained model. The hidden layers are two Dense layers, which represents the regular densely-connected Neural Network layer, and two LSTM Layers intercalated. Finally, the output layer is a Dense layer that

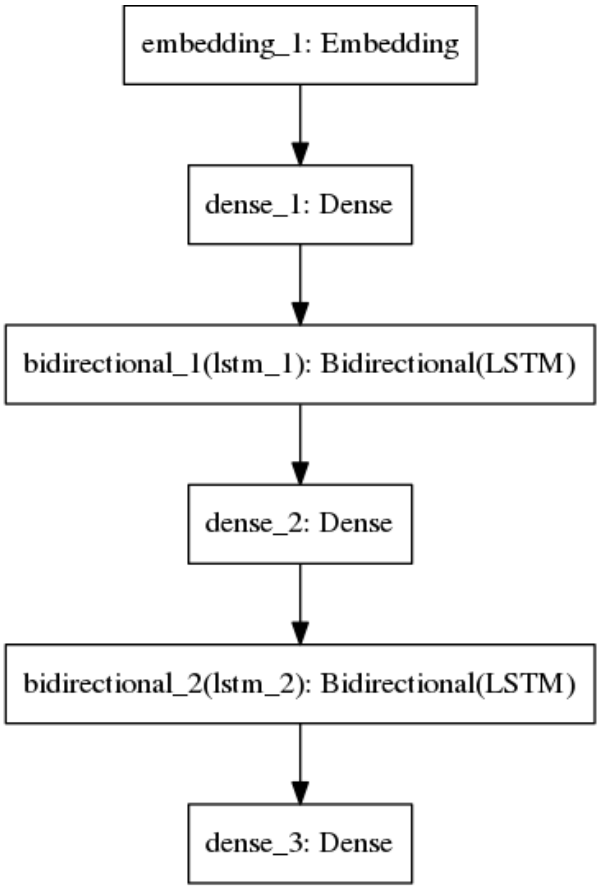

Figure 6.3: Neural Network Model Architecture

outputs the MMSE score.

This architecture is extremely similar to the example in the Keras Github[1] that classifies IMDB reviews using Bidirectional LSTM, with the hidden layers duplicated. We started by using the model used in the example, and expanded our Neural Network from there, thus the architecture shown is the architecture which gave us the best results. However, we could have tried to use other type of layers (such as Convolutional Neural Networks layers), but, due to time constraints, this was not possible.

In addition, we used the Root Mean Squared Error (RMSE) as the loss function. Therefore, given N as the number of transcripts, y_n as the true label of transcript x_n and $\hat{y}_n$ is the prediction for the given x_n transcript, the error is calculated as following:

$$RMSE = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{y}_n)^2} \tag{6.1}$$

We decided to use this loss function, instead of the MAE, due to some problems in the MAE function. Whenever we train a machine learning model, our goal is to find the point that minimizes loss function. Of course, both functions reach the minimum when the prediction is exactly equal to the true value. This point is given where the derivative of the function is zero. However, the MAE functions, like other absolute functions, e.g. $|x|$, has no derivative in the minimum point. Other problem with this function is that its

[1] https://github.com/keras-team/keras/blob/master/examples/imdb_bidirectional_lstm.py

gradient is the same throughout, which means the gradient will be large even for small loss values, which is not good for learning. On the other hand, RMSE behaves nicely in this case and will converge, since the gradient of RMSE loss is high for larger loss values and decreases as loss approaches 0. Nevertheless, to evaluate the models, we will still use the MAE, since it is the easiest to interpret.

We evaluated the neural network by using different dimensions for the embeddings vector and by using the corresponding GloVe model for the dimension chosen. The results of this evaluation are shown on Figure 6.4. The lowest possible dimension provides the less MAE, which is 4.34. When increasing the number of dimensions, the trend is to increase the error as well, with the dimension of 200 being an exception. At a dimension of 300, we reached the maximum of the error, which is 4.9. Nevertheless, even the lowest error provided by this approach, is higher than the SVR strategy presented previously by 1, which is a significant difference.

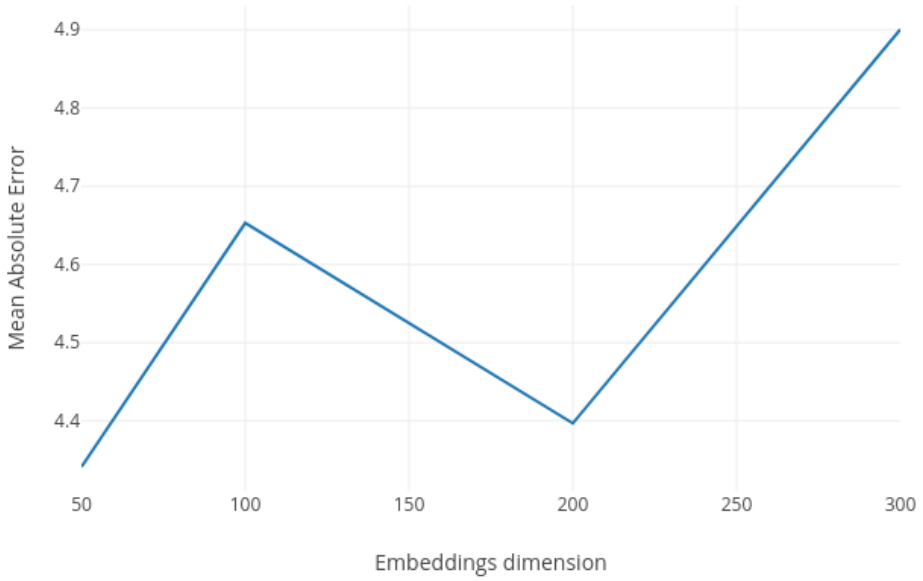

Figure 6.4: MAE using LSTM network as we vary the number of the embeddings dimensions

Nonetheless, before discarding Neural Networks as a solution to this problem, more research need to be done. Other architectures, since ours is very basic, or others types of layers can help decreasing the error. Due to time constraints, we could not experiment more with Neural Networks. However I believe there is space for growth for this approach, mainly if in the future a bigger dataset of transcripts becomes available.

