Universiteit Leiden
The Netherlands

Opleiding Informatica

Analyzing Referral Letters

Using Text Mining

Rik Zandbelt

Supervisors:
Suzan Verberne & Semiha Aydin

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

09/07/2019

Abstract

Mental health problems are common, but often not identified correctly or quickly. This research tries to contribute to this by discovering relations between the text in the referral letter for a child and the diagnosis that the child eventually gets. This research will attempt to do this by using topic modeling, classification and chi-squared tests on a dataset of patient records of children. Afterwards, the results are presented to an expert who will analyze them and decide which of them is the most informative.

Contents

1 Introduction

1.1 The situation

Mental health problems are common for both children and adults. According to a research from 2016, approximately 75% of adults with a mental health disorder did already experience issues with their mental health in their childhood or adolescence.[1]
The study has also found that approximately 13.4% of the children and adolescents in the world were experiencing some form of mental health problems. Those problems are likely to have a severe negative impact on their social life, their performance at school or the quality of their life in general. Therefore there is a major need for proper identification and treatment of those mental health issues on children and adolescents.

Although effective treatments for child and adolescent mental health disorders are easily accessible throughout the developed countries, the proper identification of mental health issues raises a big problem. According to the aforementioned study, only 25-35% of children and adolescents with mental health problems are accessing the available treatment.

General practitioners are a big factor in the identification and management of child and adolescent mental health problems, because children in developed countries go to their general practitioner often and the families generally highly value the input of the general practitioner. Unfortunately, general practitioners are often unable to identify a child or adolescent mental health problem. This is partly because children and adolescents often experience different symptoms of those problems than adults do, they may be less willing to share their issues and they are more likely to focus on their physical symptoms than on their mental ones. Furthermore, not every diagnosis has very clear symptoms, which makes identifying a diagnosis even harder.

Even if a mental health disorder is identified correctly, there often are challenges which prevent the patient from getting proper treatment right away. An example of this is that the patients that do get referred to receive a treatment often experience significant delays before they actually receive the treatment. The research has found that all of these challenges occur because there is a lack of confidence in recognizing childhood mental health issues, which is mainly caused by a lack of research on the subject.[1]

Because patients often don't receive the proper treatment quickly, it would be very useful if their diagnosis could be determined as early as possible so they can receive treatment more quickly. Because children with mental health issues are examined by their general practitioner or other referrer before they refer the children to some kind of mental health care, those referrers probably have a lot of information about the children and their referral letter for a child to a mental health care service are therefore very informative. Because of this, it might be possible to identify the diagnosis of a child better based on the referral letter. This research will attempt to investigate this possibility and answer the following question:

"*To what extent can a relation be discovered between the text in a referral letter for a child and a psychiatrist's diagnosis of the child?*"

1.2 Thesis overview

First, section 2 is going to point out the research that has been done in this field before. Then section 3 will describe how the data is collected, section 4 will go over the tasks that I am going to perform in this research and section 5 will describe the dataset. In section 6 the data is then preprocessed and used to perform the tasks and analyze the results. Finally, section 7 will contain a conclusion and pointers on potential follow-up research.

2 Related Work

Some other research has been done on the subject of mental health problems. For example, research has been done that attempts to predict mental health diagnoses of children from Great Britain and Bangladesh using data from the Strengths and Difficulties Questionnaire (SDQ). The SDQ is a brief behavioural screening questionnaire filled in by children, their parents or their teachers.[2] There is also research that used the SDQ to find a bi-directional association between psychological distress and exclusion from school.[3]

Another research tried to predict whether a patient would improve or relapse in their anxiety disorder after they had followed psychotherapy.[4] That research made predictions based on information about the patient such as the type of treatment they received, their marital status and marital tension.

However, apart from the fact that those researches have not been performed recently, they are also all different from the research that I am going to perform. They are especially different because this research is based on referral letters from general practitioners or other referrers, as opposed to the data that is used in the other researches. Furthermore, children will almost always visit their general practitioner first before filling in questionnaires such as the SDQ, which is used as data in researches mentioned before. Because of those reasons, my research could contribute by finding possibilities to predict diagnoses in an earlier stadium by looking at the text in referral letters.

A research that has been performed recently and used a method that is similar to the method that I am going to use in my research, is one that has been done on data from people who suffer from *physical* health problems and are referred to a hospital.[5] This research also examines medical letters using text mining in order to predict diagnoses, but those diagnoses are for physical health problems instead of mental health problems like in this research and such a research can't be found on mental health disorders yet. As mentioned earlier, this is a reason that this research may contribute to the early recognition of children mental health disorders.

3 Background of the Data

This research will attempt to make predictions based on data that is collected from Dutch children that have been diagnosed with a mental health problem by a psychiatrist. The data has been collected from the first time they visited their general practitioner or other referrer (henceforth combined to 'general practitioner') with symptoms related to the mental health issue until the psychiatrist diagnosed the child. Some children in the dataset have already been treated or examined several times and some children are only examined for the first time.

When a child visits the general practitioner, the general practitioner makes notes of the symptoms of the child and about what he thinks the child suffers from, if he or she even thinks the child has a mental health issue. When the general practitioner thinks that the child should be referred to a mental health service for further treatment, the general practitioner combines his or her findings in a referral letter and refers the child to a mental health service.

The child then visits the mental health service which he or she is referred to and together with the parents they decide whether or not he child will be treated there. If the child and the parents give their approval, they fill in a 'Development And Well-Being Assessment' (DAWBA), a novel package of questionnaires, interviews and rating techniques that are designed to generate psychiatric diagnoses on 5-16 year old children.[6]

After filling in the DAWBA, the child gets interviewed by a psychiatrist and afterwards the child is diagnosed. The most important parts of this research are those diagnoses and the referral letters from the general practitioner, as I am going to try and predict the former and I will use the latter to do that.

4 Methods

As stated in section 1, the main research question of this research is:

"*To what extent can a relation be discovered between the text in a referral letter for a child and a psychiatrist's diagnosis of the child?*"

In order to answer this research question, several tasks will be performed on the dataset consisting of data that is collected as described in section 3:

Firstly, I am also going to explore the data by applying topic modeling models on the dataset to investigate which words do often appear together in the referral letters. I am going to look at the comprehensibility of the topics and the coherence between the words in the same topic to judge how well a model performs. Afterwards, an expert in the field of mental health problems will assign labels to the topics in order to find out which diagnosis or diagnosis group is commonly linked to certain words. With this task, the following sub question will be answered:

"*Which words do often appear together in referral letters and is it possible to meaningfully label those groups of words that often appear together?*"

After exploring the data by applying topic modeling, I am going to train machine learning classifiers in order to predict the diagnoses of the children using the text in their referral letters. How well those classifiers perform will be measured by the accuracy score, the precision, the recall, the f1-score and the confusion matrix. Besides looking at how well the classifiers predict the diagnoses, I am also going to investigate the words in the referral letters that the best classifier mostly uses to make predictions and thus which words are commonly linked to each diagnosis. This will be more informative than just creating models to predict the diagnoses, especially if the models don't predict the diagnosis very accurately. With this task, the following sub question will be answered:

"*To what extent can relations be discovered between some relevant words that appear in referral letters and certain diagnoses?*"

Each referral letter got one or more ICPC codes assigned to them by the person who wrote the referral letter. Each ICPC code stands for a certain disorder, symptom or reason for a disorder that is named in the referral letter and all of the ICPC codes assigned to a letter form a complete description of the referral letter. Predicting the diagnosis of a child based on those ICPC codes might be easier than predicting it based on text data because text data is more sparse and the ICPC codes are often clearer than single words. Because of this, I will create a cross table between the frequently used ICPC codes and the diagnoses. Based on that table, I will be able to perform a chi-squared test that will indicate if there are any clear relations between an ICPC code and a certain diagnosis. This task will answer the following and last sub question:

"*To what extent can relations be discovered between the ICPC codes that are given to referral letters and certain diagnoses?*"

Besides answering the sub questions and the main research question, it is also interesting for the field of mental health care to know which of these tasks contributes the most to answering the question. After answering the sub questions and the main research question, I am therefore going to ask an expert to rate the informativeness of the tasks and with that they will answer the sub question of this research:

"*Which of the performed analyses is the most informative for the field of mental health care and why is the chosen analysis the most informative one?*"

5 Description of the Data

5.1 Main dataset

For this research I am using a dataset with data that is collected from children who are referred from a general practitioner to the psychiatry in roughly the way as described in Section 3. Every row in the dataset contains the data of one child and there are 1312 rows in the dataset. Each row in the dataset contains a lot of attributes, most of which are not relevant for this research. The following attributes are relevant for this research:

- The 'SamenvattingJournaal' attribute, which contains the text from the referral letter of the general practitioner.
- The 'hfdclass' attribute, which contains the main diagnosis given by a psychiatrist.
- The 'totdia' attribute, which contains the diagnosis group of the main diagnosis. This attribute has fewer different classes than the 'hfdclass' diagnosis and thus it is a more general version of the main diagnosis.
- The ICPC codes. Each row has five attributes that contain the ICPC codes that are given to the referral letter by the person who wrote the referral letter.

Unfortunately, there are some limitations to the dataset. Firstly, not every row in the dataset has a referral letter or a diagnosis. This limits the amount of rows that I can use for my research, because I need both a referral letter and a diagnosis to include them in the task of predicting the diagnosis. While a respectable 1174 out of the 1312 rows contain a referral letter, only 863 of the total amount of patients have been given a diagnosis. 811 of the rows contain both a referral letter and a main diagnosis, which is only 61.8% of the patients in the already fairly small dataset.

