Automated Assignment of ICF Functioning Levels to Clinical Notes in Dutch

Jenia Kim
jenka@protonmail.com

December 8, 2021

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The goal of the A-PROOF project is to create AI models that identify the functioning level of a patient from a free-text clinical note in Dutch. These models can be applied to a large quantity of clinical data in order to get insights into, for example, recovery patterns in specific patient populations.

ICF code	Domain	Abbrev.	Functioning levels scale
b1300	Energy level	ENR	0-4
b140	Attention functions	ATT	0-4
b152	Emotional functions	STM	0-4
b440	Respiration functions	ADM	0-4
b455	Exercise tolerance functions	INS	0-5
b530	Weight maintenance functions	MBW	0-4
d450	Walking	FAC	0-5
d550	Eating	ETN	0-4
d840-d859	Work and employment	BER	0-4

Table 1.1: Overview of the ICF domains in the project

We use the functioning categories of the *International Classification of Functioning, Disability and Health* (ICF)[1], a WHO framework for describing and measuring health and disability. Specifically, A-PROOF focuses on 9 ICF

[1] https://www.who.int/standards/classifications/international-classification-of-functioning-disability-and-health

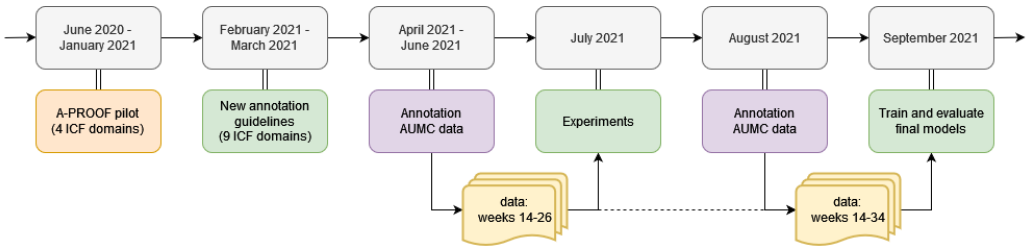

Figure 1.1: A-PROOF: overview of the timeline

categories (domains), which were chosen due to their relevance to recovery from COVID-19; see Table 1.1. For each category, the levels of functioning (qualifiers) are defined on a scale of 0-4 or 0-5; the levels indicate the extent of functioning or disability, where 4 or 5 indicates that there is no problem or limitation, and 0 indicates a total disability (in this domain).

The timeline of the project, from June 2020 to September 2021, is summarized in Figure 1.1. The project started with a pilot (proof-of-concept) phase that focused on 4 ICF domains; this report does not cover the pilot phase. From February 2021 onward, the project has been focused on building the 9-domains system; this includes creating new annotation guidelines, annotating data, experimenting with the classifiers, and building the final machine learning pipeline. These phases are discussed in detail in this report: Chapter 2 describes the annotation process, including the guidelines and the inter-annotator-agreement; Chapter 3 discusses the outputs of this process, i.e. the annotated data; Chapter 4 presents the final classifiers and their performance; Chapter 5 describes the intermediate experiments that were performed before building the final system; Chapter 6 summarizes and discusses the main results and provides suggestions for the next steps of the project. Appendix A contains links to all the publicly available resources created in the course of the project.

Chapter 2

Annotation

2.1 Introduction

The annotation phase lasted 17 weeks between April - August 2021. The annotators are six native Dutch-speaking (para)medical students, who participated in the pilot phase and were already familiar with the goals of the project and the annotation procedure. In addition, healthcare professionals (physiotherapists and dietitians) from the core project team occasionally joined the annotation effort.

The annotation was conducted using the INCEpTION[1] software (Klie et al. 2018), which was installed locally on a server of the Amsterdam UMC to ensure that the sensitive patient data do not leave the hospital's virtual environment.

Section 2.2 describes the annotation guidelines; Section 2.3 discusses the inter-annotator-agreement. For details about the data selection for annotation, see Section 3.2 in the next chapter.

2.2 Annotation Guidelines

The guidelines were created by the core project team, consisting of healthcare professionals and NLP experts. Before the production-phase with the six annotators started, the guidelines were tested on a small sample of notes by the healthcare professionals from the core team. Based on their experi-

[1] https://inception-project.github.io/

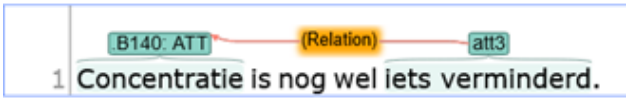

Figure 2.1: Screenshot of an annotated sentence in INCEpTION

ence, a few rounds of update-annotation-discussion occurred, resulting in the final version that was given to the annotators. Based on the first week of annotation (week 14, 2021), the guidelines were updated one last time; from then onward, no significant changes to the guidelines were made (besides small clarifications and adjustments). The full version of the final guidelines (in Dutch) is available on the project's GitHub[2]; below, the main aspects are summarized in English.

Main annotation components

During the annotation process, a clinical note is presented to the annotator in the INCEpTION interface; each sentence in the note appears on a separate line. The annotation consists of the following components:

1. Marking the phrase that indicates that the sentence is about one of the 9 domains with the domain label (e.g. 'ENR');

2. Marking the phrase that indicates the level of functioning with the level label (e.g. 'enr4');

3. Marking sentences that discuss one of the 9 domains, but are not about the current state with the label 'background' (a sentence that describes past functioning) or 'target' (a sentence that describes future functioning);

4. Marking notes that should be completely excluded from the data (e.g. notes about children under 12 years old) with the label 'disregard'.

Figure 2.1 shows an example of an annotated sentence in the INCEpTION interface. The word *concentratie* (concentration) is marked with the domain label ATT (attention functions), the phrase *iets verminderd* (slightly diminished) is marked with the level label att3 (indicating a mild functioning

[2] https://github.com/cltl/a-proof-zonmw/tree/main/resources/annotation_guidelines

problem); in addition, a relation arrow is drawn from att3 to ATT to mark that they belong together.

An important point that was emphasized in the guidelines is that only explicit, literal mentions of functioning levels should be annotated. The annotators were urged to resist the tendency to draw conclusions based on their professional knowledge and to concentrate on the text only. This was implemented as one of the conclusions from the pilot phase of the project.

Domains definitions

The definitions of the 9 domains, as they appear in the annotation guidelines, are given in Table 2.1. These are the original definitions of the ICF; for some of them, the inclusions and exclusions are explicitly mentioned in the guidelines, while for others this was not deemed necessary. The definitions appear in the guidelines in English, since we discovered some inaccuracies in the Dutch translation of the ICF.

Levels definitions

To describe the level of functioning, we use a 0-4 scale for 7 out of the 9 domains. This scale corresponds to the generic qualifiers system of the ICF (World Health Organization 2013); however, note that our scale is reversed: in the ICF scale, 0 indicates no problem and 4 indicates a total disability, while in our scale 4 indicates no problem and 0 indicates a total disability. Table 2.2 and Table 2.3 show the interpretation of the generic qualifiers per domain, as it was provided in the annotation guidelines.

For the *Walking* domain, we implemented an existing domain-specific 0-5 scale instead of the ICF qualifiers: the *Functional Ambulation Category (FAC)*; for the *Exercise tolerance* domain, we used a 0-5 scale inspired by the *Metabolic Equivalents (METs)* scores. The definitions of the levels for these two domains are shown in Table 2.4.

The annotators were instructed to always assign a level of functioning if a domain is discussed. If the level is unclear from the description (e.g. it is mentioned that the patient is fatigued but the degree is not mentioned), the instruction is to assign a middle value (i.e. 2 for the 0-4 scale).

ICF code	Domain	Definition
b1300	ENR	Energy level: Mental functions that produce vigour and stamina.
b140	ATT	Attention functions: Specific mental functions of focusing on an external stimulus or internal experience for the required period of time.
b152	STM	Emotional functions: Specific mental functions related to the feeling and affective components of the processes of the mind.
b440	ADM	Respiration functions: Functions of inhaling air into the lungs, the exchange of gases between air and blood, and exhaling air.
b455	INS	Exercise tolerance functions: Functions related to respiratory and cardiovascular capacity as required for enduring physical exertion.
b530	MBW	Weight maintenance functions: Functions of maintaining appropriate body weight, including weight gain during the developmental period.
d450	FAC	Walking: Moving along a surface on foot, step by step, so that one foot is always on the ground, such as when strolling, sauntering, walking forwards, backwards, or sideways. Include: walking short or long distances; walking on different surfaces; walking around obstacles.
d550	ETN	Eating: Carrying out the coordinated tasks and actions of eating food that has been served, bringing it to the mouth and consuming it in culturally acceptable ways, cutting or breaking food into pieces, opening bottles and cans, using eating implements, having meals, feasting or dining. Exclude: ingestion functions (chewing, swallowing, etc.), appetite
d840-d859	BER	Work and employment: apprenticeship (work preparation); acquiring, keeping and terminating a job; remunerative employment; non-remunerative employment.

