Discourse analysis for mental health monitoring

Pedro Rafael Barriga Negra Ascensão
pedro.barriga.negra@ist.utl.pt
L2f/INESC-ID, Rua Alves Redol 9, 1000-029 Lisboa, Portugal
IST/Universidade de Lisboa, Av. Rovisco Pais, 1, 1049-001 Lisboa, Portugal

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 approach, when classifying between different dementia diseases and healthy individuals, 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 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

I. 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 affects primarily senior citizens above the age of 60 where the prevalence ranges between 5%-7%, which increases with age [4]. This is worrisome, since the 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, according to a report made by OCDE in 2017 [6]. The mean of the 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.

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 [7]. Nevertheless, of the 47 million people that live with dementia nowadays, only 25% receive formal diagnosis [8]. 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 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 [8].

One solution for this problem is to develop an automated tool that can diagnose patients by detecting changes on his language. 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 [9], language impairments has been suggested as being more significant than other symptoms [10]. 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 patients speech with success.

The main goal of this project is to develop a system that allows the detection of dementia disease in early stages 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 a transcript 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 transcripts, in which, for example, higher scores mean higher cognitive capabilities, therefore, there is no separation in classes between patients.
II. MEDICAL BACKGROUND

In this section, we present an overview of dementia focusing mainly on Alzheimer’s Disease and Mild Cognitive Impairment since they will be the main focus of this document. 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. Finally, we will explain some diagnostic methods to screen dementia.

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” [11]. Furthermore, since discourse is related to cognitive processes, it is expected that dementia leads to language impairments [12]. These type of impairment is common to all dementias, although they may manifest themselves differently depending of the underlying disease [13].

A. Alzheimer’s Disease

Alzheimer’s Disease is a neurodegenerative disorder that worsens over time, which, as mentioned previously, is the most common cause of dementia, accounting for approximately 60-80% of the cases [3]. The most common early symptom 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, causing death. One of the main problems is that at early stages these symptoms may be mistaken with the natural process of ageing.

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

Semantically, studies indicate that changes in patient’s 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 [14], which make the discourse of the patients seem "empty". Moreover, dementia patients tend to repeat ideas [15] and have difficulties in maintaining a conversation, since their speech is often incoherent [16]. In addition, other functions, such as understanding metaphor and sarcasm, are often impaired in patients with dementia [17]. Finally, studies also show that these patients have difficulties in maintaining a theme throughout the conversation [18].

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 [19]. On one hand, various studies have found evidence for a decrease in syntactic complexity. For instance, Bernard Croisile et al. [20] claimed that AD patients produce fewer subordinate clauses when compared to the control subjects. On the other hand, Daniel Kempler et al. [21] observed that AD patients used the same range of syntactic constructions as control subject.

B. 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, like AD, the patients with MCI has 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 [22]. 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 [22].

Although there is 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. [23] 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. Concerning word-finding difficulty, MCI patients also showed differences when compared with healthy elderly controls. Bennet et al. [24] found out that MCI patients had a lower baseline and declined more rapidly when compared to healthy controls in tests of semantic memory.

Syntactically, Ribeiro et al. [25] 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.

C. Diagnostic

The process of getting diagnosed with almost any type of dementia is far from ideal. In the first place, if the person shows signs of cognitive impairments, the doctor administers a cognitive test. If the patients 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 the diagnosis of Alzheimers [26]. 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. [27]

There are several cognitive tests. One of the most commonly used to screen for dementia is the Mini Mental State Examination (MMSE) [28]. 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 4 categories:

- 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 or training for administration, and has both validity and reliability for the diagnosis and longitudinal assessment of Alzheimer’s disease [29]. 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 [28].

Nevertheless, 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.

III. RELATED WORK

In this section, we provide to the reader the related work on the automatic diagnosis of Alzheimer’s and other types of dementia. With the recent advances in Machine Learning and in the area of Natural Language Processing, authors have been more interested in developing models to automatically detect diseases related to dementia through discourse. As mentioned before, this problem can be seen in two different ways (classification or regression), therefore we are going to divide review of the state of the art into these two categories.

A. Diagnosis as a Classification Problem

The common approach for detecting signs of dementia diseases from speech is to treat it as a classification problem, usually binary classification between control subjects and patients with AD. However, there are authors that instead of using patients with AD, used patients with MCI.