6.3 Using MMSE Score for classification

As mentioned previously, the MMSE score can be used as a dementia diagnostic test. According to the test, if a subject has a score lower than 24 in this test, it may be indicative of having dementia. Therefore, using the regressors described above, I will use the following function for the classification:

$$c_i = \begin{cases} 0, & \text{if } \hat{y_i} \geq 24 \\ 1, & \text{if } \hat{y_i} < 24 \end{cases} \tag{6.2}$$

where c_i is the predicted label for the transcript i and $\hat{y_i}$ is the predicted MMSE score for i.

Patients with dementia of the MCI type are removed from this approach. Since the MMSE, as mentioned previously, cannot really capture the impairments in this type of dementia, these type of patients have always high scores, which would only increase the error of this approach.

Figure 6.5 depicts the performance of this approach using the models studied in the Chapters 6.1 and 6.2. Unfortunately, LSTM Neural Networks underperform when compared to the other 4 models, reaching an accuracy of 69.8%, which makes sense, since it was the model with the greatest MAE. Both the Linear Regressor and the Support Machine Regressor are the top performers, with similar performance in two metrics (Accuracy = 82.2%, AUC = 83.7%), however SVR has a better F1 measure score, 80,1% against 79,5%, therefore being the better model in general. The other two models, Random Forest Regressor and KNN Regressor, are in between, with KNN being a bit better than the Random Forest. When compared to the state-of-art classification between patients, we achieved the same results, while using a more broad spectre of dementia diseases, instead of just classifying between Alzheimer and Control, which is a harder task. In addition, both of these models are simpler when compared to the ones used on the state-of-art, where they used 200 different features, which reduces the probability of overfitting.

The approach described in Chapter 5.1 still outperforms this one, however only when classifying between patients with AD and healthy subjects. The performance is similar when classifying between subjects with dementia in general and control individuals.

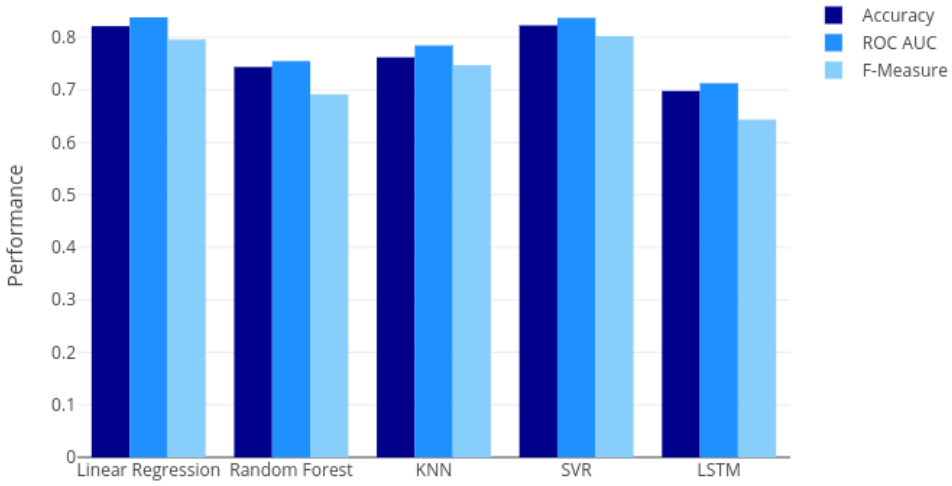

Figure 6.5: Accuracy, AUC and F-measure for each model

Moreover, Figure 6.6 depicts the confusion matrix for the SVR model, mentioned above. By analyzing the image, we can understand that most of the misclassification comes from the fact that a significant amount, 65, of transcripts produced by patients with dementia, are classified as being healthy. This

makes sense, when considering that between the disease types of 'ProbableAD' and 'PossibleAD' there are 53 patients with these type of dementia that have a MMSE equal or higher than 24. Therefore, even if our regressor had a MAE of 0, we would misclassify all these patients by using this approach. However, our regressor has an error superior than 0, so some of these patients can have a predicted MMSE below 24, thus being classified correctly. Actually, from these 53 patients, our regressor classified 35 transcripts as being produced by patients with dementia, and misclassified 18 transcripts. Thus, if we removed these transcripts, the accuracy of our model would increase from 82.2% to 84.3%, an increase of 2.1%. Regarding the classification of the transcripts produced by control subjects, the false positive rate (almost 10%) is low, similar to the other approaches discussed previously. Despite one can argue that we are manipulating our dataset to suit our needs, this approach surpasses the performance of the state-of-art model while classifiying in a more general manner (control vs different dementia diseases).

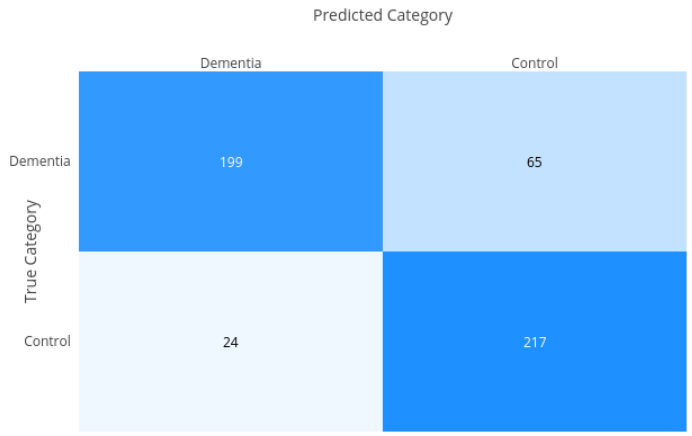

Figure 6.6: Confusion matrix for the SVR model

Concluding, in terms of regression, every one of our models underperforms when compared to the state-of-art regressors that uses image recognition for MRI exams. In comparison to the work by Maria Yancheva [60], which uses a dynamic Bayes Network to predict the MMSE scores from the transcripts of the DementiaBank dataset, our SVR approach achieved similar results in terms of MAE. In addition, in terms of classification, our approach of classifying the patients based on the predicted MMSE score outperforms the state of art models and our embeddings approach mentioned on the previous chapter. The only approach, up to our knowledge, that has better performance than this strategy, is the Bag-of-Words approach discussed, once again, in the previous approach, although it used a more restrict group of dementia diseases.