The children that did not a diagnosis have either had no interview with a psychiatrist at all (in which case the diagnosis field is empty) or the psychiatrist has determined that their issue is not a psychiatric one (the diagnosis field in the data is filled with the 'no diagnosis' diagnosis). I left the 'no diagnosis' label out of my research because the label occurs more times than any other label and including it would make the models strongly biased towards it.

Besides the limited number of rows that contain a diagnosis, there is also a limitation to the referral letters. While almost all of the rows in the dataset contain one, they are rather short on average. In the box plot in figure 1, it is shown that the average length of the referral letters is fewer than 50 words. There are some outliers that contain a lot of words, but most of the letters contain only a few sentences.

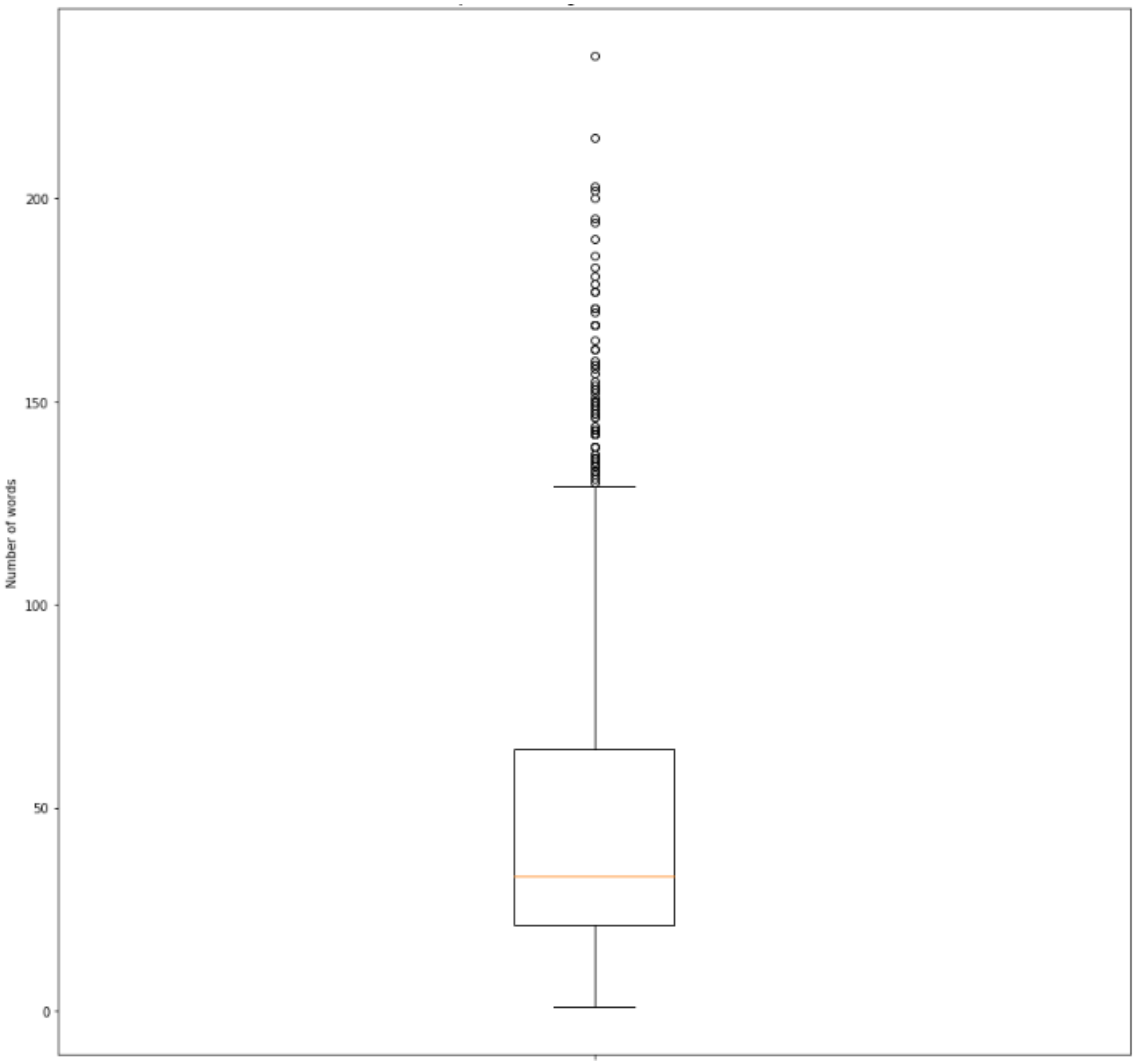

Figure 1: *Box plot of the referral letter lengths.*

Another limitation of the dataset can be found in the distribution of the diagnoses (the 'hfdclass' attribute) in the dataset. As seen in the bar chart in figure 2, some diagnoses appear a lot in the dataset. The three diagnoses that have by far the most appearances are adhd, autism and developmental disorders. This will make the prediction of using machine learning classifiers more difficult because the classifiers will tend to predict that a row has one of those three diagnoses, as a lot of rows in the training set also had one of those diagnoses and because of that the model is biased. Figure 2 also shows that a lot of diagnoses have barely any appearances. The rows that contain a diagnosis that is used less than 10 times cannot be used in the classification process, because they are not represented well enough to be predicted accurately. Leaving out the rows that contain one of those diagnoses reduces our dataset even further, which also makes it more difficult to predict the diagnoses properly.

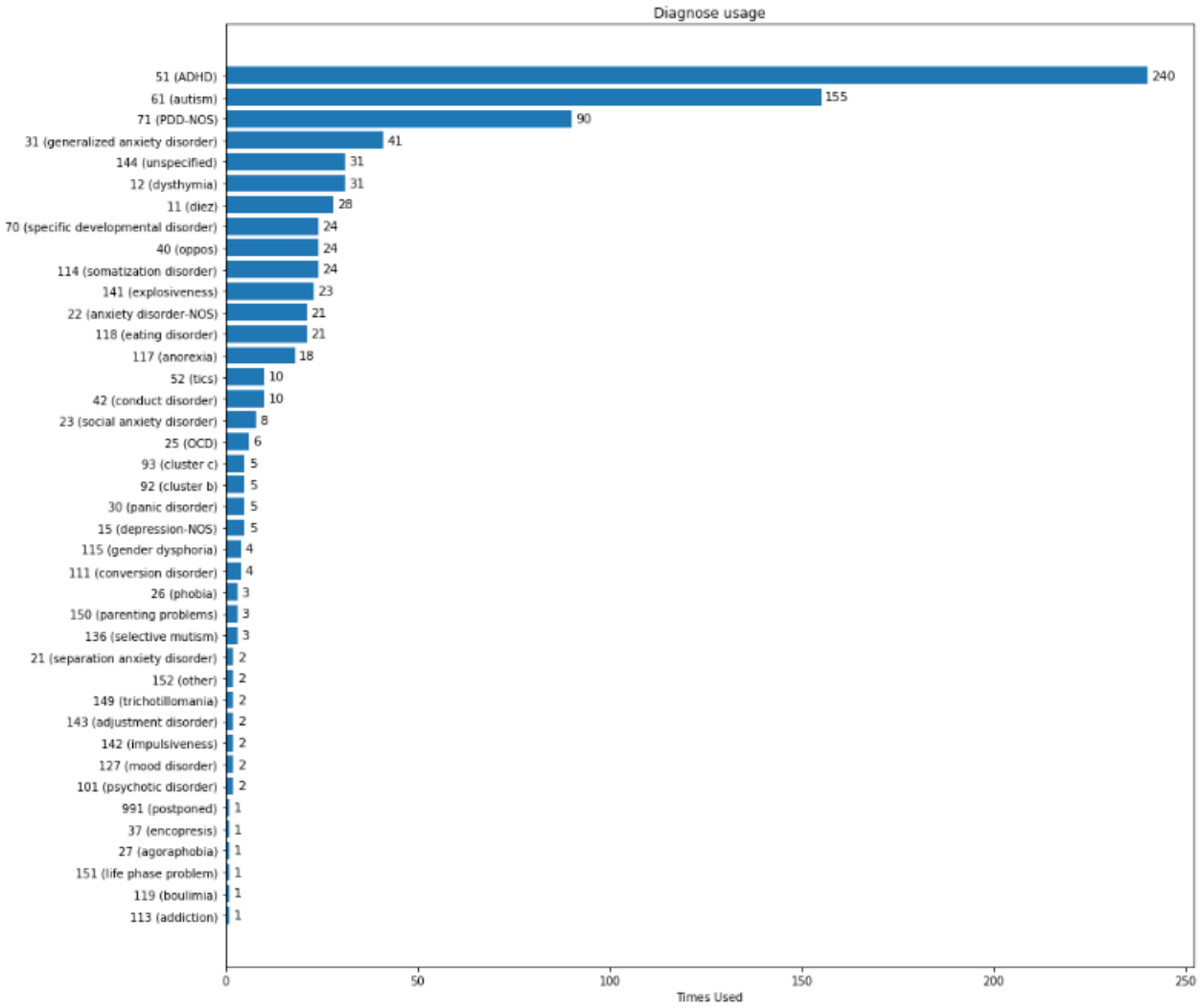

Figure 2: *The diagnosis frequency in the dataset.*

However, the distribution of the diagnosis groups (the 'totdia' attribute) in the dataset is much more balanced. This can be seen in the bar chart in figure 3. As seen in the bar chart, there are a lot fewer classes. This is mostly because some diagnoses have been merged into one diagnosis group. Also, almost all of the classes are well represented and there are only four classes with less than 10 appearances. This will probably cause classification models to be less biased than in the case that they are trained on the 'hfdclass' diagnosis attribute. Furthermore, more of the rows in the dataset can be used in the models because there are only a four diagnoses that are represented too poorly to be used, all of which including less than four rows. There will still be some bias when models are trained with this attribute as target class, because not every diagnosis is used the same number of times. For example, the adhd diagnosis is still appears the most in the data.