Kathleen Fraser et al. [18] considered 370 features of different types in order to capture 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. Regarding the results, using the top 35 features, a maximum average accuracy of 81.92% was achieved, with a recall of 81.9%. These values make this work the state of the art, since, to our 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. [30] 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 phenomenon common on patients with Alzheimer and with other dementia diseases, as previous studies have shown [31], [32]. The authors divided the figure in study into halves, strips and quadrants, and 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. [18], mentioned previously, increased the F-measure obtained by Kathleen Fraser et al. from 82.4% to 84.6%.

In addition, Maria Yancheva and Frank Rudzicz [14] 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. 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. Using this method, the authors reached some conclusions. Firstly, all of the control clusters words 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).

Although researchers focus 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, as mentioned previously, MCI is less symptomatic when compared to AD.

Vaden Masrani et al. [33] tackled this task using the DementiaBank 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 [34] which has been shown to be effective on a wide range of datasets, and CORAL [35], 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 [18]. 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.
B. 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. [36] 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. This is a bayesian network (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 [37]. 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.

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 [38], [39]. Lei Huang et al. [40] 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.

IV. METHODOLOGY

In this section, we focus on 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.

A. Dataset

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

We used a publicly available dataset called DementiaBank. 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 1 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.

Fig. 1: Image of the Cookie Theft test

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 ??.

Regarding the distribution of samples per patients/controls, 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 2 that there is a misalignment between them. This can influence our results, 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.

Fig. 2: Boxplot representing the age distribution between the two groups

Similarly, it is possible to see in Figure 3 that there is also a difference in the average word length between the two groups. Those with dementia are older and more likely to use shorter words. Therefore, this two characteristics seem to be good features to use in order 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. [18]. However, their distributions were similar between the two groups, which suggests that they are not good characteristics to separate the two groups.
Figure 4 represents the distribution of MMSE score between 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 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 for deciding if a patient is dement or not, all patients with MCI are going to be classified as being healthy. Furthermore, if we try this approach, 25% of Probable AD and 25% Possible AD will be classified as being healthy, which is a problem again. Nevertheless, we are still going to try this approach, since most of patients with dementia are below the MMSE score of 24 if we remove the MCI patients.

For that reason, all the approaches mentioned in this document were built in 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, thus all the transcripts in the dataset are preprocess to obtain the transcripts in their raw format. We remove all human annotations and lower all characters so "The" becomes the same word as "the".

C. Models

We used several models present in the scikit-learn package, 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. The hyperparameters of these models are obtained using a Random Search algorithm, which is described in [1].

D. 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.

V. DIAGNOSIS AS CLASSIFICATION

In this section, we propose and evaluate different approaches to classify patients with AD or not using speech. This includes a bag-of-words approach using Term FrequencyInverse Document Frequency (TF-IDF) and other using word embeddings, described in sections V-A and V-B, respectively. In this document, we will only use patients with AD or healthy subjects, since these are the most predominant classes in our dataset.

A. Bag-of-words approach

Following the work of Kairit Sirts et al. [42], 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 [43]. This value is calculated in our experiment as follow:

$$ tfidf_{t,d} = \log_2(1 + t_{f,t,d}) \times \log_2(1 + \frac{N}{df_t}) $$

where $t$ and $d$ are the term and document for which we are computing a feature, $t_{f,t,d}$ 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 $t$. 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.
By applying TF-IDF to our transcripts, we obtain a sparse matrix (most of the elements are zero) with 551 rows which correspond to our transcripts, and 2000 columns which correspond 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 \(j\).

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 kinds 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 \(\chi^2\) formula. The \(\chi^2\) 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 \(H_0\): the variables are independent.

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

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.

Furthermore, previously, in Section IV-A, we concluded that both the age and the mean average word length were good features to distinguish between control subjects and patients with dementia.

Therefore, Figure 5 represents the Logistic Regression classifier using the TF-IDF approach varying the number of words used using the \(\chi^2\) test, with or without the additional features (age or age and average word length).

![Fig. 5: Accuracy for different combinations of features as we vary the number of words included](image)

By applying TF-IDF to our transcripts, we obtain a sparse matrix (most of the elements are zero) with 551 rows which correspond to our transcripts, and 2000 columns which correspond 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 \(j\).

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 kinds 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 \(\chi^2\) formula.

The \(\chi^2\) 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 \(H_0\): the variables are independent.

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

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.