Nonetheless, this does not mean that this approach could not be improved, as we will discuss in the next subsection.

6.4 Discussion

There are some problems and limitations by using this regression approach. Primarily, the dataset used in combination with the MMSE score presents a big problem for this strategy. Firstly, patients with MCI

are not detected by this test, therefore they have scores above 24, which increases the inaccuracy of our classification. Secondly, some of the transcripts produced by patients with Alzheimer, either probable or possible, have a MMSE score higher than 24, that, once again, adds inaccuracy to our model. Therefore, the results for the performance of the models explained previously should be taken carefully, since the dataset and the MMSE test makes them not reliable.

Nevertheless, the approach that uses the SVR in combination with Tf-IDF and the χ^2 test proved to be, once again, a good approach for the problem, with a MAE similar to the work of Maria Yancheva [60]. Even in terms of classification, it performed better than the state of art models but worse than using the same approach but instead classifying based on the MMSE score, the classification comes directly from the Tf-IDF matrix filtered by the χ^2 test. This may be due to problem mentioned above, where the MMSE score leads to inaccuracy for our models. Nonetheless, more experiments should be done, such as, varying the number of n-grams used, or even add new features and study their impact on the performance of the model.

Moreover, the deep learning approach using LSTM Neural Networks to predict the MMSE scores did not provide any improvements to the previous one. Overall, it had worse performance in either regression or classification, while being a more complex approach with more probability to overfit. However, few architectures were studied, and the one chosen is a very simple one based on a Keras example. Therefore, there is still hope for this kind of approach, since before discarding it completely, more experiments with others architectures and different layers should be done. Additionally, the architecture chosen has an embedding layer that uses the GloVe model, which, as we already discussed on the previous Mean Embeddings approach presents some problems for this task, and therefore, experimenting with other embeddings model is advised.

Nevertheless, by treating our problem as a regression one, we reached an approach with performance worse than the state-of-art. However, by using this regression approach for the classification, the model had a better performance than the state-of-art, yet it performed worse than the previous BoW approach for classification discussed in the previous chapter. Furthermore, since it adds one more operation (the conversion of the MMSE score to the diagnosis), it turns the model more complex and dependent of a test with deficiencies. For example, this model, unlike the ones in the previous chapter, can not detect people with MCI.

6.5 Summary

In this chapter, we presented two different methods to predict the MMSE score of patients with dementia or not using speech. These include a bag-of-words approach using TF-IDF and other using an LSTM neural network.

The TF-IDF approach in combination with the χ^2 test for word selection surpasses the performance of the work of Maria Yancheva et al. [60], with an MAE of 3.31. The LSTM neural network had a higher error, maybe due to its simple architecture.

In addition, we used these approaches for classification between dementia patients and control sub-

jects based on the predicted MMSE score. By using the TF-IDF approach for the prediction of the MMSE scores, we reached an accuracy of 82.2%, which is higher than the state of the art, however it is lower than using the TF-IDF directly for classification, as reported in Chapter 5.1.

Chapter 7

Conclusions and Future Work

Detection of dementia, and more importantly early detection of dementia, is important not only for patients for whom a diagnosis is the first step for appropriate support and treatment, but for researchers, who claim that early diagnosis is an important step for finding a cure [8]. Recently, there have been authors with successful results using machine learning and natural language processing techniques to automatically detect dementia from speech. This thesis has made one main contribution towards this effort by using a bag-of-words approach.

First, we analyzed the dataset used in previous research, DementiaBank, in order to find useful features and/or limitations. Foremost, this dataset includes different kinds of diseases related to dementia, such as, Alzheimer's, Parkison's, Vascular Dementia and MCI. However, the most predominant disease in the dataset is Alzheimer's, with the other types of diseases not having a significant expression in the dataset, which means that multiclass labeling is a type of problem almost impossible with this information. Regarding the features, we found two features that have different distributions in the group of healthy subjects and in the group of patients with dementia, which are the age of the patient and the average word length of the transcript. Nevertheless, this dataset also presents some difficulties and limitations. In the first place, the patients with MCI and some with Alzheimer's have a MMSE score higher than the threshold to be considered as a dement subject, which make difficult the task of classifying patients in this dataset based on a predicted MMSE score grounded on the subjects' transcripts. In the second place, the DementiaBank dataset is small, which increases the probability of overfitting and outliers and noise have more impact on the data. Furthermore, the size of dataset also removes the possibility of some experiments, such as, classification based on deep learning models. Finally, when gathering the data, the characteristics of the patients, such as, age, education, etc., were not taken into account, which means that the differences found in the discourse of these two groups can be due to the demographic differences and not due to having dementia or not.

Regarding the problem at hand, I used two different strategies: classification and regression. By treating the problem as a classification one, I experimented a BoW approach using TF-IDF, since it is easy to implement and BoW approaches already proved to be a good solution for the problem. Surprisingly, this method had a performance only 6% lower in terms of accuracy, while classifying in a more

general setting. Therefore, I tried, successfully, to expand this approach by adding new features and by selecting the words for the tf-idf matrix using the χ^2 test. This technique increased all evaluation metrics and reached similar performance to the state-of-art, while, once again, classifying in a more general setting. Actually, when doing the same classification as the state-of-art, Alzheimer's versus Control, I surpassed the performance of the state-of-art by 4% in terms of accuracy and almost 3% in terms of F-measure. Other strategy using the mean or sum of the embeddings present on the transcripts was experimented without any success.