Table 2.1: Definitions of the ICF domains

level	generic qualifier	ENR	ATT	STM	ADM
4	no problem	No problem with the energy level.	No problem with concentrating / directing / holding / dividing attention.	No problem with emotional functioning: emotions are appropriate, well regulated, etc.	No problem with respiration, and/or respiratory rate is normal (EWS: 9-20).
3	mild problem	Slight fatigue that causes mild limitations.	Slight problem with concentrating / directing / holding / dividing attention for a longer period of time or for complex tasks.	Slight problem with emotional functioning: irritable, gloomy, etc.	Shortness of breath in exercise (saturation $\geq$90), and/or respiratory rate is slightly increased (EWS: 21-30).
2	moderate problem	Moderate fatigue; the patient gets easily tired from light activities or needs a long time to recover after an activity.	Can concentrate / direct / hold / divide attention only for a short time.	Moderate problem with emotional functioning: negative emotions, such as fear, anger, sadness, etc.	Shortness of breath in rest (saturation $\geq$90), and/or respiratory rate is fairly increased (EWS: 31-35).
1	severe problem	Severe fatigue; the patient is capable of very little.	Can barely concentrate / direct / hold / divide attention.	Severe problem with emotional functioning: intense negative emotions, such as fear, anger, sadness, etc.	Needs oxygen at rest or during exercise (saturation <90), and/or respiratory rate >35.
0	complete problem	Very severe fatigue; unable to do anything and mostly lays in bed.	Unable to concentrate / direct / hold / divide attention.	Flat affect, apathy, unstable, inappropriate emotions.	Mechanical ventilation is needed.

Table 2.2: Definitions of the levels of functioning; part 1

level	generic qualifier	MBW	ETN	BER
4	no problem	Healthy weight, no unintentional weight loss or gain, SNAQ 0 or 1.	Can eat independently (in culturally acceptable ways), good intake, eats according to her/his needs.	Can work/study fully (like when healthy).
3	mild problem	Some unintentional weight loss or gain, or lost a lot of weight but gained some of it back afterwards.	Can eat independently but with adjustments, and/or somewhat reduced intake (>75% of her/his needs), and/or good intake can be achieved with proper advice.	Can work/study almost fully.
2	moderate problem	Moderate unintentional weight loss or gain (more than 3 kg in the last month), SNAQ 2.	Reduced intake, and/or stimulus / feeding modules / nutrition drinks are needed (but not tube feeding / TPN).	Can work/study only for about 50%, or can only work at home and cannot go to school / office.
1	severe problem	Severe unintentional weight loss or gain (more than 6 kg in the last 6 months), SNAQ $\geq$ 3.	Intake is severely reduced (<50% of her/his needs), and/or tube feeding / TPN is needed.	Work/study is severely limited.
0	complete problem	Severe unintentional weight loss or gain (more than 6 kg in the last 6 months) and admitted to ICU.	Cannot eat, and/or fully dependent on tube feeding / TPN.	Cannot work/study.

Table 2.3: Definitions of the levels of functioning; part 2

level	generic qualifier	FAC	INS
5	no problem	Patient can walk independently anywhere: level surface, uneven surface, slopes, stairs.	MET>6. Can tolerate jogging, hard exercises, running, climbing stairs fast, sports.
4	mild problem	Patient can walk independently on level surface but requires help on stairs, inclines, uneven surface; or, patient can walk independently, but the walking is not fully normal.	4≤MET<6. Can tolerate walking / cycling at a brisk pace, considerable effort (e.g. cycling from 16 km/h), heavy housework.
3	moderate problem	Patient requires verbal supervision for walking, without physical contact.	3≤MET<4. Can tolerate walking / cycling at a normal pace, gardening, exercises without equipment.
2	moderate / severe problem	Patient needs continuous or intermittent support of one person to help with balance and coordination.	2≤MET<3. Can tolerate walking at a slow to moderate pace, grocery shopping, light housework.
1	severe problem	Patient needs firm continuous support from one person who helps carrying weight and with balance.	1≤MET<2. Can tolerate sitting activities.
0	complete problem	Patient cannot walk or needs help from two or more people; or, patient walks on a treadmill.	0≤MET<1. Can physically tolerate only recumbent activities.

Table 2.4: Definitions of the levels of functioning; part 3

2.3 Inter-Annotator Agreement

Inter-annotator agreement (also known as inter-rater reliability) is the degree of agreement among independent annotators working on the same task. It is a way to assess how reliable and consistent the 'gold labels' produced by the annotators are. Since the machine learning models are trained on the gold labels, inconsistency in these labels necessarily leads to reduced performance of the models.

In the first 7 weeks of the annotation phase (weeks 14-20, 2021), the weekly annotation batch included between 3 to 5 notes that were the same for all annotators (the annotators were not aware which notes are shared). A sample of the labels assigned in these notes were compared and discussed on a weekly basis in a meeting with all the annotators and the core team members. This procedure was meant to improve the agreement by identifying and discussing difficult or confusing examples. Later annotation batches also contained shared notes, but these were not discussed. All in all, there are 35 notes from all stages of the annotation phase that were annotated by all 6 annotators. These 35 notes (henceforth referred to as 'IAA notes') are used to calculate the quantitative metrics for inter-annotator agreement.

To quantitatively assess the inter-annotator agreement, we use the same metrics that are used to evaluate the performance of the machine learning models: *F1-score* for the domains labels, and *mean absolute error* for the levels labels. The advantage of this approach is that it allows direct comparison between the performance of the model and human performance.

Agreement over domain labels

F1-score is the harmonic mean of the *precision* and the *recall*. In the context of machine learning, these metrics are calculated by comparing two sets of values: gold labels and the labels predicted by the algorithm. Precision is the proportion of cases predicted as positive that are actually positive in the gold standard; it is equivalent to *positive predictive value*. Recall is the proportion of positive cases in the gold standard that were predicted as positive; it is equivalent to *sensitivity*.

These metrics can be used to calculate inter-annotator agreement using the following method (Hripcsak and Rothschild 2005): take a pair of annotators, first treat one of them as 'gold' and the other as 'predictions' and calculate precision, recall and F1-score, then switch the roles and calculate

gold	predict	domain	precision	recall	F1-score
A	K	ADM	0.72	0.58	0.65
K	A	ADM	0.58	0.72	0.65

Table 2.5: Example of pairwise metrics

the metrics again. When this is done for all possible annotator pairs, calculate the average F1-score among all pairs; this average F1-score quantifies the agreement: a higher F1-score indicates better agreement (and thus better reliability of the gold labels).

For example, consider the annotations of the annotators A and K for the ADM domain. In the 35 IAA notes, A labeled 36 sentences as ADM and K labeled 29 sentences as ADM; 21 of those sentences are shared between them, i.e. these are the sentences they agree on.

- When A is treated as the gold standard, there are 21 correctly predicted positives (true positives) out of all 29 predicted positives, i.e. the precision is 21/29=0.72. Out of the 36 gold positives, 21 were correctly predicted, i.e. the recall is 21/36=0.58. The F1-score is 2*0.72*0.58/(0.72+0.58)=0.65.

- When K is treated as the gold standard, the precision and recall are reversed: there are 21 correctly predicted positives (true positives) out of all 36 predicted positives, i.e. the precision is 21/36=0.58. Out of the 29 gold positives, 21 were correctly predicted, i.e. the recall is 21/29=0.72. The F1-score is therefore the same 2*0.72*0.58/(0.72+0.58)=0.65.

The example is summarized in Table 2.5. As we can see, for the pairwise F1-score, it is sufficient to calculate only one direction (e.g. A is the gold).

For 6 annotators, there are 15 possible combinations of pairs. The F1-score per domain for each of the 15 pairs is shown in Figure 2.2; the bottom row shows the mean F1-score that quantifies the overall agreement for this domain.

For the domains ATT, BER and MBW, there are very few annotation examples in the IAA notes. For ATT, the annotators labeled between 1-4 sentences in total, for BER 1-5 sentences and for MBW 3-6 sentences. For this reason, the F1-scores for these domains should be considered cautiously;

annotator1	annotator2	ADM	ATT	BER	ENR	ETN	FAC	INS	MBW	STM
a	k	0.65	0.0	0.33	0.8	0.51	0.69	0.3	0.91	0.5
a	m	0.75	0.57	0.5	0.62	0.53	0.87	0.43	0.67	0.53
a	o	0.63	0.33	0.25	0.64	0.39	0.85	0.087	0.6	0.72
a	s	0.62	0.75	0.57	0.57	0.53	0.96	0.48	0.67	0.45
a	v	0.61	0.57	0.29	0.7	0.54	0.87	0.61	0.73	0.55
k	m	0.65	0.5	0.5	0.59	0.45	0.6	0.36	0.73	0.57
k	o	0.56	0.67	0.5	0.7	0.36	0.62	0.25	0.44	0.61
k	s	0.63	0.4	0.67	0.63	0.41	0.72	0.44	0.75	0.55
k	v	0.62	0.5	0.0	0.76	0.31	0.6	0.44	0.8	0.65
m	o	0.67	0.8	0.67	0.62	0.38	0.8	0.11	0.6	0.64
m	s	0.69	0.86	0.8	0.57	0.36	0.83	0.5	0.44	0.45
m	v	0.58	0.67	0.4	0.73	0.48	0.94	0.43	0.73	0.61
o	s	0.65	0.67	0.4	0.76	0.57	0.8	0.14	0.29	0.46
o	v	0.64	0.8	0.4	0.7	0.42	0.8	0.14	0.22	0.5
s	v	0.67	0.57	0.0	0.55	0.52	0.83	0.33	0.75	0.76
mean		0.64	0.58	0.42	0.66	0.45	0.78	0.34	0.62	0.57

Figure 2.2: Pairwise and mean F1-scores per domain

when the sample is very small, the pairwise scores can vary significantly (e.g. between 0 and 0.86 for ATT) and the mean score is not very informative.