Furthermore, previously, in Section IV-A, we concluded that both the age and the mean average word length were good features to distinguish between control subjects and patients with dementia.

Therefore, Figure 5 represents the Logistic Regression classifier using the TF-IDF approach varying the number of words used using the \(\chi^2\) test, with or without the additional features (age or age and average word length).

![Fig. 6: Accuracy, AUC and F-measure for each model at their optimum number of words (i.e. the peak performance in Figure 5)](image)

Figure 5 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 6 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.

### B. 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. [44].

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, however, in this case, we only used the mean of all the word vectors present in that transcript.

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. 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 [45] that uses a combined corpus of Wikipedia 2014 plus Gigaword 5 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.

Figure 7 represents the Accuracy across a range of regressors using the GloVe embeddings model with different dimensions using the mean. 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%. Furthermore, 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.

Figure 8 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 BoW 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.

Overall, this approach has worse performance than the bag-of-words 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.

C. 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 \( \chi^2 \) test, when compared to the state-of-art, is, without a doubt, an impressive 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 \( \chi^2 \) test we could only 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.

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. This poor performance may be due to using a written text embeddings model in combination of assigning random vector representations to the words not present in the model. More importantly, we could have tried another model, a model that is trained on spoken text. 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.

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.

VI. Diagnosis as Regression

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

A. Bag-of-words Baseline

Since TF-IDF plus the chi squared test as feature selection had a great performance in the classification task, we used the same method as baseline for predicting the MMSE scores.

As we can see on Figure 9, SVR outperforms the other regressors with a MAE of 3.31. When compared to the work of Maria Yancheva et al. [36] 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.

Be that as it may, our results are still far from the ones obtained by Lei Huang et al. [40], who obtained a MAE of 1.68 using the MRI images, which represents a decrease of the error of 1.63 when compared to our approach. Nevertheless, we still think this a good result, since when solely comparing to the research predicting the MMSE score using spontaneous speech, we have similar or better results while using a simpler approach.

B. LSTM Neural Networks

Long Short-Term Memory networks are a special kind of a Recurrent Neural Network (RNN). Unlike other traditional neural networks, RNN have 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 sequential data contains 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 isn't important (forget gate) or to let it impact the output at the current time step (output gate). This type of networks are well suited to learn from important experiences that have very long time lags in between.

The architecture of the network 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.

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 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.

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 we believe there is space for growth for this approach, mainly if in the future someone releases a bigger dataset of transcripts.

C. 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. 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 10 depicts the performance of this approach using the models studied in the Chapters VI-A and VI-B. 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

\(^1\)https://github.com/keras-team/keras/blob/master/examples/imdb_bidirectional_lstm.py
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 spectrum of dementia diseases, instead of just classifying between Alzheimer and Control, which is a harder task. In addition, both of this 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.

Concluding, in terms of regression, every one of our models underperforms when compared to the state-of-art regressors that uses image recognition for the magnetic resonance imaging exams. In comparison to the work by Maria Yancheva [36], 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.

D. 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 with a grain of salt, 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 $\chi^2$ test proved to be, once again, to be a good approach for the problem, with a MAE similar to the work of Maria Yancheva [36].

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. Additionally, the architecture chosen has an embedding layer that uses the GloVe model, which, as we already discussed on the previous Mean Embeddings approach, is not suitable for this problem, 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 flaws. For example, this model, unlike the ones in the previous chapter, can not detect people with MCI.

VII. Conclusion

Regarding the problem at hand, we used two different strategies: classification and regression. By treating the problem as a classification one, we experimented using two different approaches. The first one was 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. We 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 tried without any success.

By treating the problem as regression, we developed two methods to predict the MMSE score of the patient. Firstly, we 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. [36], which uses the same dataset. A second strategy based on LSTM Neural Networks was tried without much success.

In addition, we used the predicted MMSE to classify patients, as an experiment of what was possible to do with these scores. This approach had a similar performance to the state-of-art or better if we filtered some of the transcripts.

Finally, there are multiple directions we would like to take this work in the future. Foremost, further experiments could be done with the Bag-of-Words approach, such as higher n for the n-grams and adding more features. Furthermore, regarding the embeddings model, we want to test the same approach, however using a different model for the embeddings. Concerning the LSTM Neural Network approach for the prediction of the MMSE scores, we plan to try different architectures with other types of layers and a different embeddings model.