By treating the problem as regression, I developed two methods to predict the MMSE score of the patient. Firstly, I used the BoW words approach, which had a great performance in the classification task, as a baseline, by replacing the classifier in the end by a regression model. This approach reached similar performance to the work of Maria Yancheva et al. [60], which uses the same dataset, however it had a worse performance when compared to the state-of-art of prediction of the MMSE, which uses neuroimaging. A second strategy based on LSTM Neural Networks was tried without much success. In our opinion, this could be due to little experimenting in the network architecture and its parameters, as a result of time constraints and the size of the data. In addition, I used the predicted MMSE to classify patients, as an experiment of what was possible to do with these scores, even though, as already said, classifying patients in this dataset based on MMSE score can lead to errors. Nonetheless, surprisingly, this approach had a similar performance to the state-of-art or better if I filtered out the transcripts produced by patients with dementia with an actual MMSE score equal or above 24.

Finally, using several technologies, I deployed our BoW classification model in a Heroku server, and built an Android Client that translates the speech of the user into text and presents the probability of having dementia based on the discourse produced. However, this Android application is not ready for production. It serves more as a proof-of-concept of what is possible to do if the model, such as ours, automatically parses the discourse and classify it based on simple features that can be also calculated without any human help. Additional details of this proof of concept can be found in Annex A.

7.1 Future Work

There are multiple directions I would like to take this work in the future. Foremost, further experiments could be done with the Bag-of-Words approach. During the thesis, I only used unigrams for the tf-idf matrix. Although using a higher n for the n-grams would increase the dimensionality of our matrix, which can be lead to overfitting, it could improve our results and we could remove the n-grams with low χ^2 values, which would reduce the dimensionality. Furthermore, we only studied the impact of two different features (age and mean word length), however more features related to other characteristics could have been added to our model. For example, in the state-of-art [13] they used 277 features, which then were filtered to 30 different features, that could be added to our model in order to study their impact in the performance of our model. Finally, in this approach, we could have tried ensemble learning using the different classifiers model used throughout the thesis, giving more weight to the decision for the ones with better performance.

Regarding the embeddings model, there is one major thing that can be done in the future. The GloVe model chosen was trained in a written text corpus which misses some characteristics of the spoken test, such as, disfluencies. Therefore, there are a significant amount of words in spoken test that do not appear in an embeddings model trained on a written text corpus. In this thesis, we used a random vector representation for each one of these words, however other approaches could be done. We could use a vector representation with zeros in every dimension, or ignore completely these words and remove them from the transcripts. However, these two approaches have also flaws, at least theoretically. Using a zero vector representation, would only decrease the weight of the other words in the mean strategy, and would do nothing in the sum strategy. Ignoring these words, would mean a loss of information that could be important to the model. In our opinion, for future work, the best would be to use another model that was trained on a spoken test corpus, such as CALLHOME[1], since this will have a vector representations for words that only appear in spoken test, and therefore we would reduce the ammount of words in the transcripts that are not present on the embeddings model.

In addition, concerning the LSTM Neural Network approach for the prediction of the MMSE scores, a lot more could be done. More complex architectures should be tried with different loss and activation functions. In addition, other types of layers could also be useful for this type of task. Like the embeddings approach, we should also try a different embeddings model, one that was tried in a spoken test corpus. Still, in this type of approaches we will have always have the limitations of a reduced dataset.

Finally, there are two possible important improvements to the deployment of the model and the Android client. First of all, adding a database to our model would be a huge improvement. This adds the possibility of storing the requests made by the Android client, which then, we could analyze the transcript and see if it was worth to add to our train dataset. This means that, with more users, we could increase our dataset, which could add the possibility of trying different and more complex classifiers. Second, the speech-to-text software used in the Android client is the Google default for Android, which comes with limitations for long discourses. Replacing it with a state of art speech-to-text software, would make our client more reliable and easier to use.

[1] https://catalog.ldc.upenn.edu/LDC97S42

Bibliography

[1] Robert L. Spitzer, Kurt Kroenke Md, and Janet B. W. Williams. Diagnostic and statistical manual of mental disorders, third edition. *AMERICAN PSYCHIATRIC ASSOCIATION*, 1980.

[2] A.E. Budson and P.R. Solomon. *Memory Loss: A Practical Guide for Clinicians*. ClinicalKey 2012. Elsevier Saunders, 2011. ISBN 9781416035978. URL https://books.google.pt/books?id=0eZUYgEACAAJ.

[3] World Health Organization. Dementia fact sheet 362, 2017. URL http://www.who.int/mediacentre/factsheets/fs362/en/.

[4] Martin Prince, Renata Bryce, Emiliano Albanese, Anders Wimo, Wagner Ribeiro, and Cleusa P. Ferri. The global prevalence of dementia: A systematic review and metaanalysis. *Alzheimer's & Dementia*, 9(1):63 – 75, 2013. ISSN 1552-5260. doi: https://doi.org/10.1016/j.jalz.2012.11.007. URL http://www.sciencedirect.com/science/article/pii/S1552526012025319.

[5] United Nations Department of Economic and Social Affairs Population Division. World population ageing report, 2015. URL www.un.org/en/development/desa/population/publications/pdf/ageing/WPA2015_Report.pdf.

[6] OECD. *Health at a Glance 2017*. 2017. doi: https://doi.org/https://doi.org/10.1787/health_glance-2017-en. URL https://www.oecd-ilibrary.org/content/publication/health_glance-2017-en.

[7] Martin Prince, Anders Wimo, Maëlenn Guerchet, Gemma-Claire Ali, Yu-Tzu Wu, and Matthew Prina. World alzheimer report 2015. the global impact of dementia. an analysis of prevalence, incidence, cost and trends. 08 2015.

[8] 2016 alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 12(4):459 – 509, 2016. ISSN 1552-5260. doi: https://doi.org/10.1016/j.jalz.2016.03.001. URL http://www.sciencedirect.com/science/article/pii/S1552526016000856.

[9] Alzheimer's Disease International. Dementia statistics, 2015. URL https://www.alz.co.uk/research/statistics.