The highest agreement is observed for the FAC domain (number of annotated sentences: 12-17). The mean F1-score is 0.78, but for some pairs the agreement is as high as 0.94-0.96. This suggests that it is generally clear to the annotators which sentences are related to this domain, i.e. there is little confusion regarding what should be considered as 'walking' and what should not. Interestingly, it seems that one annotator might have interpreted this domain differently than the others: all scores below 0.8 for this domain involve the annotator K. For example, the sentence *Kan enkele pasjes aan de hand lopen* (Can walk a few steps with help) was annotated by everyone except K as FAC, while the sentence *Mobiliteit: zie fysiotherapie, vandaag voor het eerst gestaan met 2 personen* (Mobility: see physiotherapy, stood for the first time today with the help of 2 people) was annotated as FAC by K, but not the others.

The lowest agreement rate is observed for the INS domain (number of annotated sentences: 2-21). The F1-score is low for all the pairs, indicating that it is unclear to the annotators which sentences should be considered as describing exercise tolerance and which should not. This difficulty was also expressed by the annotators themselves during the weekly meetings. Specifically, it was not clear to the annotators whether mentions of activities (e.g. that a patient plays football once a week) should be labeled as INS. In addition, there is a lot of overlap between the INS domain and the FAC, ADM and ENR domains; this was also often mentioned as a cause for confusion. For example, the sentence *Was zelf naar de WC gelopen, en was daarna uitgeput* (Walked to the WC by herself and was exhausted afterwards) describes walking (the patient can walk independently) and energy level (the patient is fatigued after a short walk); the confusing question is whether it should also be considered as INS (the patient cannot tolerate a short walk, i.e. can tolerate only sitting activities). Although this domain was confusing for all annotators, it seems that it was especially confusing for the annotator O: all scores below 0.3 are of pairs involving O.

ETN (number of annotated sentences: 7-25) is another domain for which the agreement is poor across all the pairs. This domain was not mentioned by the annotators as especially confusing or unclear, so the cause of the disagreement is not immediately evident. For example, the sentence *2x koffie, 750 ml water, 500 ml appelsap, vlaflip, 3x NCP* (2x coffee, 750 ml water, 500 ml apple juice, *vlaflip* (custard-like dessert), 3x *NCP* (Nutridrink Compact

Protein, brand name of a nutrition drink)) contains a mention of a nutrition drink (explicitly mentioned in the guidelines, and thus familiar to all the annotators), but it was labeled as ETN only by one out of the 6 annotators. Similarly, *Abdominaal en voeding: Sondevoeding weer herstart* (Abdominal and nutrition: Tube feeding restarted) explicitly mentions tube feeding but was labeled as ETN only by half of the annotators.

For ADM (number of annotated sentences: 21-36), ENR (number of annotated sentences: 9-18) and STM (number of annotated sentences: 10-18), the mean F1-scores range between 0.57 and 0.66. This moderate agreement suggests that the definitions for these domains are fairly clear, but still allow for quite a lot of personal interpretation. The agreement seems to be similar across all pairs (especially for ADM), meaning that no individual annotator has especially divergent views on these domains.

To conclude, assignment of domain labels seems to be a generally complicated task. The average F1-score is between 0.42 and 0.78; only 4 out of the 9 domains have agreement above 0.6. Out of the domains that have enough examples in the IAA notes, INS and ETN are especially problematic. The definitions and guidelines for these domains need to be examined and discussed further.

Agreement over level labels

The metric used to assess the agreement about the levels of functioning is *mean absolute error*: an arithmetic average of the absolute errors. For example, if one annotator labeled a sentence as att3 and the other labeled the same sentence as att2, the absolute error is $\mid 3 - 2 \mid = 1$ (or $\mid 2 - 3 \mid = 1$, the direction does not matter). Averaging over the absolute errors for all ATT sentences that these pair labeled, we get the pairwise mean absolute error (MAE). The smaller the MAE, the better the agreement is.

Figure 2.3 shows the pairwise MAE for all possible combinations of pairs; the bottom row shows the overall mean for the domain. Cells with 'N/A' mean that this pair did not have any shared sentences that both labeled with level labels for this domain; cells with 0 mean that there is perfect agreement between the two annotators regarding the levels for this domain.

Similarly to the domains labels, there were not enough examples in the IAA files of ATT_lvl (1-3 labeled sentences in total), BER_lvl (1-3 labeled sentences in total) and MBW_lvl (1-4 labeled sentences in total). Therefore, the MAE's for these domains should be considered with caution.

annotator1	annotator2	ADM_lvl	ATT_lvl	BER_lvl	ENR_lvl	ETN_lvl	FAC_lvl	INS_lvl	MBW_lvl	STM_lvl
a	k	0.23	N/A	0.0	0.2	0.27	0.0	0.0	0.5	0.22
a	m	0.19	0.0	0.5	0.44	0.56	0.27	0.3	0.0	0.44
a	o	0.18	0.0	1.0	0.43	0.33	0.33	0.0	0.0	0.39
a	s	0.2	0.33	0.5	0.44	0.5	0.18	0.4	0.5	0.49
a	v	0.32	0.5	0.0	0.18	0.25	0.17	0.6	0.33	0.29
k	m	0.34	1.0	1.0	0.44	0.2	0.0	0.0	0.5	0.26
k	o	0.37	N/A	1.0	0.29	0.25	0.0	0.0	0.5	0.048
k	s	0.33	1.0	1.0	0.25	0.33	0.1	0.0	0.67	0.35
k	v	0.26	1.0	N/A	0.18	0.2	0.22	0.6	0.75	0.33
m	o	0.091	0.0	0.0	0.6	0.33	0.1	0.0	0.0	0.12
m	s	0.28	0.33	0.0	0.88	0.38	0.1	0.38	0.33	0.36
m	v	0.24	0.0	0.0	0.38	0.46	0.29	0.5	0.25	0.22
o	s	0.23	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.083
o	v	0.1	0.0	0.0	0.25	0.0	0.36	1.0	0.0	0.43
s	v	0.34	0.0	N/A	0.44	0.12	0.36	0.71	0.5	0.58
mean		0.25	0.32	0.38	0.39	0.28	0.17	0.3	0.32	0.31

Figure 2.3: Pairwise and mean MAE per domain

Overall, there is very good agreement about the levels, across all domains and all pairs. The mean MAE's per domain are all below 0.4, which is very good, both for the 5-level scales (0-4) and especially for the 6-level scales (0-5). This result indicates that the definitions of the levels in the annotation guidelines are clear and that the levels are easily distinguishable from each other. The main difficulty of the annotation task lies in the detection of the domains, not in the assignment of functioning levels.

Chapter 3

Data

3.1 Introduction

The data for the project consists of notes from the electronic health records of the Amsterdam UMC (both the AMC and the VUmc locations). Specifically, we had access to about 4 million notes from 2017 (both locations), about 2 million notes from 2018 (AMC location only), and about 2 million notes from the first three quarters of 2020 (both locations). Due to to the project's interest in COVID-19, the 2020 data was split into two subsets: notes that belong to patients with a COVID-19 diagnosis (cov-2020), and notes that belong to patients that do not have a COVID-19 diagnosis (non-cov-2020).

From all available notes, a subset was selected for annotation; the selection procedure is detailed in Section 3.2. Section 3.3 provides descriptive statistics about the annotated data. Section 3.4 provides statistics about the data that was annotated during the pilot phase of the project.[1]

3.2 Selecting Data for Annotation

The gold-labeled data obtained from the annotation process is used for training and evaluating the machine learning models. Therefore, it should ideally have the following characteristics:

[1]Even though the pilot project is not discussed in this report, the positive examples from the pilot annotations are used in the machine learning model for domain classification (see Chapter 4), and therefore the statistics are provided.

1. There should be a sufficient number of positive examples, i.e. sentences with domain and level labels. The goal set at the beginning of the project was to obtain about 15,000 sentences with labels.

2. The labels should be well-distributed, i.e. there should be enough labels for each of the 9 domains.

3. The examples should be diverse, i.e. contain all possible phrasings that are relevant to discussing a domain in a range of different clinical notes.