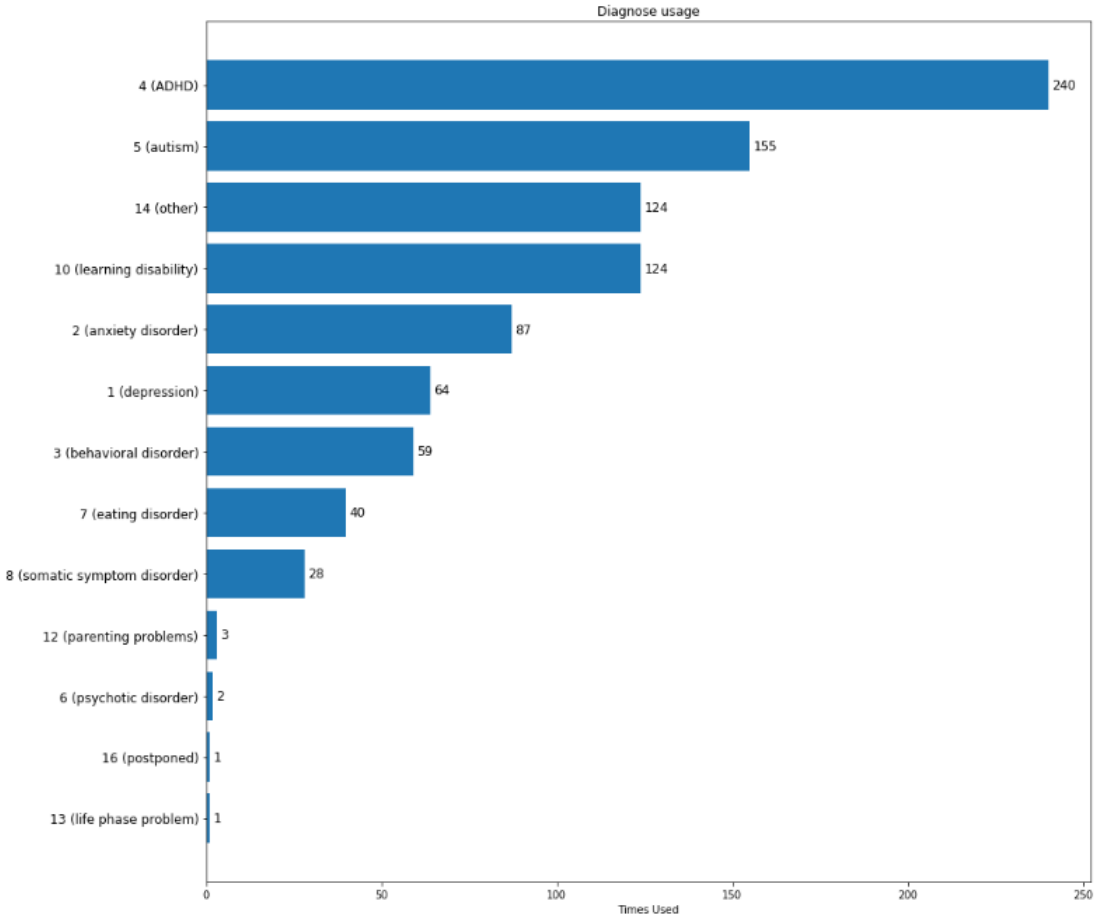

Figure 3: *The diagnosis group frequency in the dataset.*

The limited number of rows in the dataset and the skewed distribution of the diagnoses will presumably make it difficult to create models that can accurately predict many of the diagnoses using machine learning classifiers, although it does seem that the 'totdia' attribute will be a better choice as a target attribute than the 'hfdclass' attribute. Even as the models will probably not be very accurate, it will still be interesting to investigate how the models make their predictions and discover a relation between the text and the diagnoses that way. This has already been discussed in section 4.

Luckily, the limitations mentioned before are not as big of a problem when considering the other tasks that I am going to perform. The limited number of rows will of course also limit the accuracy of the results of the other tasks, but the skewed distribution of the diagnoses will be less of a problem. With topic modeling, the diagnoses will not be used at all because topic modeling does not involve a target class. The chi-squared test that I am going to perform on the ICPC codes does involve the diagnosis classes, but the skewed distribution is not a big problem for those calculations because the calculations for the test do automatically take the skewed distribution into account.

5.2 Sample set with ICPC codes

Besides the main dataset that I described before, I also received a sample set with ICPC codes of 153 rows. Each of the rows of this sample set contains ICPC codes which are assigned to a referral letter by two different coders (the referral letters in the main dataset each have only been assigned codes by one coder), as well as the IDs of those coders (the IDs ranging from 0 to 3). The first coder is always the same coder (coder 0) and the second coder varies between multiple different coders (coders 1, 2 and 3). For each row, both of the coders assigned up to 5 codes to the referral letter (attributes 'ICPC1' to 'ICPC5'). Those coders are experts who were not involved in writing the referral letter, while the ICPC codes in the main dataset are given by the person who wrote the referral letter.

When I received the sample set, I was asked to calculate the inter-coder agreement of the ICPC codes in the set to find out how well they represent the referral letters. The sample set will not be used for anything else in this research, but the calculated agreement can be used in further research.

I calculated the agreement for each ICPC code based on to what extent the second coder agrees with coder 0. If a code from coder 0 is coded in the same spot by the second coder or if it differs by one spot, they have a 100% agreement for that code. If the second coder coded it two spots away from where coder 0 coded it, they have a 75% agreement. If the codes are three or four places apart, they have a 50% agreement and if the second coder didn't code a code from coder 0 at all, they have a 0% agreement on that code. I calculated the average agreement of the 5 codes to get the agreement of the full row.

Because it was agreed between the coders that the codes are ordered in decreasing importance, I assigned weights on each of the codes before calculating the agreement of the first row. The first two codes have a weight of 1, the third code has a weight of 0.5 and the last two codes have a weight of 0.25. The weights and the agreement ratings are provided by coder 0, who knows a great deal about those codes and is convinced that those values should result in a representative agreement value. The calculated agreement scores are shown in the table below:

The average agreement between coder 0 and coder 1	79.04%
The average agreement between coder 0 and coder 2	83.13%
The average agreement between coder 0 and coder 3	82.79%
The average agreement between coder 0 and the other coders	81.65%

The agreement between the main coder and the other coders is high, so the ICPC codes from the main coder are representing the referral letters well.

6 Methods and results

6.1 Preprocessing

Before using the dataset to perform any task, I preprocessed the data (especially the referral letters) in several ways:

The text in the data was converted to unicode in the UTF-8 format, so that there would not be any issues regarding certain special characters that are not in the ASCII table, such as the 'ë' character. This was especially important because the referral letters are written in Dutch, a language that contains many special characters like 'ë'.

All characters in the referral letters were set to lowercase. This is very useful because if it is not done, the computer regards a word with a capital letter (e.g. at the start of a sentence) as a different word than the same word without a capital letter. This would make analyzing the text harder and the results worse.

All punctuation characters are removed from the referral letters. The reason for this is similar to the reason for setting all characters to lowercase: to make the computer treat two equal words as the same word, regardless of punctuation characters they may have attached to them.

I took a Dutch list of 'stop words' from the internet [1] and added more stop words to it myself because they appeared in the results and were not useful. The stop words that I used, can be found in appendix A. The terms I added myself are years, commonly used abbreviations and the single letters of the alphabet. The other stop words are general words that have nothing to do with mental health issues and have no predictive value about the diagnosis. The words in the list of stop words were filtered out of the referral letters, so they would not be used in the tasks or show up in the results.

I have also performed lemmatization on the words in the referral letters. This means that each word is replaced by its stem (e.g. "he walks" is replaced by "he walk"). This is very useful because the texts can be analyzed better if two words that are different forms of the same word, are treated as the same word. This is especially important in small datasets as this one, because many words are already not featured many times in the texts and lemmatization helps with that. I lemmatized the words in the referral letters using the Dutch lemmatizer from the 'Pattern' package in Python.

[1] https://eikhart.com/blog/dutch-stopwords-list

6.2 Topic Modeling

Determining the amount of topics

As mentioned in section 4, I first applied topic modeling on the referral letters. Each topic contains words that often appear together in a referral letter. This way the referral letters will be divided into categories and then an expert will try to assign adequate labels to the topics. The manually assigned labels will indicate to which category the referral letter belongs. It is also interesting to see which words do often appear together. Before performing these tasks, the text in the referral letters is preprocessed as described in section 6.1.