[10] Jamie Reilly, Joshua Troche, and Murray Grossman. *Language Processing in Dementia*, pages

336–368. Wiley-Blackwell, 2011. ISBN 9781444344110. doi: 10.1002/9781444344110.ch12. URL http://dx.doi.org/10.1002/9781444344110.ch12.

[11] S. H. Ferris and M. Farlow. Language impairment in Alzheimer's disease and benefits of acetylcholinesterase inhibitors. *Clin Interv Aging*, 8:1007–1014, 2013.

[12] E. Schwam and Y. Xu. Cognition and function in Alzheimer's disease: identifying the transitions from moderate to severe disease. *Dement Geriatr Cogn Disord*, 29(4):309–316, 2010.

[13] Kathleen C Fraser, Jed A Meltzer, and Frank Rudzicz. Linguistic features identify alzheimer's disease in narrative speech. *Journal of Alzheimer's disease : JAD*, 49(2):407—422, 2016. ISSN 1387-2877. doi: 10.3233/jad-150520. URL https://doi.org/10.3233/JAD-150520.

[14] V. Rentoumi, L. Raoufian, S. Ahmed, C. A. de Jager, and P. Garrard. Features and machine learning classification of connected speech samples from patients with autopsy proven Alzheimer's disease with and without additional vascular pathology. *J. Alzheimers Dis.*, 42 Suppl 3:3–17, 2014.

[15] American Psychiatric Association. *Diagnostic and statistical manual of mental disorders: DSM-5*. Autor, Washington, DC, 5th ed. edition, 2013.

[16] Sebastian Wankerl, Elmar Nöth, and Stefan Evert. An n-gram based approach to the automatic diagnosis of alzheimer's disease from spoken language. pages 3162–3166, 08 2017.

[17] Blanka Klimova and Kamil Kuca. Speech and language impairments in dementia. *Journal of Applied Biomedicine*, 14(2):97 – 103, 2016. ISSN 1214-021X. doi: https://doi.org/10.1016/j.jab.2016.02.002. URL http://www.sciencedirect.com/science/article/pii/S1214021X15300387.

[18] ADEAR: "Alzheimer's Disease Education and a service of the National Institute on Aging." Referral Center. Diagram of the brain of a person with alzheimer's disease, 2017. URL https://upload.wikimedia.org/wikipedia/commons/1/14/Alzheimer%27s_disease_brain_severe.jpg.

[19] Maria Yancheva and Frank Rudzicz. Vector-space topic models for detecting alzheimer's disease. In *ACL*, 2016.

[20] Marjorie Nicholas, Loraine Obler, Martin Albert, and Nancy Helm-Estabrooks. Empty speech in alzheimer's disease and fluent aphasia. 28:405–10, 10 1985.

[21] Katinka Dijkstra, Michelle S Bourgeois, Rebecca S Allen, and Louis D Burgio. Conversational coherence: discourse analysis of older adults with and without dementia. *Journal of Neurolinguistics*, 17(4):263 – 283, 2004. ISSN 0911-6044. doi: https://doi.org/10.1016/S0911-6044(03)00048-4. URL http://www.sciencedirect.com/science/article/pii/S0911604403000484.

[22] Alexander Rapp and Barbara Wild. Nonliteral language in alzheimer dementia: A review. 17: 207–18, 03 2011.

[23] Bernard Croisile, Bernadette Ska, Marie-Josée Brabant, Annick Duchene, Yves Lepage, Gilbert Aimard, and Marc Trillet. Comparative study of oral and written picture description in patients with alzheimer's disease. 53:1–19, 05 1996.

[24] Seyed Sajjadi, Karalyn Patterson, Michal Tomek, and Peter Nestor. Abnormalities of connected speech in semantic dementia vs alzheimer's disease. 26:1–20, 06 2012.

[25] D Kempler, S Curtiss, and C Jackson. Syntactic preservation in Alzheimer's disease. *Journal of speech and hearing research*, 30(3):343–50, 1987. ISSN 0022-4685. doi: 10.1044/jshr.3003.343. URL http://www.ncbi.nlm.nih.gov/pubmed/3669641.

[26] Guila Glosser and Toni Deser. Patterns of discourse production among neurological patients with fluent language disorders. *Brain and Language*, 40(1):67 – 88, 1991. ISSN 0093-934X. doi: https://doi.org/10.1016/0093-934X(91)90117-J. URL http://www.sciencedirect.com/science/article/pii/0093934X9190117J.

[27] John R. Hodges and Karalyn Patterson. Is semantic memory consistently impaired early in the course of alzheimer's disease? neuroanatomical and diagnostic implications. *Neuropsychologia*, 33(4):441 – 459, 1995. ISSN 0028-3932. doi: https://doi.org/10.1016/0028-3932(94)00127-B. URL http://www.sciencedirect.com/science/article/pii/002839329400127B.

[28] Petersen RC, Smith GE, Waring SC, Ivnik RJ, Tangalos EG, and Kokmen E. Mild cognitive impairment: Clinical characterization and outcome. *Archives of Neurology*, 56(3):303–308, 1999. doi: 10.1001/archneur.56.3.303. URL +http://dx.doi.org/10.1001/archneur.56.3.303.

[29] A. Cunje, D. W. Molloy, T. I. Standish, and D. L. Lewis. Alternate forms of logical memory and verbal fluency tasks for repeated testing in early cognitive changes. *Int Psychogeriatr*, 19(1):65–75, Feb 2007.

[30] K. J. Murphy, J. B. Rich, and A. K. Troyer. Verbal fluency patterns in amnestic mild cognitive impairment are characteristic of Alzheimer's type dementia. *J Int Neuropsychol Soc*, 12(4):570–574, Jul 2006.

[31] D. A. Bennett, R. S. Wilson, J. A. Schneider, D. A. Evans, L. A. Beckett, N. T. Aggarwal, L. L. Barnes, J. H. Fox, and J. Bach. Natural history of mild cognitive impairment in older persons. *Neurology*, 59(2):198–205, Jul 2002.