To facilitate the first goal, a keyword-based search was implemented in the data selection procedure. A list of keywords related to each of the domains was compiled by the members of the core team, based on their professional knowledge and experience; the (various versions of the) list can be found on the project's GitHub[2]. For each annotation batch, the following parameters for the keyword search could be configured:

- The proportion of notes in the batch that contains keywords; the rest of the notes are selected randomly to ensure diversity (the risk in selecting all notes with keywords is that it might create a bias towards specific phrasings).

- The specific domains whose keywords should be used; this allows to control the distribution of the obtained labels (for example, if a domain is under-represented, we can try to obtain more examples by searching specifically for notes that contain keywords for this domain).

- The minimum number of matched domains in a note; i.e. only notes that contain keywords from at least X different domains are selected. This is done to ensure that the annotated notes are "rich" in relevant sentences.

In addition to these keyword-related parameters, the batch could be configured in terms of the proportion of COVID data (i.e. notes of patients with a COVID-19 diagnosis) that it contains and the type of notes that are selected (in general, there are about 60 different note types in the data).

Table 3.1 summarizes the settings used for all of the batches throughout the project. In the first weeks, 50% of the notes selected for annotation were related to COVID patients, 80% of the notes were selected with keywords

[2] https://github.com/cltl/a-proof-zonmw/tree/main/resources/keywords

Batch	% COVID	% Kwd	Kwd version	Matched doms	Min matched doms	Note types
w14 - w15	0.5	0.8	v2	all	4	all
w16 - w19	0.5	0.8	v3	all	4	all
w20 - w22(a)	0.3	0.7	v3	ATT, BER, MBW	1	*Consulten (niet-arts)*
w22(b) - w26	0.3	0.7	v4	all except ADM	3	all
w27 - w34	0.4	0.8	v4	all except ADM	3	all

Table 3.1: Data selection parameters for different annotation batches

(of all 9 domains, so that each note contained at least 4 different domains), and all note types were selected. After the first 4 weeks of annotation, the annotated data was analyzed, with the following findings:

- The distribution of the labels obtained in the first 4 weeks is highly imbalanced: the ADM domain is very dominant (41% of all labels), while the ATT, BER and MBW domains are very rare (2%-4% of all labels).
- The proportion of the ADM domain is especially big in the COVID data (49% of all labels).

Based on these findings, it was decided to experiment with different data selection settings for 3 weeks; see 'w20-w22(a)' in Table 3.1. The goal of the new settings was to try to increase the proportion of the ATT, BER and MBW domains in the annotated data and decrease the proportion of ADM:

- The proportion of COVID notes is reduced to 30%.
- Only keywords for ATT, BER and MBW are used for the keyword selection (a note is selected if it contains keywords from at least one of these domains).
- Only notes of type *Consulten (niet-arts)* (Consultations (non-doctor)) are selected. The idea is that the 3 under-represented domains are likely to be discussed in non-doctor consultations: MBW is likely to be discussed in dietitian notes, ATT is likely to be discussed in occupational therapists notes, and BER is likely to be discussed in social workers notes.
- The proportion of randomly-selected notes is increased to 30% to balance the effect of the other adjustments.

Analysis of the labels collected in weeks 20-21 revealed that the adjusted settings helped to collect a substantial quantity of MBW and ETN labels, but the number of ATT and BER labels remained low. This is partially due to the fact that the note type *Consulten (niet-arts)* turned out to contain a lot of dietitian consultations.

After this experiment, and until the end of the annotation phase, we went back to selecting from all types of notes. To temper the dominance of

dataset	N Total	N Annotated	N Disregard	% Disregard
2017	1,262	1,089	173	14%
2018	1,225	1,053	172	14%
cov_2020	2,399	2,340	59	3%
non_cov_2020	1,278	1,072	206	16%
total	6,164	5,554	610	10%

Table 3.2: Weeks 14-34: Number of annotated notes (incl. disregard)

ADM labels, the proportion of COVID data was kept to 30-40% and ADM keywords were not used in the keyword search. In addition, the keyword list for ATT, BER and MBW was updated with new keywords based on the annotations so far.

3.3 Annotated Data

This section describes the final outputs of the annotation phase that took place in weeks 14-34, 2021.

Number of Notes & Number of Sentences

In total, about 6,000 clinical notes were annotated; 10% were marked 'disregard' and therefore removed from the final dataset (disregarded notes include, for example, notes about children under 12 years old, listings of medications, etc.). Table 3.2 shows the number of total and 'disregard' notes per dataset; as evident, there are significantly less 'disregard' notes in the COVID dataset (3%), compared to the other datasets (about 15%).

The 5,554 non-disregard notes contain in total about 286,000 sentences; 5% out of these sentences contain at least one domain label (see Table 3.3). This means that about 15,000 sentences with domain labels were obtained in the current annotation effort, in accordance with the goal that had been set at the beginning of the project.

Distribution of Domains

Figure 3.1 shows the total number of labeled sentences per domain. Note that a sentence can contain more than one label, and therefore some sentences are

dataset	N total	N with labels	% with labels
2017	52,454	2,339	5%
2018	53,187	2,249	4%
cov_2020	124,418	7,803	6%
non_cov_2020	56,056	2,295	4%
total	286,115	14,686	5%

Table 3.3: Weeks 14-34: Number of sentences: total and with domain labels

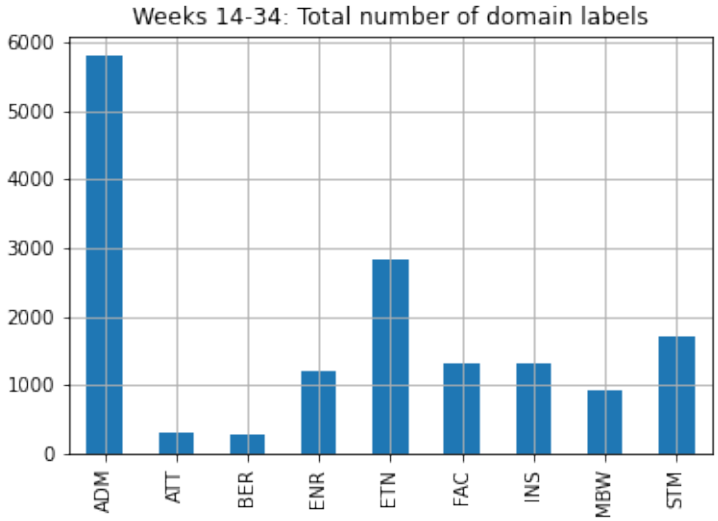

Figure 3.1: Weeks 14-34: Number of labeled sentences per domain

dataset	ADM		ATT		BER		ENR		ETN		FAC		INS		MBW		STM	
	n	%	n	%	n	%	n	%	n	%	n	%	n	%	n	%	n	%
2017	567	23	58	2	59	2	221	9	527	21	246	10	187	7	253	10	389	16
2018	597	25	41	2	66	3	194	8	501	21	274	11	232	10	198	8	308	13
cov_2020	3,940	47	142	2	91	1	591	7	1,376	17	541	7	694	8	314	4	626	8
non_cov_2020	700	29	60	2	52	2	186	8	419	17	257	11	214	9	150	6	395	16
total	5,804	37	301	2	268	2	1,192	8	2,823	18	1,318	8	1,327	8	915	6	1,718	11

Table 3.4: Weeks 14-34: Distribution of domains

counted more than once. Despite the attempts to balance the distribution of the labels, as described in Section 3.2, ADM is still very dominant in the data (∼5,800 sentences are labeled as ADM), while ATT and BER are rare (∼300 sentences for each). The rest of the domains are more balanced: ∼1,000 sentences for MBW, FAC and INS, ∼1,700 sentences for STM and ∼2,800 sentences for ENR.

Table 3.4 shows more details about the distribution of domains per dataset. In each row, we can see the percentage of a specific domain out of all labeled sentences in a dataset. For example, 47% of the labels in the COVID dataset are ADM labels; this is very different from the other datasets, where ADM comprises between 23-29% of the labels. For the rest of the domains, the differences between the datasets are less significant.

Distribution of Levels

Figure 3.2 and Figure 3.3 show the distribution of the level labels in each domain, per dataset. A few observations can be made:

- ADM: the COVID data shows a different distribution, compared to the other 3 datasets; there are a lot more 0 and 1, and a lot less 4. However, when all datasets are considered together, the distribution of the different levels is quite balanced.
- ATT: level 2 is very dominant, in all datasets.
- BER: levels 4 and 0 are dominant, in all datasets.
- ENR: levels 1 and 2 are dominant, in all datasets.
- FAC: level 4 is very dominant, in all datasets.
- INS: the COVID data shows a different distribution, compared to the other 3 datasets; there are more 0 and 1, and a lot less 4 and 5.
- MBW: there almost no examples of level 0.
- STM: level 2 is very dominant, in all datasets.