Before creating the topics, it needs to be determined in how many topics the referral letters should be divided. In order to determine this, I will be looking at the coherence value of different models. This means that for every number between 2 and 20, a topic model is created with that number of topics. The coherence value is then calculated for each model, which indicates how often the words in the topics appear together ans thus how good the topics are. For this, the 'CoherenceModel' function from the 'gensim' Python package is used. This function works only for the 'LDA' (Latent Dirichlet Allocation) topic modeling approach, so this is used in the function. I created a graph of all the coherence values, which is shown in figure 4. The graph shows that the coherence does not vary much between the different models, but some models do have higher scores than others. The model with 14 topics has the highest coherence score and after comparing the words in that model with those in other models with relatively high scores, I found the one with 14 topics to be the one with the topics that make the most sense.

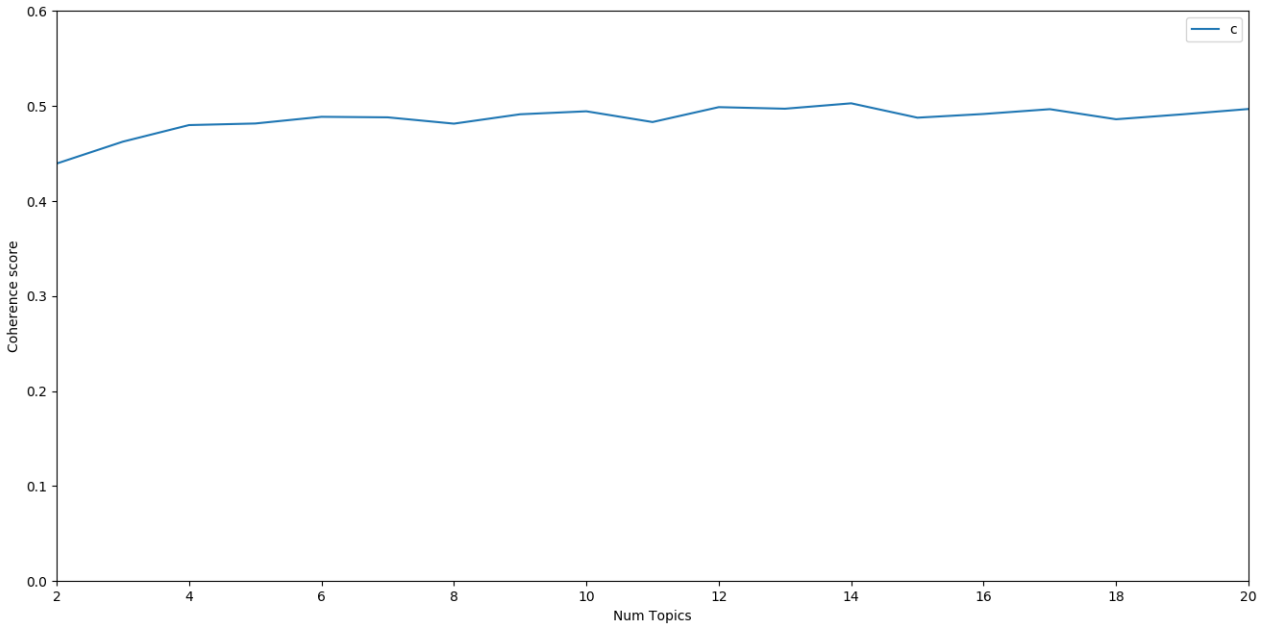

Figure 4: *The graph of the coherence values (c) of the different models.*

Creating the model

After determining the optimal number of topics for LDA, I created two models with 14 topics. One model uses the LDA approach and the other one uses the 'NMF' (Non-negative Matrix factorization) approach. It turned out that the NMF approach is the better one in this case, as the different topics of the LDA model are very similar. In this model, some words were even featured in almost all of the topics. Because the difference between the different topics is important in topic modeling, I chose the NMF model as the final model. The words in the topics of that model are shown in table 1.

Topic	**Words**	**Label (assigned by expert)**
problem group: 1	ADHD, medication, suspicion, evaluation, to start, problem, to diagnose, restless, to come, concentration issues	suspicion of ADHD
2	behavior, difficult, concern, behavioral problem, school, angry, noisy, agressive, accompaniment, restless	difficult behavior
3	trouble, distracted, concentration, restless, attention, social, angry, to sleep, class, head	identify problems
4	dyslexia, ASD, iq, to sign up, NOS, PDD, autism, clinic, developmental case history, learning disorder	developmental issues
5	psychologist, complaints, depressed, mental health care, sadness, thought, death, depression, school absenteeism, fear	complaints (withdrawn child)
referral group: 6	available, further, information, referral letter, brother, present, to examine, psychiatric, to acknowledge, content	basic short referral
7	Curium, referral, to advise, to request, anxiety disorder, to consult, stepmother, diagnostics, mother, municipal health care	more extensive referral (to Curium)
8	to refer, pediatrician, child psychologist, dietician, registration form, medication, advice, intake, clinic, mental health care	referral with history in medical health care
9	treatment, psychologist, mental health care, depressed, eating disorder, anxiety disorder, Rivierduinen, specialist, autistic, disorder	referral with explicitly naming suspicions
10	ADD, presumption, concentration, to determine, evaluation, memory, LUMC, dreamy, medication, to calculate	referral for medication
context group: 11	school, to go, problem, home, bullied, evaluation, dyslexia, group, concentration issues, conversation	context of problems
12	mother, father, very, evaluation, concern, to tell, son, tics, angry, brother	context: family problems
13	trouble, home, family, behavioral problem, noisy, angry, school, agressive, mother, to hit	context: problems at home
14	parents, divorce, problem, family, child, concern, to live, to involve, to argue, education	relationship problems of parents

Table 1: *The 14 topics of the NMF model.*

Analyzing the model

After creating the model, I consulted an expert in order to analyze and label the topics as good as possible. The labels are also visible in table 1 and they are based on the view of the expert, just like the following analysis. The expert saw that there were roughly three types of topics: topics about the problems of a child, topics about the context of the problems and topics about referring the child.

Because there are three types of topics and the coherence value for three topics is not much lower of the coherence value for 14 topics (see figure 4), I also created a model with three topics to see if the same division is visible there. As shown in table 9 in appendix B, there was no such division in that model. This shows that this topic modeling task is very uncertain and the division is not visible in every model. However, as the 14 topic model has the highest coherence value, I am sticking with that model and the three discovered groups of topics.

All of those groups of topics contain a number of topics, and the writing style between those topics seemed to be somewhat different. For example, the 'problems' group consists of the topics 1 to 5. In topics 1, 4 and 5 the suspected diagnosis is explicitly visible, while in 2 and 3 it is not. The expert thinks that this might be due to different referrers having different writing styles.

A similar difference in writing style might be present in the 'referral' group, consisting of topics 6 to 10. The words in some topics (6 and 7) seem to be rather basic, but the other topics in this group contain words that point towards more context in the letter such as history in mental health care.

This leaves topics 11 to 14 for the 'context' group. Those topics seem to be divided by the type of context of the issue.

Unfortunately only a few topics are really clear and this analysis is partly based on guessing. This makes this kind of unsupervised learning not as useful as supervised learning such as classification, but it is still useful to see which words appear together a lot and into which subjects the letters could be divided.

6.3 Classification

6.3.1 Feature extraction

As mentioned in section 4, I am going to train machine learning classifiers in order to predict the diagnoses of the children using the text in their referral letters. After preprocessing the data (see section 6.1), I have to extract features from the referral letters in order to train the models. Since text data is used to train the models, the goal of the feature extraction is to create a term-document matrix. This is a matrix which rows represent the referral letters and the columns represent all the words that appear in at least one of the referral letters and do not appear in the list of stop words (see section 6.1). Each of those words is a feature and the last column, the target class, is the diagnosis given to the patient.

Each value under a column (word) in the matrix indicates how many times the word appears in the referral letter of the corresponding row. Those values serve as the weights of those words and the models will take those weights into account when deciding which words provide good evidence towards a certain diagnosis and which do not. The term-document matrix is sparse, as a lot of the words do by far not appear in all of the letters and thus a lot of values in the matrix will be 0.

The term-document matrix is created by the 'CountVectorizer' function from the 'scikit-learn' Python package. After the CountVectorizer has created the matrix, the 'TfidfTransformer' from the same Python package will adjust the values in the matrix. 'Tf-idf' stands for 'Term Frequency' and 'Inverse Document Frequency'. The former function will adjust the weight of a word for a letter according to the length of the letter, making the frequently appearing words more important than those that rarely appear. The 'idf' function reduces the values in the matrix of words that appear in many different documents. This is because a word that appears in many documents is less informative than a word that only occurs in a few of the referral letters.

6.3.2 Classifier training and evaluation

Now that the term-document matrix has been created, I am going to use it to train different classifiers to create models:

- Multinomial Naive Bayes
- Linear Support Vector Classification
- Stochastic Gradient Descent
- Logistic Regression

Before creating any of those models, the data will first be divided into a training set and a testing set. Then the 'GridSearchCV' function from the scikit-learn package is used to determine the best values of the parameters. The function uses an exhaustive search over a parameter grid to determine the combination of parameter values that give the best results. This is accomplished by using 5-fold cross validation over the training set. After the best parameter values are determined and the model is created, the original testing set is used as a validation set for the model and the results are returned.

The training set is used to train the models for predicting diagnoses by giving each combination of a word and a diagnosis a weight that indicates to what extent the word points to that particular diagnosis. The model will then predict the diagnoses of the testing set and those predictions will form the results of the model. When dividing the data between a training set and a testing set, I am using stratification to make sure that each diagnosis is well represented in both the training and the testing set.