[32] Vanessa Taler and Natalie A. Phillips. Language performance in alzheimer's disease and mild cognitive impairment: A comparative review. *Journal of Clinical and Experimental Neuropsychology*, 30(5):501–556, 2008. doi: 10.1080/13803390701550128. URL https://doi.org/10.1080/13803390701550128. PMID: 18569251.

[33] F. Ribeiro, A. de Mendonca, and M. Guerreiro. Mild cognitive impairment: deficits in cognitive domains other than memory. *Dement Geriatr Cogn Disord*, 21(5-6):284–290, 2006.

[34] Bill Gates. Why diagnosing alzheimer's today is so difficult—and how we can do better, 2018. URL https://www.gatesnotes.com/Health/A-better-way-of-diagnosing-Alzheimers.

[35] J T O'BRIEN. Role of imaging techniques in the diagnosis of dementia. *The British Journal of Radiology*, 80(special_issue_2):S71–S77, 2007. doi: 10.1259/bjr/33117326. URL https://doi.org/10.1259/bjr/33117326. PMID: 18445747.

[36] Tom N. Tombaugh and Nancy J. McIntyre. The Mini-Mental State Examination: A Comprehensive Review. *Journal of the American Geriatrics Society*, 40(9):922–935, 1992. ISSN 15325415. doi: 10.1111/j.1532-5415.1992.tb01992.x.

[37] Lindy Harrell, Daniel Marson, Anjan Chatterjee, and Jo Ann Parrish. The severe mini-mental state examination: A new neuropsychologic instrument for the bedside assessment of severely impaired patients with alzheimer disease. 14:168–75, 07 2000.

[38] Eric M. Reiman, Yakeel T. Quiroz, Adam S. Fleisher, Kewei Chen, Carlos Velez-Pardo, Marlene Jimenez-Del-Rio, Anne M. Fagan, Aarti R. Shah, Sergio Alvarez, Andrés Arbelaez, Margarita Giraldo, Natalia Acosta-Baena, Reisa A. Sperling, Brad Dickerson, Chantal E. Stern, Victoria Tirado, Claudia Munoz, Rebecca A. Reiman, Matthew J. Huentelman, Gene E. Alexander, Jessica B.S. Langbaum, Kenneth S. Kosik, Pierre N. Tariot, and Francisco Lopera. Brain imaging and fluid biomarker analysis in young adults at genetic risk for autosomal dominant Alzheimer's disease in the presenilin 1 E280A kindred: A case-control study. *The Lancet Neurology*, 2012. ISSN 14744422. doi: 10.1016/S1474-4422(12)70228-4.

[39] Svenja Adolphs, Brian Brown, Ronald Carter, Paul Crawford, Opinder Sahota, Mariya Limerick, Alison Pilnick, Linda Gibson, and Stacy Johnson. Applying corpus linguistics in a health care context. 1, 01 2004.

[40] Charlene Pope and Boyd Davis. Finding a balance: The carolinas conversation collection. 7: 143–161, 05 2011.

[41] James Becker, Francois Boller, O.L. Lopez, J Saxton, and K.L. McGonigle. The natural history of alzheimer's disease: Description of study cohort and accuracy of diagnosis. 51:585–94, 07 1994.

[42] Thalia S. Field, Vaden Masrani, Gabriel Murray, and Giuseppe Carenini. Improving diagnostic accuracy of alzheimer's disease from speech analysis using markers of hemispatial neglect. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, 13(7):P157–P158, Jul 2017. ISSN 1552-5260. doi: 10.1016/j.jalz.2017.06.2598. URL http://dx.doi.org/10.1016/j.jalz.2017.06.2598.

[43] M. F. Mendez, M. M. Cherrier, and J. S. Cymerman. Hemispatial neglect on visual search tasks in Alzheimer's disease. *Neuropsychiatry Neuropsychol Behav Neurol*, 10(3):203–208, Jul 1997.

[44] S. Ishiai, R. Okiyama, Y. Koyama, and K. Seki. Unilateral spatial neglect in Alzheimer's disease. A line bisection study. *Acta Neurol. Scand.*, 93(2-3):219–224, 1996.

[45] Frank Rudzicz, Leila Chan Currie, Andrew Danks, Tejas Mehta, and Shunan Zhao. Automatically identifying trouble-indicating speech behaviors in alzheimer's disease. In *Proceedings of the 16th International ACM SIGACCESS Conference on Computers & Accessibility*, ASSETS '14, pages 241–242, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2720-6. doi: 10.1145/2661334.2661382. URL http://doi.acm.org/10.1145/2661334.2661382.

[46] Kairit Sirts, Olivier Piguet, and Mark Johnson. Idea density for predicting alzheimer's disease from transcribed speech. *CoRR*, abs/1706.04473, 2017. URL http://arxiv.org/abs/1706.04473.

[47] David A Snowdon, Susan J Kemper, James A Mortimer, Lydia H Greiner, David R Wekstein, and William R Markesbery. Linguistic Ability in Early Life and Cognitive Function and Alzheimer's Disease in Late Life. *JAMA*, 275(7):528–532, 1996. ISSN 0098-7484. doi: 10.1001/jama.1996.03530310034029.

[48] Leandro Borges dos Santos, Edilson Anselmo Corrêa Júnior, Osvaldo N. Oliveira Jr., Diego R. Amancio, Leticia Lessa Mansur, and Sandra M. Aluísio. Enriching complex networks with word embeddings for detecting mild cognitive impairment from speech transcripts. *CoRR*, abs/1704.08088, 2017. URL http://arxiv.org/abs/1704.08088.

[49] Daniel B. Hier, Karen Hagenlocker, and Andrea Gellin Shindler. Language disintegration in dementia: Effects of etiology and severity. *Brain and Language*, 25(1):117 – 133, 1985. ISSN 0093-934X. doi: https://doi.org/10.1016/0093-934X(85)90124-5. URL http://www.sciencedirect.com/science/article/pii/0093934X85901245.