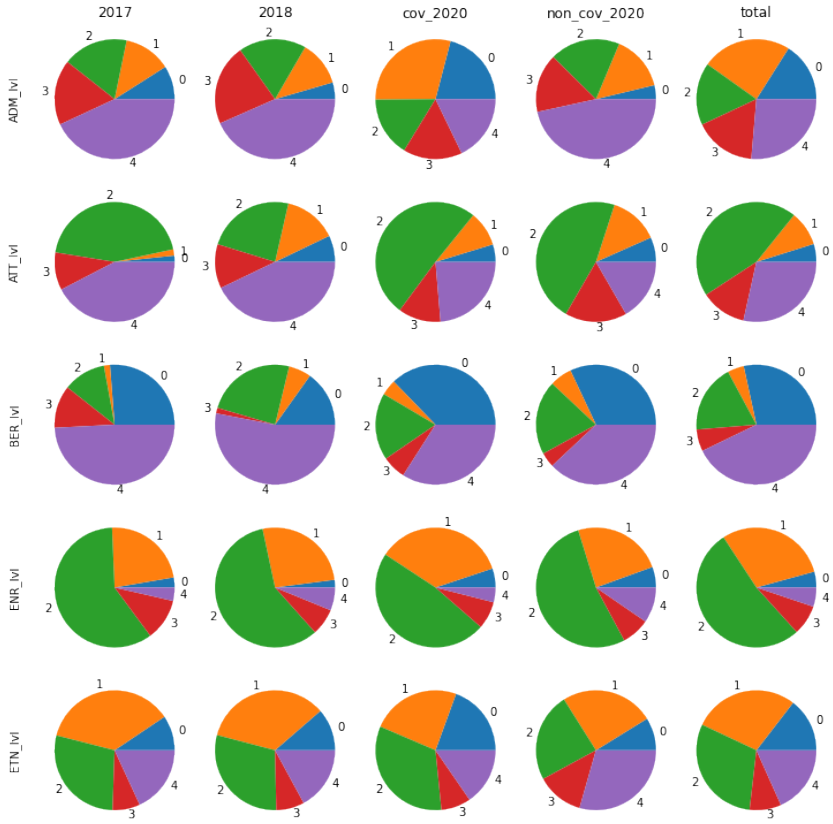

Figure 3.2: Distribution of levels per domain (Part 1)

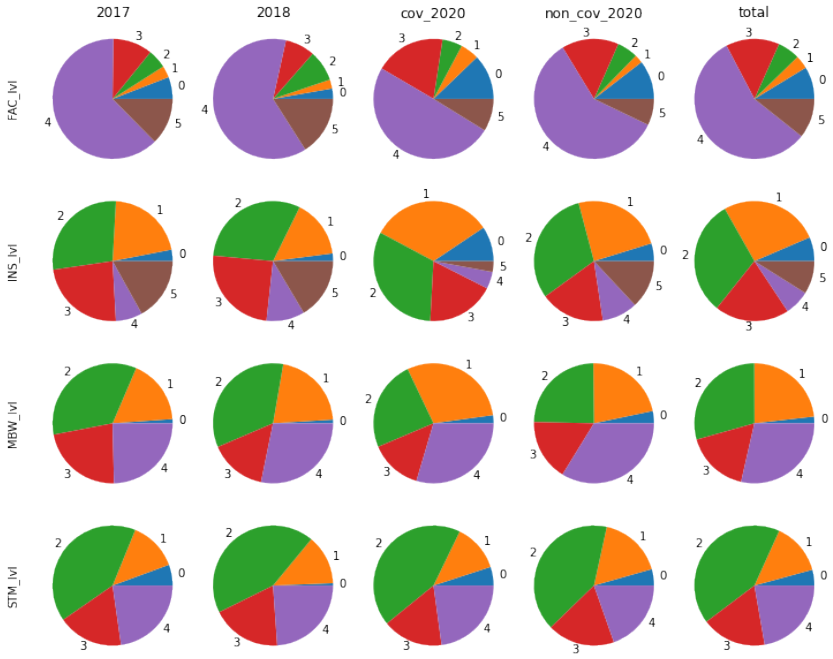

Figure 3.3: Distribution of levels per domain (Part 2)

	kwd	rndm
% disregard notes	10.4	8.4
% sentences with labels	5.2	4.8
% ADM	37.0	37.7
% ATT	2.1	0.8
% BER	1.7	1.8
% ENR	7.7	6.8
% ETN	17.6	20.9
% FAC	8.3	9.4
% INS	8.5	8.0
% MBW	6.0	4.8
% STM	11.2	9.7

Table 3.5: Weeks 14-34: Comparison between randomly-selected and keyword-selected notes

dataset	N Total	N Annotated	N Disregard	% Disregard
2017	3,377	3,048	329	10%
cov_2020	1,687	1,583	104	6%
total	5,064	4,631	433	9%

Table 3.6: Pilot: Number of annotated notes (incl. disregard)

Randomly-selected vs. Keyword-selected Notes

To assess the effect of the keyword-based data selection procedure (explained in Section 3.2), Table 3.5 compares notes that were selected with keywords vs. notes that were selected randomly. Both types of notes contain the same percentage of sentences with labels (about 5%); this means that the keyword-based selection does not actually contribute to obtaining more labeled sentences (i.e. positive examples), which was the initial goal of implementing this method. However, the method seems to somewhat help with obtaining more labels for certain domains, specifically ATT, MBW and STM; the proportion of labels for these three domains is slightly higher in keyword-selected notes than in randomly-selected notes.

3.4 Data from the Pilot Project

As mentioned above, the positive examples from the pilot project are used for training the model for domain classification. Therefore, the descriptive statistics for this data are provided here as well.

As shown in Table 3.6, about 5,000 notes were annotated in the pilot project. Similarly to the current project, 9% of the notes were marked 'disregard'. The non-disregard notes consist of about 313,000 sentences, ∼4,000 of which contain a domain label. The percentage of sentences with labels is lower compared to the current project (1% vs. 5%) because less domains were annotated in the pilot.

Figure 3.4 shows the total number of labeled sentences per domain. Similarly to the current project, the BER domain is under-represented; only 273 sentences were labeled with this domain. There are sufficient examples for the other 3 domains: 1,431 sentences for FAC, 902 sentences for INS and 1,992 sentences for STM.

dataset	N total	N with labels	% with labels
2017	218,863	3,419	2%
cov_2020	94,431	1,016	1%
total	313,294	4,435	1%

Table 3.7: Pilot: Number of sentences: total and with domain labels

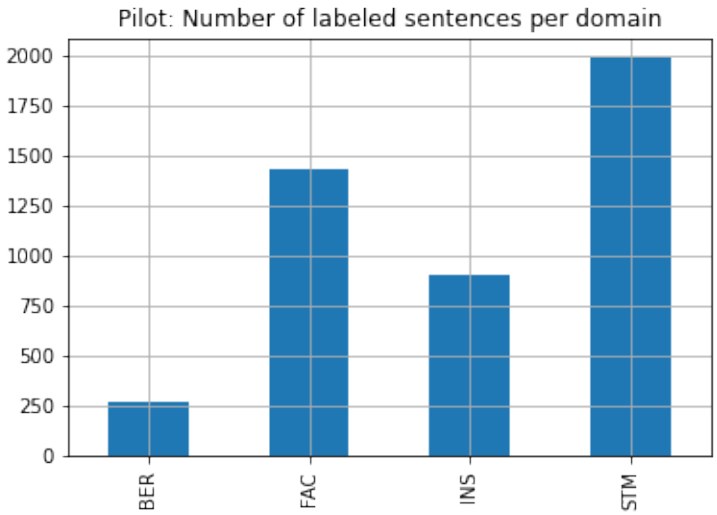

Figure 3.4: Pilot: Number of labeled sentences per domain

The levels labels from the pilot were not used in the current project, since the scales for some of the domains had been adjusted after the pilot and therefore were no longer compatible with each other.

Chapter 4

Final Models

4.1 Introduction

This chapter presents the primary output of the project: a machine learning pipeline that reads a clinical note in Dutch and, based on the textual description, assigns one or more ICF functioning levels to it. The pipeline includes a multi-label classification model that detects the domains mentioned in a sentence, and 9 regression models that assign a level to sentences in which a specific domain was detected. Section 4.2 presents an overview of the pipeline. Section 4.3 provides details about the domains classification model, including evaluation of its performance. Section 4.4 provides details about the levels models, including their performance. Section 4.5 summarizes the main findings discussed in the chapter.