I will train and test each model twice, once with the 'hfdclass' diagnosis as target class and once with the more general 'totdia' attribute as target class. While I have concluded in section 5 that the 'totdia' attribute is more likely to create a model with the least amount of bias, I will still create models with both diagnoses and choose which kind of diagnosis is the best target class. For the best model, I am going to investigate the weights that the model gave to the words and which words got the highest weight for each diagnosis.

6.3.3 Multinomial Naive Bayes

The first classifier that I have trained is the Multinomial Naive Bayes classifier. Naive Bayes classifiers are probabilistic classifiers. They apply Bayes' Theorem and assume that the features are independent. I am using the Multinomial Naive Bayes classifier because there are more than 2 different target classes (diagnoses), which makes this a multiclass or multinomial classification task.

First I trained a Naive Bayes classifier with the specific 'hfdclass' diagnosis as target class. The only parameter I tuned using GridSearchCV was the 'alpha' parameter and the best value seems to be 0.1. As shown in table 2, this first model turned out to be very inaccurate. The accuracy and recall are only 0.329, and the precision and f1-score are even worse (0.217 and 0.238). Those scores are weighted averages of those scores for all of the labels. When looking at the confusion matrix, the clear reason for those low scores shows. Almost all referral letters are predicted to be either diagnosis 51 (adhd) or 61 (autism). As seen in figure 2, those diagnoses appear by far the most in the dataset. This means that the model is biased towards the diagnoses that appear the most.

Target class	**Parameters**	**Accuracy**	**Precision**	**Recall**	**F1-score**
'hfdclass' (specific)	alpha: 0.1	0.329	0.217	0.329	0.238
'totdia' (general)	alpha: 0.05	0.347	0.324	0.347	0.299

Table 2: *The results of the Multinomial Naive Bayes classifier.*

Because the distribution of the 'totdia' diagnoses in the data is less skewed than the distribution of the 'hfdclass' attribute, I thought the classifier with the former as its target class would be less biased and would therefore yield better results. Surely, table 2 shows that the results were indeed better. Apart from the scores, the value of the 'alpha' parameter is also different (0.05 instead of 0.1). The real difference in the results is shown in the confusion matrix of the model: instead of barely predicting any diagnosis other that adhd or autism like the first model, this model does also assign almost all other labels on a number of occasions.

The only label that it does not assign is diagnosis 8 (somatic symptom disorder), which occurs less than any other label (see figure 3). Still, adhd is predicted in most cases, even if it is not the real label. This seems to be inevitable as that diagnosis occurs the most and the Naive Bayes model are at least somewhat biased towards it.

6.3.4 Linear Support Vector Classification

The next classifier that I have trained, is the Linear Support Vector Classification (Linear SVC) model. It is a type of support-vector machine, which is a non-probabilistic classifier (although it can also be used in a probabilistic way). Linear SVC is a linear classifier, as the name suggests. It creates a representation of the data instances as points in a space. The points are mapped in such a way that points with a different label are as far from each other as possible. That way the different labels (in this case diagnoses) are divided by clear gaps in the space. Then new data points are also mapped into the space and their label is predicted based on the cluster of points that they are closest to.

When I trained a Linear SVC classifier with the 'hfdclass' diagnosis as target attribute, I encountered the same issue that I found with the Naive Bayes classifier: the model is biased towards the attributes that occur the most amount of times, being adhd and autism. Again this can be clearly seen in the confusion matrix and it causes the classifier to yield low scores as shown in table 3. Most of the scores are even slightly less than those from the Naive Bayes model. The table also shows that the GridSearchCV function found the best parameter values to be 0.5 for the 'C' parameter and 0.0001 for 'tol'.

Target class	**Parameters**	**Accuracy**	**Precision**	**Recall**	**F1-score**
'hfdclass' (specific)	C: 0.5, tol: 0.0001	0.313	0.229	0.313	0.256
'totdia' (general)	C: 0.5, tol: 0.1	0.318	0.338	0.318	0.292

Table 3: *The results of the Linear Support Vector Classification classifier.*

When I trained the Linear SVC classifier with the 'totdia' diagnosis as target class, the results changed in a similar way as the Naive Bayes classifier. The scores increased and the confusion matrix shows that many more of the different labels are now being predicted in at least one instance. Even the 'somatic symptom disorder' attribute has been predicted now, although not correctly. Compared to the Naive Bayes Classifier, the scores of the Linear SVC model are a bit lower. In this model, the 'C' parameter is the same (0.5) as in the 'hfdclass' model and the 'tol' parameter is set to 0.1. All in all, the Linear SVC classifier performs slightly worse on this dataset in comparison to the Naive Bayes classifier and the scores are again rather low. The improvement when changing from the 'hfdclass' diagnoses to the 'totdia' diagnoses is similar to this improvement in the Naive Bayes classifiers.

6.3.5 Logistic Regression

The next classifier that I trained is a Logistic Regression model. It is a statistical model that is used for predicting categorical labels such as the diagnoses. A Logistic Regression classifier first makes a linear regression model of the data and it then uses a logistic function to predict the target class.

Aside from the usual parameters 'C' and 'tol', which GridSearchCV tuned, the Logistic Regression classifier has another parameter, which is very useful. This is the 'class_weight' parameter. When it is not set , the parameter does nothing. However, it can be set to 'balanced'. This makes sure that the weights that are assigned to the words for a diagnosis, are adjusted so that they are smaller if the diagnosis appears more often in the data and vice versa. This could reduce the bias that the previous models had towards the frequently appearing diagnosis, so I made sure GridSearchCV took this parameter into account as well when trying to create the best model. The Linear SVC model does have this parameter as well, but when I set it to 'balanced' the scores and the bias only got worse.

When I trained the Logistic Regression classifier with the 'hfdclass' attribute, it showed that the scores are better than those from the previous models that I trained with the same target class. Table 4 shows that the accuracy and recall are 0.32 (about the same as those from the Linear SVC model and the SGD model with the same target class) and it also shows that the precision and the f1-score of the Logistic Regression model (0.24 and 0.26 respectively) are higher that those scores from the other models.

The confusion matrix of the model shows that the classifier is also less biased in comparison to the previous ones with the same target class, because this model predicts almost all diagnoses at least once, instead of only predicting the ones that appear the most. As previously discussed, this has probably something to do with the 'class_weight' parameter being set to 'balanced'. Furthermore, there are some other interesting things to be seen in the confusion matrix. It shows that some diagnoses are in fact often being confused for each other by the model. For example, the diagnosis 117 (anorexia nervosa) is often predicted to be diagnosis 118 (eating disorder) and vice versa. Since anorexia is a type of eating disorder, those labels do not differ very much and thus it is to be expected that they are confused with each other. This indicates that it is very hard to predict diagnosis accurately and that it is best to use the more general diagnosis groups as the target class instead. The 'totdia' attribute actually combines all of the eating disorder classes into one class.

Target class	**Parameters**	**Accuracy**	**Precision**	**Recall**	**F1-score**
'hfdclass' (specific)	C: 0.7, tol: 0.0001, class_weight: 'balanced'	0.322	0.244	0.322	0.265
'totdia' (general)	C: 0.3, tol: 0.01, class_weight: 'balanced'	0.363	0.327	0.363	0.312

Table 4: *The results of the Logistic Regression classifier.*

When I did test a model with 'totdia' as the target class, again with the 'class_weight' parameter set to 'balanced', the scores showed that Logistic Regression performs better than the previous models that I have trained and this does also show in the confusion matrix. It is noticeable that there is still some bias towards the adhd diagnosis, but there is less bias than in any of the other models.

6.3.6 Stochastic Gradient Descent

Stochastic Gradient Descent (SGD) is not a classifier, but it is an optimization method for linear classifiers. The scikit-learn package provides the 'SGDClassifier', which is a linear classifier that is enhanced with SGD. SGD typically works well on sparse data. As discussed in the introduction of the classifier section (this section), text data like the referral letters in this data is sparse. Because of this, it seemed likely that the SGD enhanced linear classifier will yield better results than the linear classifier that I tried before, being the LinearSVC classifier. It turned out that the SGD classifier did perform better than the LinearSVC classifier and also when compared to the other classifiers that I tried. Because of that, the confusion matrices of this classifier are shown in the text.

When I trained the SGD classifier with the 'hfdclass' attribute, the scores were higher than any other model with the 'hfdclass' target class (see table 5). Furthermore, the confusion matrix does show a little more variation in the predictions when compared to the first two classifiers. It also shows a confusion between anorexia and the general eating disorder diagnosis like the Logistic Regression classifier also had. The confusion matrix of this model is shown in figure 5. The SGD classifier does have the option of setting the 'class_weight' parameter to 'balanced', but it reacted the same as the other linear classifier (LinearSVC) and the bias actually got worse.