[50] Bernard Croisile, Bernadette Ska, Marie-Josee Brabant, Annick Duchene, Yves Lepage, Gilbert Aimard, and Marc Trillet. Comparative study of oral and written picture description in patients with alzheimer's disease. *Brain and Language*, 53(1):1 – 19, 1996. ISSN 0093-934X. doi: https://doi.org/10.1006/brln.1996.0033. URL http://www.sciencedirect.com/science/article/pii/S0093934X96900334.

[51] K. E. Forbes-McKay and A. Venneri. Detecting subtle spontaneous language decline in early alzheimer's disease with a picture description task. *Neurological Sciences*, 26(4):243–254, Oct 2005. ISSN 1590-3478. doi: 10.1007/s10072-005-0467-9. URL https://doi.org/10.1007/s10072-005-0467-9.

[52] Yi hsiu Lai, Hsiu hua Pai, and Yu te Lin. To be semantically-impaired or to be syntactically-impaired: Linguistic patterns in chinese-speaking persons with or without dementia. *Journal of Neurolinguistics*, 22(5):465 – 475, 2009. ISSN 0911-6044. doi: https://doi.org/10.1016/j.jneuroling.2009.03.004. URL http://www.sciencedirect.com/science/article/pii/S0911604409000268.

[53] B. Roark, M. Mitchell, J. P. Hosom, K. Hollingshead, and J. Kaye. Spoken Language Derived Measures for Detecting Mild Cognitive Impairment. *IEEE Trans Audio Speech Lang Process*, 19 (7):2081–2090, Sep 2011.

[54] Vaden Masrani, Gabriel Murray, Thalia Shoshana Field, and Giuseppe Carenini. *Domain Adaptation for Detecting Mild Cognitive Impairment*, pages 248–259. Springer International Publishing, Cham, 2017. ISBN 978-3-319-57351-9. doi: 10.1007/978-3-319-57351-9_29. URL https://doi.org/10.1007/978-3-319-57351-9_29.

[55] Hal Daumé III. Frustratingly easy domain adaptation. *CoRR*, abs/0907.1815, 2009. URL http://arxiv.org/abs/0907.1815.

[56] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. *CoRR*, abs/1511.05547, 2015. URL http://arxiv.org/abs/1511.05547.

[57] László Tóth, Gábor Gosztolya, Veronika Vincze, Ildikó Hoffmann, Gréta Szatlóczki, Edit Biró, Fruzsina Zsura, Magdolna Pákáski, and János Kálmán. Automatic detection of mild cognitive impairment from spontaneous speech using asr. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2015-January, pages 2694–2698. International Speech and Communication Association, 2015.

[58] Alexandra König, Aharon Satt, Alexander Sorin, Ron Hoory, Orith Toledo-Ronen, Alexandre Derreumaux, Valeria Manera, Frans Verhey, Pauline Aalten, Phillipe H. Robert, and Renaud David. Automatic speech analysis for the assessment of patients with predementia and alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 1(1):112 – 124, 2015. ISSN 2352-8729. doi: https://doi.org/10.1016/j.dadm.2014.11.012. URL http://www.sciencedirect.com/science/article/pii/S2352872915000160.

[59] Aharon Satt, Alexander Sorin, Orith Toledo-Ronen, Oren Barkan, Ioannis Kompatsiaris, Athina Kokonozi, and Magda Tsolaki. Evaluation of speech-based protocol for detection of early-stage dementia, 08 2013.

[60] Maria Yancheva, Kathleen C. Fraser, and Frank Rudzicz. Using linguistic features longitudinally to predict clinical scores for alzheimer's disease and related dementias. In *SLPAT@Interspeech*, 2015.

[61] Kevin Patrick Murphy. Dynamic bayesian networks: Representation, inference and learning. 01 2002.

[62] N. Linz, J. Tröger, J. Alexandersson, M. Wolters, A. König, and P. Robert. Predicting dementia screening and staging scores from semantic verbal fluency performance. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 719–728, Nov 2017. doi: 10.1109/ICDMW.2017.100.

[63] Anastasios Karakostas, Alexia Briassouli, Konstantinos Avgerinakis, Ioannis Kompatsiaris, and Magda Tsolaki. The dem@care experiments and datasets: a technical report. *CoRR*, abs/1701.01142, 2017. URL http://arxiv.org/abs/1701.01142.

[64] Johannes Tröger, Nicklas Linz, Jan Alexandersson, Alexandra König, and Philippe Robert. Automated speech-based screening for alzheimer's disease in a care service scenario. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare*, PervasiveHealth '17, pages 292–297, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-6363-1. doi: 10.1145/3154862.3154915. URL http://doi.acm.org/10.1145/3154862.3154915.

[65] Kim-Han Thung, Chong-Yaw Wee, Pew-Thian Yap, and Dinggang Shen. Identification of alzheimer's disease using incomplete multimodal dataset via matrix shrinkage and completion. In Guorong Wu, Daoqiang Zhang, Dinggang Shen, Pingkun Yan, Kenji Suzuki, and Fei Wang, editors, *Machine Learning in Medical Imaging*, pages 163–170, Cham, 2013. Springer International Publishing. ISBN 978-3-319-02267-3.

[66] Xiaofeng Zhu, Heung Il Suk, and Dinggang Shen. A novel matrix-similarity based loss function for joint regression and classification in AD diagnosis. *NeuroImage*, 100:91–105, 2014. ISSN 10959572. doi: 10.1016/j.neuroimage.2014.05.078.

[67] Lei Huang, Yan Jin, Yaozong Gao, Kim-Han Thung, and Dinggang Shen. Longitudinal clinical score prediction in alzheimer's disease with soft-split sparse regression based random forest. *Neurobiology of Aging*, 46:180 – 191, 2016. ISSN 0197-4580. doi: https://doi.org/10.1016/j.neurobiolaging.2016.07.005. URL http://www.sciencedirect.com/science/article/pii/S0197458016301282.

[68] Kathleen C. Fraser and Graeme Hirst. Detecting semantic changes in alzheimer's disease with vector space models. 2016.