4.2 Machine Learning Pipeline

Figure 4.1 shows an overview of all the steps that are involved in assigning functioning levels to a clinical note. The first step is the anonymization of the note. This is necessary because our classification models are built on top of a pre-trained medical language model (Verkijk 2021); since the language model is anonymized, our text needs to go through the same procedure to be compatible with it. The anonymization is performed with spaCy[1] Named Entity Recognition (NER) model; all words that the NER model identifies

[1] https://spacy.io/

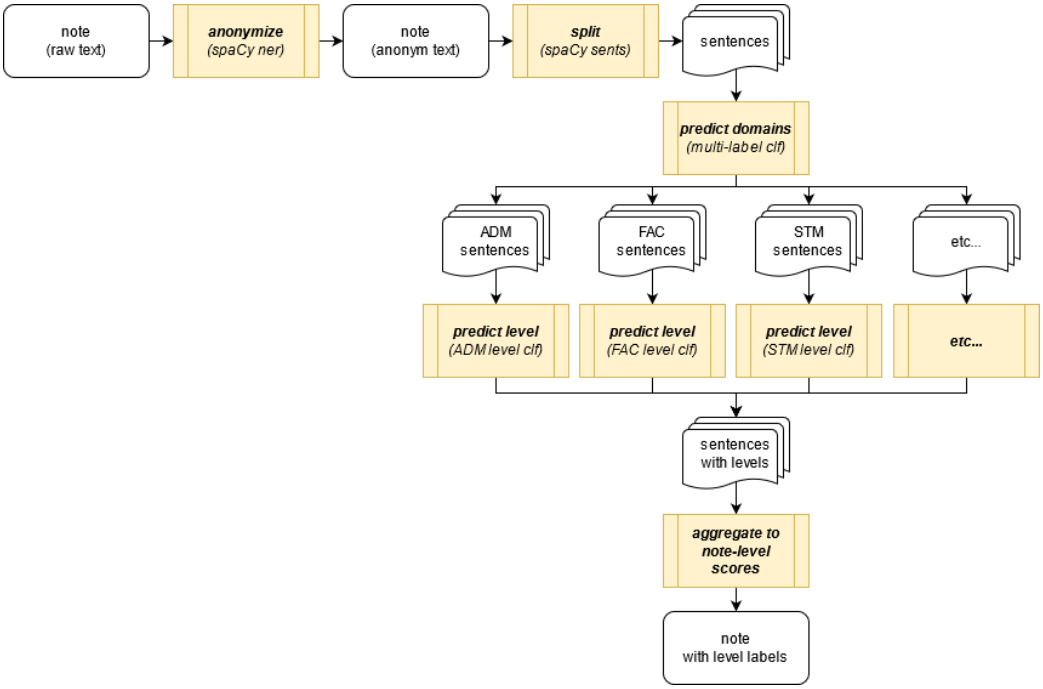

Figure 4.1: Overview of the machine learning pipeline

as entities of type PERSON (i.e. names of people) are replaced with the string 'PERSON', and all words that the NER model identifies as entities of type GPE (i.e. names of cities, countries, etc.) are replaced with the string 'GPE'. See Verkijk (2021) for further information about the anonymization procedure.

After the anonymization, the note is split into sentences, which is also done with spaCy. All sentences are then sent to the domains classification model, which assigns between 0-9 domains to each sentence; for example, the sentence *Loopt, eet, drinkt, geen dyspnoe* (Walks, eats, drinks, no dyspnea) is assigned 3 domains by the classifier: ADM, ETN, FAC.

Next, all sentences that were labeled with a specific domain are sent to the the regression model that assigns them a functioning level with regards to this domain. For example, the abovementioned sentence goes to 3 regression models – ADM level, ETN level, FAC level – and gets a score from each, e.g. ADM level 4.1, ETN level 3.8, FAC level 4.6.

Finally, all the sentences belonging to the same note are aggregated and a note-level score for each domain is calculated. The note-level score is the average of all the sentence-level scores; for example, if a note contains 3 sentences with ADM levels, the ADM level of the note is the average of the ADM levels of the sentences.

4.3 Domains: Multi-label Classification Model

4.3.1 Method

For detection of the domain(s) mentioned in a sentence, a multi-label classification model was trained. A 'multi-label' means that all 9 domains are represented in one label; if a sentence is a positive example for domain X, it will be marked as 1. For example, the label for the abovementioned sentence *Loopt, eet, drinkt, geen dyspnoe* is:

ADM	ATT	BER	ENR	ETN	FAC	INS	MBW	STM
1	0	0	0	1	1	0	0	0

To train the classifier, a state-of-the-art NLP method is applied: fine-tuning of a pre-trained language model. This is implemented with the Python library Simple Transformers[2]. As the pre-trained language model, we use the 'from scratch' medical language model of Verkijk (2021); the reasons for this choice

[2] https://simpletransformers.ai/

	Total number sentences	Total number notes
train	239,153	6,821
dev	21,742	431
test	22,082	431
total	282,977	7,683

Table 4.1: Domain classification: total number of sentences and notes

are discussed in Section 5.2.3 below. The default hyperparameters values of Simple Transformers are applied:

Optimizer: AdamW
Learning rate: 4e-5
Number train epochs: 1
Train batch size: 8

Data

All sentences from the current annotation round (n=286,115) – both the positive examples (i.e. sentences with domain labels) and the negative examples (i.e. sentences without domain labels) – are split into a training set (80% of the sentences), a development set (10% of the sentences), and a test set (10% of the sentences). The development set is used for evaluation during the intermediate experiments described in Chapter 5; the test set is used for the final evaluation described below. After the split, the following additional steps are applied:

- Sentences that are labeled as background/target (and do not contain any domain labels) are removed from the training set (n=7,548). The reason for this is discussed in Section 5.2.1.
- The positive examples from the pilot project are added to the training set (n=4,410). The reason for this is discussed in Section 5.2.2.
- Part of the test set is re-annotated by one of the annotators. The reason for this is that during the initial evaluation with the test set, a lot of inconsistencies and annotation mistakes were observed; specifically, the number of "false positives" that turned out to be not real false positives but annotation mistakes was extremely high. Inconsistencies in the

	ADM	ATT	BER	ENR	ETN	FAC	INS	MBW	STM
train	4,988	247	486	989	2,420	2,489	1,967	755	3,390
dev	411	22	29	105	225	119	127	96	147
test	775	39	54	160	382	253	287	125	181
total	6,174	308	569	1,254	3,027	2,861	2,381	976	3,718

Table 4.2: Domain classification: sentences with labels (positive examples)

	ADM	ATT	BER	ENR	ETN	FAC	INS	MBW	STM
train	2,345	175	381	707	1,416	1,631	1,260	546	1,989
dev	188	17	25	71	128	75	78	71	83
test	231	27	34	92	165	95	116	64	94
total	2,764	219	440	870	1,709	1,801	1,454	681	2,166

Table 4.3: Domain classification: notes with labels (positive examples)

gold labels are inevitable (see Section 2.3 above); however, if they are very dominant in the test set, it prevents us from accurately assessing the performance of the model. Therefore, it was decided to partially re-annotate the test set, with emphasis on accuracy and consistency. An additional special instruction for the re-annotation was to label the domains (but not the levels) in background/target sentences as well (the reason for this is discussed in Section 5.2.1).

After the application of the above steps, the training, development and test datasets consist of a total of 7,683 notes which contain 282,977 sentences; see Table 4.1. Most of these sentences are negative examples, i.e. they do not contain any domain labels. The number of positive examples for each domain is shown in Table 4.2 (sentence-level) and Table 4.3 (note-level).

4.3.2 Results

Precision, Recall, F1-score

When fine-tuning a pre-trained language model, some weights are being randomly initialized; this is a characteristic of all deep-learning methods. Because of this random initialization, two models that are trained on exactly

	ADM	ATT	BER	ENR	ETN	FAC	INS	MBW	STM
precision	1.0	1.0	0.66	0.96	0.95	0.84	0.95	0.87	0.80
recall	0.89	0.56	0.44	0.70	0.72	0.89	0.46	0.87	0.87
F1-score	0.94	0.71	0.50	0.81	0.82	0.86	0.61	0.87	0.84
support	231	27	34	92	165	95	116	64	94

Table 4.4: Domain classification: evaluation on test set, note-level

	ADM	ATT	BER	ENR	ETN	FAC	INS	MBW	STM
precision	0.98	0.98	0.56	0.96	0.92	0.84	0.89	0.79	0.70
recall	0.49	0.41	0.29	0.57	0.49	0.71	0.26	0.62	0.75
F1-score	0.66	0.58	0.35	0.72	0.63	0.76	0.41	0.70	0.72
support	775	39	54	160	382	253	287	125	181

Table 4.5: Domain classification: evaluation on test set, sentence-level

the same data and with exactly the same hyperparameters might still differ from each other in performance. Therefore, the commonly used practice is to train two (or more) models that are identical in everything except for the random initialization, evaluate both on the same test set and average the performance. This procedure is followed here as well; the metrics that are presented and discussed below are an average of the performance of two models.

The performance that is relevant for the healthcare professionals is on note-level. Therefore, although the classification is performed on sentence-level, the aggregated note-level results are also reported. Table 4.4 shows the precision, recall and F1-score on a note-level. The performance for ADM, ENR, ETN, FAC, MBW and STM is good, with an F1-score above 0.8. The domains ATT, BER and INS, on the other hand, perform poorly, especially in terms of their recall. To understand these results, the sentence-level performance is analyzed in detail below.

Table 4.5 shows the metrics on a sentence-level. For the two domains for which there is little data – ATT and BER – the results are very unstable across the two models; the F1-score of ATT is 0.53 in one model and 0.62 in the other, and the F1-score of BER is 0.44 in one model and 0.26 in the other. For comparison, the average difference in F1-score for the other domains is

0.017. This indicates that the models did not manage to consistently learn a pattern for these two domains; most likely because there are not enough training examples.