Target class	**Parameters**	**Accuracy**	**Precision**	**Recall**	**F1-score**
'hfdclass' (specific)	alpha: 0.01, tol: 0.001, average: False	0.352	0.256	0.352	0.284
'totdia' (general)	alpha: 0.01, tol: 0.0001, average: True	0.385	0.373	0.385	0.341

Table 5: *The results of the Stochastic Gradient Descent classifier.*

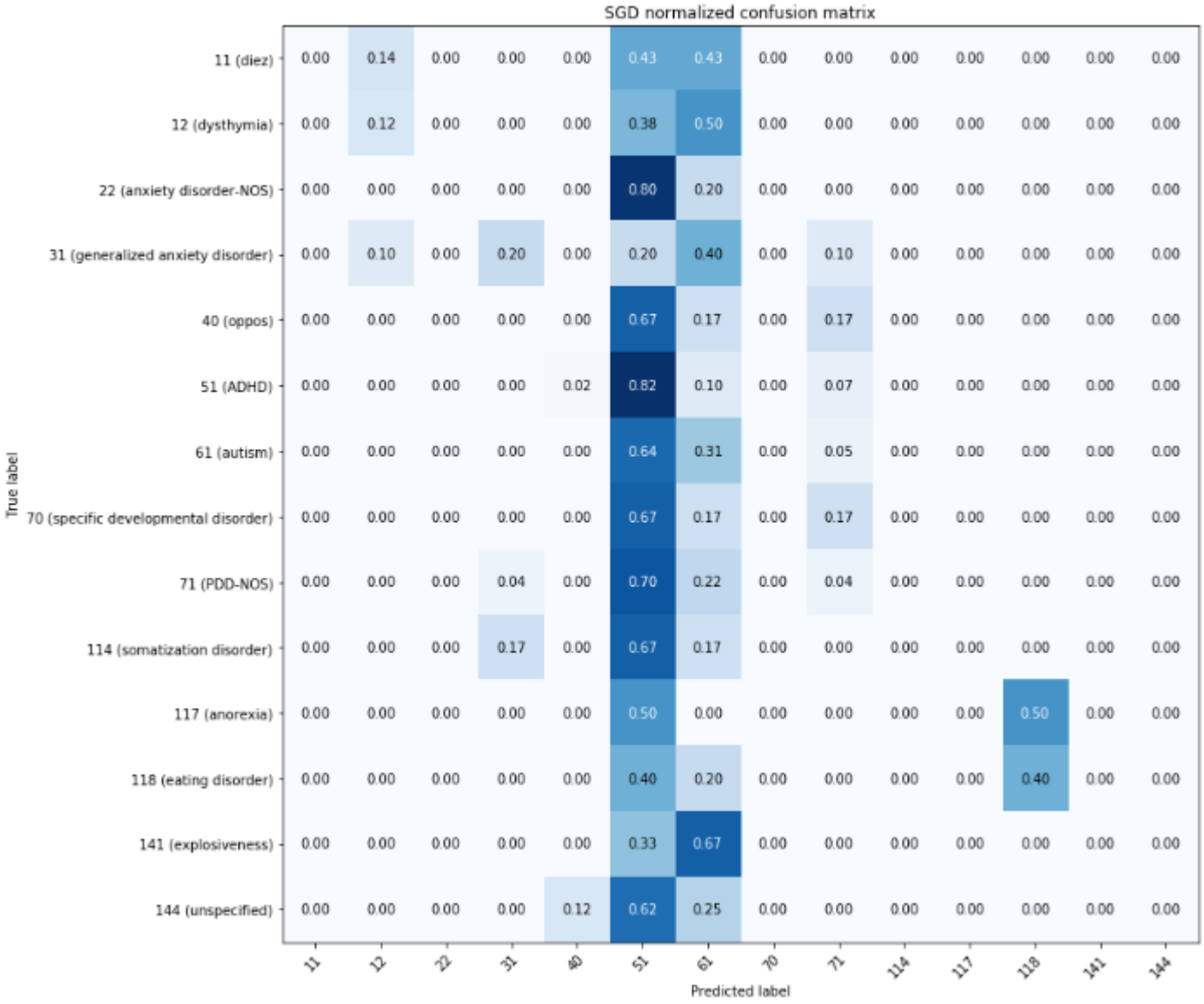

Figure 5: *The normalized 'hfdclass' confusion matrix for the SGD classifier.*

The SGD model with the 'totdia' diagnosis as target class, showed a great improvement in comparison to the model with the 'hfdclass' target class. The improvement is especially visible in the precision and f1-score results, which have risen by about ten percent. Those scores do actually make this model the best one that I have tried. The confusion matrix also shows that the model is at least not as biased as the first two classifiers. The Logistic Regression classifier does also have such an unbiased confusion matrix, but the scores in table 5 make for the SGD model to be the best classifier for this dataset. The confusion matrix of the model is shown in figure 6.

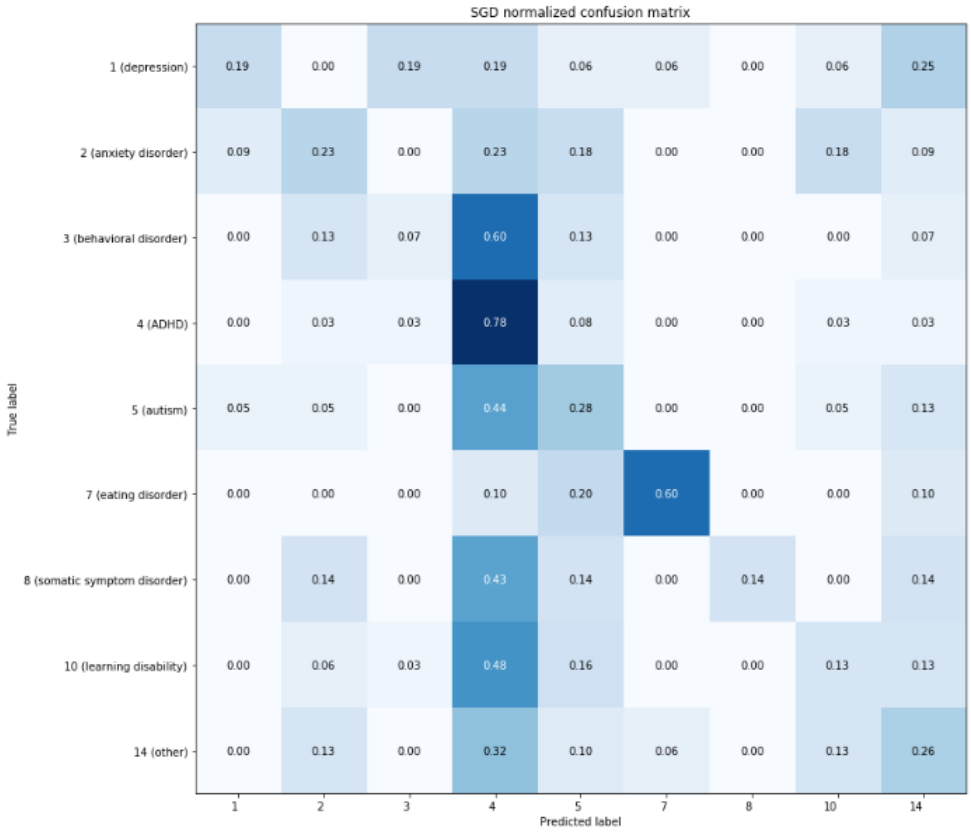

Figure 6: *The normalized 'totdia' confusion matrix for the SGD classifier.*

It says a lot about the difficulty of predicting diagnoses that the best classifier only has an accuracy score of 0.38. However, this classifier still has some potential of predicting diagnoses correctly. Therefore it might be more interesting to look at the way the model predicts the diagnoses rather than merely looking at the results that the classifier yields. Given the mediocre results of the classifier, it is important to keep in mind that the results of the analysis of the way the classifier predicts the diagnoses is not always accurate as well. Despite this, it is still interesting to look at the words in the referral letters which the classifier believes to point strongly in the direction of a certain diagnosis.

6.3.7 Feature analysis

For each diagnosis, the weights of each word that occurs in at least one of the referral letters is stored in the 'coef_' attribute of the classifier. The higher the weight, the more likely it seems to the classifier that the diagnosis is the right one for that letter when that word appears in the letter. I used the 'coef_' attribute to print the ten words that have the highest weight of that diagnosis and thus have the highest predictive value for that diagnosis. Then I translated those Dutch words to English and those translated words and their weights can be seen in table 6. This analysis has been done on the SGD classifier because it performed the best in comparison to the other classifiers that I tried.

ADHD	**Autism**	**Learning Disability**	**Anxiety Disorder**	**Depression**
ADD - 0.57	spectrum - 0.14	tics - 0.22	fear - 0.17	to feel - 0.12
presumption - 0.55	autistic - 0.14	head - 0.18	to dare - 0.13	tired - 0.11
ADHD - 0.54	special - 0.13	dyslexia - 0.15	school absenteeism - 0.12	depressed - 0.10
medication - 0.38	clinic - 0.12	motor skill - 0.14	anxiety disorder - 0.10	sufficient - 0.10
restless - 0.22	specialist - 0.11	to close - 0.11	anxiety complaints - 0.10	to cut - 0.09
teacher - 0.22	to sit - 0.11	quick - 0.10	scared - 0.09	dysthymia - 0.09
to establish - 0.19	remedial educationalist - 0.11	NOS - 0.10 (Not Otherwise Specified)	people - 0.09	day - 0.09
advice - 0.17	adjustment disorder - 0.11	to want - 0.09	panic disorder - 0.09	to live - 0.08
concentration - 0.16	hand - 0.10	distracted - 0.09	exam - 0.08	want - 0.08
suspicion - 0.16	Molemann - 0.09	overview - 0.09	EMDR - 0.08	homework - 0.08