[69] T. Bschor, K. P. Kuhl, and F. M. Reischies. Spontaneous speech of patients with dementia of the Alzheimer type and mild cognitive impairment. *Int Psychogeriatr*, 13(3):289–298, Sep 2001.

[70] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008. ISBN 0521865719, 9780521865715.

[71] Anand Rajaraman and Jeffrey David Ullman. *Data Mining*, page 1–17. Cambridge University Press, 2011. doi: 10.1017/CBO9781139058452.002.

[72] John A. Bullinaria and Joseph P. Levy. Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and svd. *Behavior Research Methods*, 44(3):890–907, Sep 2012. ISSN 1554-3528. doi: 10.3758/s13428-011-0183-8. URL https://doi.org/10.3758/s13428-011-0183-8.

[73] Richard Socher, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 1201–1211, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=2390948.2391084.

[74] W.Y. Zou, R Socher, Daniel Cer, and C.D. Manning. Bilingual word embeddings for phrase-based machine translation. pages 1393–1398, 01 2013.

[75] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL http://www.aclweb.org/anthology/D14-1162.

Appendix A

Deployment and Client

This appendix presents how the model was deployed to a server and how an Android client was developed to interact with this web server. Figure A.1 shows a diagram that represents the interaction between the Android Client and the Server. In addition, this is just a proof of concept to show that it is possible to build an application that uses the models studied during the thesis, so, a lot of improvements would be needed before launching such application.

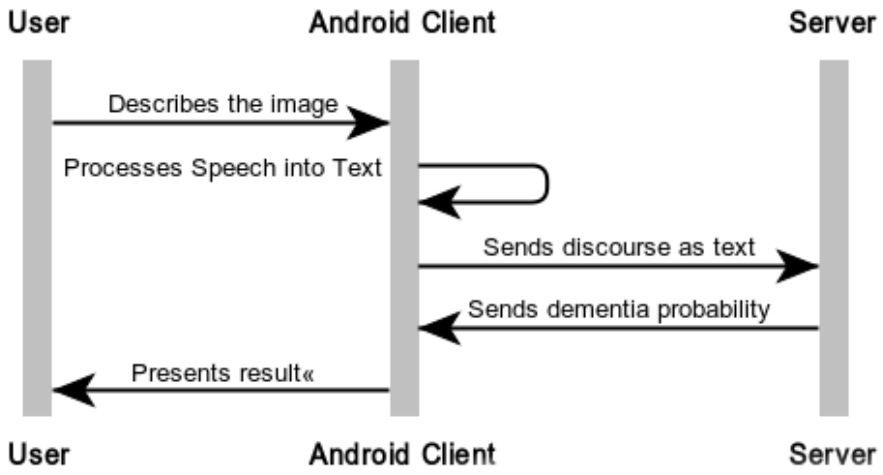

Figure A.1: Interaction between User, Android Client and the Server

A.1 Deployment

The Logistic Regression model described in 5.1 was trained and serialized in order to be used by the server, which was built using Flask. Flask is a lightweight Web Server Gateway Interface web application framework, that makes web development easy and quick, which was the reason we chose it. The server built using Flask expects to receive a JSON with the field ”transcript” and the respective value (the transcript as a string) to the gateway point ”/predict” and returns the probability of having dementia which is given by the serialized model. Figure A.2 shows one requests and the respective response.

```
→ curl -X POST http://localhost:5000/predict -d '{"transcript": "the mother doing the dishes"}' -H "Content-Type:application/json"
{
  "proba": 0.6585349460750317
}
```

Figure A.2: Example of a request to the Flask localhost server and respective probability response.

This server was then deployed in Heroku, which is a container-based cloud platform. Since this service is fully managed by the Heroku company concerning all the infrastructure needed, we just needed to deploy our Flask server to this service in order to have a HTTP server accessible to all devices instead of running the server on localhost. The reader, if curious, can send a request similar to the Figure A.2 to ”http://dementia-model.herokuapp.com/predict”.

Finally, there are some issues with deploying in Heroku. In a free tier account, there is no server redundancy, therefore it does not tolerate server faults. Moreover, its logging capabilities could be better, however, for a proof of concept was sufficient.

A.2 Android Client

Since the server only allows the transcript as text, a client that allows interaction with the users and converts their speech discourse into text. We chose to build this client in Android, however any other platform, that has Speech-To-Text recognition and can make HTTP requests, could be used.

In general, this client shows to the user the Cookie Theft Picture (Figure 2.2) and asks the user to describe the picture. When ready, the user can start describing the picture which, when finished, the speech is transcribed to text and then sent via an HTTP request to the Heroku server. Finally, the client presents to the user the result given by the server.

Concerning the speech recognition, Google provides a Speech-To-Text API called Speech Recognizer that allows access to the speech recognizer made by Google. The implementation of this API allows streaming audio to remote servers to perform speech recognition, which can consume a significant amount of battery and bandwidth. Nevertheless, if the user has downloaded these speech recognition models before, the problem disappears. However, to our knowledge, this API has two major problems:

1. The API stops streaming audio if the user pauses during a considerable amount. This means that, if our user stops to think about more things to say about the picture, the client will stop steaming audio to the speech recognition model, and will send the first portion send to the Heroku Server, ignoring everything the user has said after the long pause.

2. The API will not recognize utterances like filled pauses (e.g. ”ummmm”), which means that a relevant set of words used to train the model will never appear in these transcripts.

Regarding the HTTP Requests, we used Volley[1], which is an open-source HTTP library for Android applications. Volley excels at RPC-type operations used to populate a UI, such as fetching a page of search results as structured data. It integrates easily with any protocol and comes out of the box with

[1] https://github.com/google/volley

support for raw strings, images, and JSON, which is exactly what we need. In general, this library has the downside of not being suitable for large download or streaming operations, since Volley holds all responses in memory during parsing, however that is not important in our case.

In conclusion, despite some limitations, we built a mobile application that works with the model researched during the thesis that provides the result to the user in real time.