Another domain that performs poorly is INS. Its low F1-score is due to an extremely low recall (0.26), meaning that the model does not manage to detect instances of this domain (the sentences that the model does label as INS are mostly correct, as indicated by the high precision score). As mentioned in Section 2.3, the annotators have indicated that this domain was difficult to annotate, which is also reflected in a very low inter-annotator agreement (0.34). It is therefore not surprising that the model has difficulty with identifying INS examples: (a) the gold labels for this domain are inconsistent, and (b) the vocabulary that is used to describe it is very diverse, since it includes a wide range of activities (walking, household, groceries, cycling, jogging, football, etc.).

For the 6 domains that do perform well, the high F1-score is mostly due to a very high precision: above 0.9 for ADM, ENR and ETN, 0.84 for FAC, 0.79 for MBW (see Table 4.5); STM is the only one of the six with a lower precision score of 0.7. High precision means that the sentences that the model identifies as domain X are mostly also annotated as such, i.e. there are not many false positives. The recall, on the other hand, is significantly lower: under 0.6 for ADM, ENR and ETN, 0.71 for FAC, 0.62 for MBW; STM is the only one that has a relatively high recall of 0.75. Low recall means that the model does not manage to find many of the sentences that are annotated as domain X; this might suggest that the model is over-fitted to specific phrasings encountered in the training data and has difficulty with generalizing to new, previously unseen phrasings. It should be noted, however, that on the note-level the recall recovers to above 0.7 for all six domains. This can be attributed to two main factors: (a) there is more than one sentence in a note that discusses a specific domains, and the model manages to detect at least one of these sentences, and/or (b) due to incorrect sentence-segmentation (which is not uncommon in clinical notes because of non-standard punctuation), the gold label and the model's label sometimes end up in different segments of the same sentence, which is regarded as an error on a sentence-level but is not problematic on a note-level.

Table 4.6 compares between the F1-score of the inter-annotator-agreement (see Section 2.3) and the F1-score achieved by the model; both on a sentence-level. In general, the performance of the model is very similar to the IAA score; in some cases (especially ETN and STM), the model's F1-score is

	ADM	ATT	BER	ENR	ETN	FAC	INS	MBW	STM
IAA	0.64	0.58	0.42	0.66	0.45	0.78	0.34	0.62	0.57
model	0.66	0.58	0.35	0.72	0.63	0.76	0.41	0.70	0.72

Table 4.6: F1-score: inter-annotator agreement vs. model performance

even quite a bit higher than the IAA.[3] As discussed in Section 2.3, low IAA means that the gold labels on which the model is trained are inconsistent. Therefore, in the best-case scenario we can expect the model to perform on a human-like level, i.e. have an F1-score comparable to the IAA F1-score; this seems to be the case for the domains classifier.

Error Analysis

Figure 4.2 shows a confusion matrix of the model's predictions on the test set; the columns are the gold labels and the rows are the predicted labels.[4] This representation allows to analyze whether the model tends to "confuse" between certain domains. For example, out of the 775 sentences that have a gold ADM label, 368 were correctly assigned this label by the model (first column, first row); 403 out of the 775 sentences were not assigned any label at all (first column, last row); 2 sentences were incorrectly labeled as FAC, 1 sentence was incorrectly labeled as INS, and 1 sentence was incorrectly labeled as STM.[5] This means that the predominant error that the model makes in regards to ADM sentences is to not label them at all; there is almost no confusion with other domains.

The same observation holds for all the other domains as well. Almost all cases of misclassification are into "none" (false negatives); the confusion between the different domains is minimal. The only cases with slightly more examples are: confusion between ETN and MBW (n=5), confusion between INS and FAC (n=5), and confusion between INS and BER (n=5). These examples are shown in Table 4.7, Table 4.8 and Table 4.9.

[3]Note that the number of sentences on which the scores are calculated is different in the two cases, so this is not a direct comparison, but rather a general indication.

[4]The confusion matrix and the examples below are based on the predictions of one of the two models whose metrics were averaged in the previous section.

[5]Note that cases where a sentence is **correctly** labeled with another domain label, e.g. both ADM and FAC labels are correct, are not counted as confusion.

	ADM	ATT	BER	ENR	ETN	FAC	INS	MBW	STM	<none>
ADM	368	0	0	0	0	0	0	0	0	9
ATT	0	14	0	0	0	0	0	0	0	0
BER	0	0	22	1	0	0	4	0	0	24
ENR	0	0	0	90	0	0	0	0	0	2
ETN	0	0	0	0	186	0	0	1	0	15
FAC	2	0	0	1	1	182	5	0	0	41
INS	1	0	1	0	0	0	83	0	0	18
MBW	0	0	0	1	4	0	0	77	0	22
STM	1	0	1	1	1	0	1	0	138	65
<none>	403	25	30	66	190	71	194	47	43	0

Figure 4.2: Domain classification: confusion matrix

Confusion between ETN and MBW

A certain degree of confusion between the ETN and the MBW domains would not be surprising since both are related to nutrition. The difference, as defined in the annotation guidelines, is that ETN focuses on the eating activity, the intake and the need for medical food supplements, whereas MBW focuses on the ability to maintain body weight.

However, there is only one example of a real confusion between these two domains in the test set: the first row in Table 4.7. This sentence discusses a risk of malnutrition and the need for nutritional supplements, ant it is therefore labeled as ETN. There is no explicit mention of problems with maintaining the body weight, even though it is implied that without the supplements the weight cannot be maintained. The model labels this sentence as MBW instead of ETN, probably because of the presence of the word 'weight'.

In all the other "confusion" examples, the algorithm's prediction is actually correct:

- The second sentence in Table 4.7, which is incorrectly segmented into two "sentences", discusses both weight loss and reduced intake; according to the annotation guidelines, it should be labeled as both ETN and MBW. The algorithm correctly assigns the MBW label (to both seg-

ments) but does not assign the ETN label; this means that this is an example of a false negative ETN, rather than an example of confusion.

- The third and fourth sentences in Table 4.7 also discuss both weight loss and intake problems (difficulties with eating), and they should be labeled as both ETN and MBW. The fact that MBW is not in the gold labels is an annotation mistake in these cases; the model's prediction of MBW is correct. Again, these are examples of a false negative ETN, rather than an example of confusion.

- The last sentence in Table 4.7 states that the person has a *slechte voedingstoestand* (poor nutritional status); this term can refer to both problems with intake (ETN) and problems with maintaining body weight (MBW). So in this case, the sentence itself is vague and it is not necessarily an example of confusion.

Confusion between INS and FAC

Confusion between INS and FAC has been mentioned as a problem by the annotators. In the annotation guidelines, some of the levels of exercise tolerance are defined in terms of the type of walking that can be tolerated; for example, being able to walk at a slow or moderate pace is ins2, while being able to walk at a normal pace is ins3. Therefore, sentences that mention walking are labeled as INS, since they provide explicit indication that the patient can tolerate this activity level; however, it was not always clear to the annotators whether these sentences should also be labeled as FAC. Following the weekly discussions with the annotators, it was decided that sentences that mention walking but do not specify the type of walking (independent / with support / stairs etc.) are not to be labeled as FAC.

Therefore, the first 4 examples in Table 4.8, which state that walking can be tolerated by the patient but do not explicitly discuss the level of walking, have the gold label INS but not FAC. The algorithm does not label these sentences as INS, but rather assigns them the FAC label. This is probably because the algorithm has a very strong association between the word *lopen* (walk) and the FAC label. These are examples of real confusion between the two domains.

The last example in Table 4.8 is interesting: the algorithm assigns a FAC label to a sentence that describes that a patient's saturation drops and then "crawls up" again. This sentence has nothing to do with walking, but it

sen id	text (original)	text (translation)	gold label(s)	predicted label(s)
414192335_0023	Er is sprake van risico op ondervoeding, chronische ziekte met inflammatie obv stabiel gewicht bij het gebruik van aanvullende medische voeding.	There is a risk of malnutrition, chronic disease with inflammation based on stable weight when using additional medical nutrition.	ETN	MBW
386958639_0009	Gewichtsverlies	Weight loss	MBW	MBW
386958639_0010	∼5 Kg en algehele zwakte bij verminderde intake 5.	∼ 5kg and general weakness with reduced intake 5.	ETN	MBW
444788701_0041	Patiënt is veel afgevallen maar zijn gewicht stijgt weer nu hij beter eet.	The patient has lost a lot of weight but his weight is increasing again now that he is eating better.	ETN	MBW
259938195_0066	Door de stress vindt hij het wel heel erg moeilijk om te eten waarbij hij ongeveer 2-3 kg is afgevallen.	Due to the stress he finds it very difficult to eat and he has lost about 2-3 kg.	ETN, STM	MBW, STM
154178975_0004	PERSON heeft een stabiele slechte voedingstoestand.	PERSON has a stable poor nutritional status.	MBW	ETN