Behavioral Disorder	**Eating Disorder**	**Somatic Symptom Disorder**	**Other Disorders**
PTSD - 0.12	to lose weight - 0.35	chronical - 0.13	child - 0.12
emotion - 0.10	weight - 0.29	stomach ache - 0.09	opinion - 0.12
domestic - 0.09	eating disorder - 0.23	headache - 0.08	middle - 0.11
psychiatric - 0.09	anorexia - 0.16	conversation - 0.07	second - 0.11
alcohol - 0.08	nervosa - 0.15	conclusion - 0.06	info - 0.11
incident - 0.08	to work out - 0.14	line - 0.06	area - 0.11
safety - 0.07	to hospitalize - 0.11	nausea - 0.06	class - 0.10
attachment disorder - 0.07	underweight - 0.10	underlying - 0.06	sound - 0.10
to control - 0.07	Ursula - 0.09	asset - 0.05	to talk - 0.10
crisis relief - 0.07	dietician - 0.08	new - 0.05	back - 0.09

Table 6: *The most informative words and their weights based on the SGD classifier.*

The first noticeable thing in the table is that the most informative words for ADHD have a far higher weight than the words for any of the other diagnoses. This is because the ADHD diagnosis occurs the most, so the model is biased towards the words that occur often in referral letter for children who get diagnosed with ADHD and it gives those words a higher weight. This causes the ADHD diagnosis to be predicted more often because the weight for ADHD increases more easily than the weights for the other diagnoses. The most important terms that have a positive association with ADHD and are not very general terms are 'ADD', 'ADHD' , 'restless' and 'concentration'. If those words occur in a referral letter, the child is likely to get the 'ADHD' diagnosis.

Although the autism diagnosis occurs the most after ADHD, the weights of the most informative words are significantly lower in comparison to those for ADHD. However, most of the words in the list do make sense. The most important, non-general terms are 'spectrum', 'autistic' (often combined to 'autistic spectrum'), 'remedial educationalist' and 'adjustment disorder'.

The weights for the learning disability diagnosis are similar to those for autism. This list features more general words than the previous two, so there are only three terms that are useful for predicting a learning disability. Those terms are 'tics', 'dyslexia', 'motor skill' and 'distracted'. This is completely different for the anxiety diagnosis. Those weights are similar to those for the autism and learning disability diagnoses, but the list features way more useful words. In fact, all of the words in the list for anxiety disorder are useful for predicting that diagnosis. So if any of those terms occur in a referral letter, the child is likely to be diagnosed with some kind of anxiety disorder. It is remarkable that the term 'school absenteeism' is in the list. This may indicate that children with a kind of anxiety disorder are likely to skip their classes.

The weights of the most informative words for the depression diagnosis are low compared to the weights for the diagnoses that I have discussed so far. Although this list contains some general terms that are not very useful on their own, together they do indicate symptoms of depression. The terms 'to feel', 'tired', 'to live' and 'want' indicate the symptoms of being in a tired or sad mood and the lack of a will to live, which are common symptoms for depression. Other important words for depression are 'depressed', 'dysthymia' and 'to cut'. The last word indicates that depressed children often cut or otherwise hurt themselves.

Like the weights of the words for depression, the weights for the behavioral disorder diagnosis are also low. However, most of the terms in the list seem very useful for predicting the diagnosis. The most useful terms are 'PTSD', 'emotion', 'domestic', 'alcohol', 'safety', 'attachment disorder' and 'crisis relief'. Especially the words 'alcohol' and 'domestic' are remarkable. Those words may indicate that behavioral disorders often find their origin in domestic issues or alcohol problems, so it seems like the issues of parents often play a significant part in the development of such a disorder.

The list of most informative terms for the eating disorder diagnosis exists only of terms that have a very logical connection to eating disorders. The term 'Ursula' is the name of a Dutch clinic that treats eating disorders. It is remarkable that the term 'hospitalized' occurs in the list. This may indicate that children who suffer from eating disorders often end up in the hospital. This indicates that it is very important to help children with those disorders as quickly and as good as possible.

The somatic symptom disorder is the diagnosis that occurs the least in the data and this is reflected by the low weights. There are some important features in the list: 'chronic', 'stomach ache', 'headache' and 'nausea'. However, with this little data to work with it is very hard to predict this diagnosis. The 'other diagnosis' diagnosis is also very hard to predict, because it is not known which diagnoses could be in that category and the diagnoses that are in the class could be completely different when compared to each other. The words that occur in the list do not give much information anyway, so I cannot say anything about this label.

6.4 ICPC codes

As mentioned in section 4, the last analysis method is to apply a chi-squared test to attempt to find relations between ICPC codes that are given to the referral letters and the diagnosis. Each row of the main dataset contains up to five ICPC codes, given by the person who wrote the referral letter. To make an analysis of the codes, I made a cross table between the 'totdia' diagnoses and the ICPC codes that are assigned to the referral letter. This cross table is shown in table 7 and it shows how often each ICPC code occurs in combination with a certain diagnosis and how often it occurs in total. As there are a lot of codes that only appear a few times, the table only features the codes that appear at least 10 times.

ICPC/totdia	**1 dep**	**2 anx**	**3 beh**	**4 adhd**	**5 aut**	**7 eat**	**8 ssd**	**10 lea**	**14 oth**	**15 non**	**Total**
P21 (overactive)	5	3	0	**25**	**10**	1	0	6	2	**16**	**68**
R96 (allergic asthma)	7	3	0	6	7	0	1	7	7	8	**46**
P24 (learning problem)	1	1	2	6	9	0	0	**11**	5	6	**41**
S87 (eczema)	3	4	1	7	6	0	1	5	1	8	**36**
P22 (behavioral concerns)	1	1	5	6	6	0	0	4	5	6	**34**
R97 (hay fever)	3	6	0	4	3	0	0	1	3	4	**24**
P74 (anxiety disorder)	0	**12**	3	1	0	0	0	2	2	3	**23**
P99 (other disorders, e.g. autism)	1	0	0	0	3	0	0	8	5	3	**20**
P76 (depression)	**10**	4	0	0	2	0	0	0	1	2	**19**
A12 (allergic reaction)	4	1	1	1	3	1	0	2	1	3	**17**
T06 (eating disorder)	0	0	0	0	0	9	0	1	2	2	**14**
P01 (anxiousness)	1	2	1	0	0	1	1	2	1	3	**12**
P20 (problem with memory/ concentration/orientation)	1	0	0	5	0	1	0	1	2	2	**12**
Total	**37**	**37**	**13**	**61**	**49**	**13**	**3**	**50**	**37**	**66**	**366**

Table 7: *The cross table between the ICPC codes and the 'totdia' diagnoses.*

In order to find out if there are any relations between certain ICPC codes and certain diagnosis, I am going to perform a Chi-squared test for each of the ICPC codes. First, I formulated a null hypothesis that assumes that the occurrences of each ICPC code in the data are randomly distributed among the diagnoses and therefore that there is no association between the ICPC code and some diagnosis:

$H_0 : P_{diagnosis\_m} = P_{diagnosis\_n}$ for each combination of diagnoses m and n.

I also formulated an alternative hypothesis, which states that there is some kind of relation between the ICPC code and some diagnosis and which will be accepted when the null hypothesis gets rejected:

$H_A : P_{diagnosis\_m} > P_{diagnosis\_n}$ for some combination of diagnoses m and n.

I am going to test this hypothesis with $\alpha = 0.05$ and df (degrees of freedom) = 10-1 = 9 (because there are 10 different diagnoses). The formula for the Chi-squared value of an ICPC code is:

$\chi^2 = \sum_{k=1}^{n} \frac{(O_k - E_k)^2}{E_k}$, which is the sum of all values in the row of the ICPC code in table 8. The values in that table are calculated by subtracting the expected value E_k (which is determined by the total occurrences of the code and the diagnosis) from the observed value O_k (the value in table 7), then taking the square and then dividing it by the expected value.

The values in the last column in table 8 are the chi-squared values of each ICPC code. If that value for an ICPC code is higher than the critical value for the aforementioned combination of α and df, the null hypothesis is rejected and it can be concluded that there is some kind of relation between that ICPC code and some diagnoses. The critical value is found in the general critical value table for the chi-squared test [2] and is equal to *16.919*.

ICPC/totdia	**1 dep**	**2 anx**	**3 beh**	**4 adhd**	**5 aut**	**7 eat**	**8 ssd**	**10 lea**	**14 oth**	**15 non**	**Total**
P21 (overactive)	0.5	2.2	2.4	**16.5**	0.1	0.8	0.6	1.2	3.5	1.1	**28.9**
R96 (allergic asthma)	1.1	0.6	1.6	0.4	0.1	1.6	0.9	0.1	1.1	0.0	**7.5**
P24 (learning problem)	2.3	2.3	0.2	0.1	2.2	1.5	0.3	5.2	0.2	0.3	**14.6**
S87 (eczema)	0.1	0.0	0.1	0.2	0.3	1.3	1.6	0.0	1.9	0.4	**5.9**
P22 (behavioral concern)	1.7	1.7	**12.0**	0.0	0.4	1.2	0.3	0.1	0.8	0.0	**18.2**
R97 (hay fever)	0.2	5.4	0.9	0.0	0.0	0.9	0.2	1.6	0.2	0.0	**9.4**
P74 (anxiety disorder)	2.3	**40.9**	6.1	2.1	3.1	0.8	0.2	0.4	0.0	0.3	**56.2**
P99 (other disorders, e.g. autism)	0.5	2.0	0.7	3.3	0.0	0.7	0.2	**10.4**	4.5	0.1	**22.4**
P76 (depression)	**34.5**	2.3	0.7	3.2	0.1	0.7	0.2	2.6	0.4	0.6	**45.3**
A12 (allergic reaction)	3.1	0.3	0.3	1.1	0.2	0.3	0.1	0.0	0.3	0.0	**5.7**
T06 (eating disorder)	1.4	1.4	0.5	2.3	1.9	**144.5**	0.1	0.4	0.3	0.1	**152.9**
P01 (anxiousness)	0.0	0.5	0.9	2.0	1.6	0.9	8.1	0.1	0.0	0.3	**14.4**
P20 (problem with memory/concentration/ orientation)	0.0	1.2	0.4	4.5	1.6	0.9	0.1	0.2	0.5	0.0	**9.4**

Table 8: *The Chi-squared values of the cross table.*

The last column of table 8 shows that codes R96, P24, S87, R97, A12, P01 and P20 do not have any significant relation to some diagnoses, because their chi-squared values are lower than the critical value. The other codes do have a significant relation to some diagnosis. P21 (overactivity) especially has a high chi-squared value for the ADHD diagnosis (diagnosis 4), which makes sense. So the first relation that is established, is the one between the ICPC code for overactivity and the ADHD diagnosis. A relation means that this ICPC code appears a lot in combination with that particular diagnosis. A similar relation can be determined between the P22 code (behavioral concerns) and the behavioral disorder diagnosis (3). The chi-squared value of that combination is high in comparison to the other values in that row. This relation also makes sense. Another discovered relation that makes sense, is the one between the ICPC code for an anxiety disorder (P74) and the diagnosis for the same disorder (2) This one is even clearer than the ones discovered before, as the value is significantly higher (40.9).

[2] http://www.ttable.org/chi-square-table.html

A remarkable relation is the one between the learning disability diagnosis (10) and the P99 ICPC code. The P99 ICPC code stands for all of the mental disorder that are not described by another ICPC code. This code is mostly used to indicate autism, so it is remarkable that the code appears more in combination with learning disabilities than it does in combination with autism.

The P76 code (depression) has a very clear relation with the depression diagnosis (1), which makes sense. However, the relation between the eating disorder code (T06) and diagnosis (7) is by far the clearest with a value of 144.5. This value is very high because both the code and the diagnosis do not appear a lot of times in the data and when they appear, they almost always appear together (there are only a few false negatives).

This chi-squared test has indicated that there are indeed several significant relations between ICPC codes and diagnoses. Most of those relations make a lot of sense, so the conclusion is that the ICPC codes are useful for predicting diagnoses. Because the referrer who writes a referral letter does also assign the ICPC codes, it does also mean that the referrer is often on the right track regarding the identification of the diagnosis.

7 Conclusions and Further Research

Although the dataset that is used in this research is too small to really help identifying diagnoses correctly in the future, all of the analyses that are performed in this research are potentially useful for future research.

First, the topic modeling task has discovered a potential division of the letters into three types of topics: 'problems', 'context' and 'referral'. Those three groups also contain multiple topics with different writing styles and different content. Unfortunately many topics are not very clear, so it was rather hard to meaningfully label them. To confirm the three different groups, future work could perform hierarchical topic modeling on a bigger dataset.

The research on the ICPC codes was useful because it showed interesting relations between the ICPC codes that are given to a referral letter and the diagnosis of that patient. This task has also showed that the ICPC codes could indeed be used to better determine a diagnosis. However, further work needs to be done on bigger datasets to confirm this.

The classification task has yielded the most interesting results. Although the accuracy of the classification models was rather low, further analysis of the best performing model (SGD) identified several potentially useful relations between words and diagnoses. Because of this, the expert has indicated that the classification analysis was the most useful in the field of research on mental health issues of all of the tasks, because the discovered relations between words in the referral letters and the diagnoses are very interesting. Furthermore, the analyses of the other tasks were either more uncertain (topic modeling) or based on fewer data (chi-squared test). The expert has also said that all of the analyses do need to be investigated further with bigger datasets to better substantiate the results, but this research is useful for further research into this topic.

References

[1] D. O'Brien, K. Harvey, J. Howse, T. Reardon, and C. Creswell. Barriers to managing child and adolescent mental health problems: a systematic review of primary care practitioners perceptions. *British Journal of General Practice*, 66:693–707, 2016.

[2] R. Goodman, T. Ford, H. Simmons, R. Gatward, and H. Meltzer. Using the strengths and difficulties questionnaire (sdq) to screen for child psychiatric disorders in a community sample. *British Journal of Psychiatry*, 177:534–539, 2000.

[3] T. Ford, C. Parker, J. Salim, R. Goodman, S. Logan, and W. Henley. The relationship between exclusion from school and mental health: a secondary analysis of the british child and adolescent mental health surveys 2004 and 2007. *Psychological medicine*, 48:629–641, 2018.

[4] R.C. Durham, T. Allan, and C.A. Hackett. On predicting improvement and relapse in generalized anxiety disorder following psychotherapy. *British Journal of Clinical Psychology*, 36:109–119, 1997.

[5] Y. Hu. Automatic icd-10 codes recognition system using machine learning methods. 2019.

[6] R. Goodman, T. Ford, H. Richards, R. Gatward, and H. Meltzer. The development and well-being assessment: Description and initial validation of an integrated assessment of child and adolescent psychopathology. *Journal of Child Psychology and Psychiatry*, 41:645–655, 2000.

Appendices

A Stop words

Words in the list from Eikhart:

a aan aangaande aangezien achter achterna aen af afd afgelopen agter al aldaar aldus alhoewel alias alle allebei alleen alleenlyk allen alles als alsnog altijd altoos altyd ander andere anderen anders anderszins anm b behalve behoudens beide beiden ben beneden bent bepaald beter betere betreffende bij bijna bijvoorbeeld bijv binnen binnenin bizonder bizondere bl blz boven bovenal bovendien bovengenoemd bovenstaand bovenvermeld buiten by daar daarheen daarin daarna daarnet daarom daarop daarvanlangs daer dan dat de deeze den der ders derzelver des deszelfs deszelvs deze dezelfde dezelve dezelven dezen dezer dezulke die dien dikwijls dikwyls dit dl doch doen doet dog door doorgaand doorgaans dr dra ds dus echter ed een eene eenen eener eenig eenige eens eer eerdat eerder eerlang eerst eerste eersten effe egter eigen eigene elk elkanderen elkanderens elke en enig enige enigerlei enigszins enkel enkele enz er erdoor et etc even eveneens evenwel ff gauw ge gebragt gedurende geen geene geenen gegeven gehad geheel geheele gekund geleden gelijk gelyk gemoeten gemogen geven geweest gewoon gewoonweg geworden gezegt gij gt gy haar had hadden hadt haer haere haeren haerer hans hare heb hebben hebt heeft hele hem hen het hier hierbeneden hierboven hierin hij hoe hoewel hun hunne hunner hy ibid idd ieder iemand iet iets ii iig ik ikke ikzelf in indien inmiddels inz inzake is ja je jezelf jij jijzelf jou jouw jouwe juist jullie kan klaar kon konden krachtens kunnen kunt laetste lang later liet liever like m maar maeken maer mag martin me mede meer meesten men menigwerf met mezelf mij mijn mijnent mijner mijzelf min minder misschien mocht mochten moest moesten moet moeten mogelijk mogelyk mogen my myn myne mynen myner myzelf na naar nabij nadat naer net niet niets nimmer nit no noch nog nogal nooit nr nu o of ofschoon om omdat omhoog omlaag omstreeks omtrent omver onder ondertussen ongeveer ons onszelf onze onzen onzer ooit ook oorspr op opdat opnieuw opzij opzy over overeind overigens p pas pp precies pres prof publ reeds rond rondom rug s sedert sinds sindsdien sl slechts sommige spoedig st steeds sy t tamelijk tamelyk te tegen tegens ten tenzij ter terwijl terwyl thans tijdens toch toe toen toenmaals toenmalig tot totdat tusschen tussen tydens u uit uitg uitgezonderd uw uwe uwen uwer vaak vaakwat vakgr van vanaf vandaan vanuit vanwege veel veeleer veelen verder verre vert vervolgens vgl vol volgens voor vooraf vooral vooralsnog voorbij voorby voordat voordezen voordien voorheen voorop voort voortgez voorts voortz vooruit vrij vroeg vry waar waarom wanneer want waren was wat we weer weg wege wegens weinig weinige wel weldra welk welke welken welker werd werden werdt wezen wie wiens wier wierd wierden wij wijzelf wil wilde worden wordt wy wyze wyzelf zal ze zeer zei zeker zekere zelf zelfde zelfs zelve zelven zelvs zich zichzelf zichzelve zichzelven zie zig zij zijn zijnde zijne zijner zo zo'n zoals zodra zommige zommigen zonder zoo zou zoude zouden zoveel zowat zulk zulke zulks zullen zult zy zyn zynde zyne zynen zyner zyns

Terms I added myself:

2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020
a b c d e f g h i j k l m n o p q r s t u v w x y z
pt pte ha ivm vb vwb jgt journaal

B Topic Modeling

Topic	Words
1	available, further, information, referral letter, to stand, present, to confess, brother, to refer, evaluation
2	school, behavior, mother, tired, parents, home, hole, to go, problem, year
3	ADHD, Curium, treatment, referral, diagnosis, to confess, to investigate, to refer, medication, autism

Table 9: *The NMF model with 3 topics.*