Table 4.7: Error analysis: confusion between ETN and MBW

sen id	text (original)	text (translation)	gold label(s)	predicted label(s)
419862512_0039	S: geen I: 6 meter lopen.	S: no I: walk 6 meters.	INS	FAC
419862512_0047	Bij lopen saturatie 85% Transfers (bij transfers 2 vpk, arts en FT aanwezig).	When walking saturation 85% Transfers (for transfers 2 nurses, a doctor and a physiotherapist are present).	ADM, INS	FAC
439607230_0042	Probeert af en toe te wandelen, maar is veel binnen.	Occasionally tries to go out for a walk, but stays indoors a lot.	INS	FAC
428039733_0061	Vervoer: lopen, fietsen, autorijden	Transport: walking, cycling, driving	INS	FAC
419862512_0044	Tijdens transfer naar bedrand sat dipt naar 86% en kruipt langzaam weer op.	During transfer to bedside sat[uration] dips to 86% and slowly crawls up again.	ADM, INS	*none*
419862512_0045			ADM, INS	FAC

Table 4.8: Error analysis: confusion between INS and FAC

sen id	text (original)	text (translation)	gold label(s)	predicted label(s)
408933242_0006	Psychosociaal Echtgenoot, 2 kinderen, eigen telefoonwinkel, fiets elke dag 7-8 km naar zijn werk, wel elektrische fiets.	Psychosocial: Spouse, 2 children, owns a telephone shop, cycles 7-8 km to work every day, but on an electric bicycle.	INS	BER
449453883_0043	Zij fietst naar het werk en terug, maar is nadien bek af en heeft snel last van verzuring.	She cycles to work and back, but afterwards is tired and suffers from muscle acidification.	ENR, INS	BER
449453883_0053	Beweegt iedere dag 20-40 minuten en probeert dit ook aan te houden op de dagen dat zij niet naar haar werk gaat.	Moves for 20-40 minutes every day and tries to keep it up on the days she doesn't go to work as well.	INS	BER
213567340_0012	Hij werkt sinds 1 jaar in de beton industrie (GPE), zwaar fysiek werk.	He has been working in the last year in the concrete industry (GPE), heavy physical work.	INS	BER
399996679_0072	In het dagelijks leven gaat het redelijk, kan een trap oplopen, helpt haar dochter in de groothandel naar kunnen.	In the daily life things are going reasonably well, she can climb stairs, helps her daughter in the wholesale when she can.	BER, FAC	FAC, INS

Table 4.9: Error analysis: confusion between INS and BER

metaphorically uses a motion-related verb (crawl) to describe changes in the saturation levels.

Confusion between INS and BER

Conceptually, INS and BER are not similar to each other; rather, the connection between them is that exercise tolerance can be expressed through the ability to work or to travel to work. The fact that someone can cycle to work, like in the first three examples in Table 4.9, explicitly says something about their level of exercise tolerance and implicitly says that they are able to work. Since the ability to work is not discussed explicitly (e.g. we don't know if they work at the same capacity like before the sickness, or only part-time), the annotator did not label it as BER; the algorithm, however, recognizes the work-related words and marks the sentences as BER. With regards to INS, the algorithm's predictions for these examples are false negatives (i.e. it fails to detect that the sentences should be labeled INS).

The fourth sentence in Table 4.9 is different since it does discuss work ability explicitly; here, the fact that the algorithm assigns a BER label seems to be correct (in terms of INS, it is a false negative). The last sentence in Table 4.9 describes a walking ability (can climb stairs) and a work ability (can help at the store), therefore it was annotated as BER and FAC. The algorithm correctly predicts the FAC label, fails to assign the BER label, but does assign an additional INS label; this is probably related to climbing the stairs, which can be arguably viewed as an indication for the level of exercise tolerance. Therefore, this example is not necessarily incorrect in relation to the predicted INS label.

False negatives

As evident from the confusion matrix in Figure 4.2 (last row), as well as from the low recall scores discussed above, the main problem of the model is that it does not manage to detect many of the sentences that belong to a certain domain; in other words, there are a lot of false negatives. In fact, for 5 out of the 9 domains, there are more false negatives than true positives. Table 4.10 shows a randomly sampled example of a false negative for each domain.

The ADM example (first row) is another case of incorrect segmentation; the relevant string is *Ademhal. 16* (respiratory rate 16), and it is erroneously cut into two segments. While the annotator marks the phrase across the segment border (i.e. both segments are labeled ADM), the algorithm marks only the first segment. This type of "false negative" is not problematic, since

sen id	text (original)	text (translation)	gold label(s)	predicted label(s)
408669301_0024	36,6 °C | Ademhal.	36.6 °C | Respiratory rate	ADM	ADM
408669301_0025	16 | SpO2	16 | SpO2	ADM	*none*
408669301_0026	96%	96%	*none*	*none*
425960851_0051	Herstel verloopt Wel voorspoedig Huidige problemen: - concentratieproblemen, snel overprikkeld - conditie aan het opbouwen mbv fysio in reade - wens psychologische ondersteuning, momenteel in Reade met psycholoog contact Beleid: Controle poli afspraak 3 maanden vanaf ontslagdatum.	Recovery is going well. Current problems: - Concentration problems, easily overstimulated - Building up fitness with a physiotherapist in Reade - Would like psychological support, currently in contact with a psychologist in Reade. Policy: Outpatient checkup appointment 3 months from discharge date.	ATT	*none*
445965542_0004	Na 2 weken hervatten activiteiten en (thuis) werken.	After 2 weeks resumed activities and work (from home).	BER	*none*
429987181_0004	Moe/slap.	Tired/weak.	ENR	*none*
417019856_0121	: ja, Heeft u hulp nodig bij het eten?:	: yes, Needs help with eating?:	ETN	*none*
417019856_0122	nee, KATZ score:	no, KATZ score:	ETN	*none*
414190442_0003	Nog zeer zwak op de benen.	Still very weak in the legs.	FAC	*none*
419862512_0017	Hele dag op de been, vooral veegwerk.	On his feet the whole day, mostly sweeping.	INS	*none*
383074874_0028	Gewicht: in 4 weken van 78 naar 74 kg.	Weight: in 4 weeks from 78 to 74 kg.	MBW	*none*
408924107_0038	Emotineel laag belastbaar en onzeker tijdens mobiliseren.	Emotionally hypersensitive and insecure when moving around.	STM	*none*

Table 4.10: Error analysis: random sample of false negatives

sen id	text (original)	text (translation)	gold label(s)	predicted label(s)
405497458_0042	Op de afdeling sterkte hij langzaam aan en kon de zuurstof langzaam afgebouwd worden waarna hij in de avond van 1 april met ontslag kon naar het Hof van Sloten voor verdere revalidatie en verder afbouwen van zuurstof.	In the ward he gradually strengthened and the oxygen could be gradually reduced, after which he was discharged to the Hof van Sloten in the evening of 1 April for further rehabilitation and further reduction of oxygen.	*none*	ADM
403593331_0034	Sociale anamnese Getrouwd, werkt in de huishouding, kinderen.	Social anamnesis: Married, works in the household, children.	*none*	BER
395536606_0035	Heeft het idee dat het vaker optreedt bij minder slapen of moe zijn.	Thinks that it occurs more often with less sleep or being tired.	*none*	ENR
265083319_0017	Algemeen Laat de sondevoeding op kamertemperatuur inlopen.	General: Let the feeding tube flow at room temperature.	*none*	ETN
176571330_0024	Belde regelmatig en wilde dan weten wanneer het bloed in was gelopen.	Called regularly and wanted to know when the blood was administered.	*none*	FAC
396143675_0032	Wijt dit zelf aan slechte conditie.	Attributes it to poor fitness.	*none*	INS
201946063_0036	mevrouw heeft een verminderde voedingstoestand, ziekte gerelateerde ondervoeding zonder inflammatie	Mrs. has a reduced nutritional status, disease-related malnutrition without inflammation	ETN	ETN, MBW
444008207_0025	Bij psychiatrisch onderzoek is er sprake van een geagiteerd angstig-depressief toestandsbeeld met suicidaliteit, geen psychose.	Psychiatric examination reveals an agitated anxious-depressive state with suicidality, no psychosis.	*none*	STM

Table 4.11: Error analysis: random sample of false positives

it is correct on the note-level, which is the unit of interest for the healthcare professionals.

The rest of the examples in Figure 4.2 are real false negatives. Some of them might be considered more "difficult" because they contain vague or less standard phrasings; for example *zwak op de benen* (weak in the legs) is a somewhat idiomatic expression which is used to describe someone who cannot stand or walk (FAC). It is not a standard way to describe a walking function (at least not as standard as using the verb *lopen*) and therefore it is understandable that this example is less clear for the model.

However, most of the other examples are very standard: the ATT example contains the word *concentratieproblemen* (concentration problems), the BER example contains the verb *werken* (work), the ENR example contains the word *moe* (tired), the ETN example contains the verb *eten* (eat), the MBW example contains both *gewicht* (weight) and *kg*. Based on the presence of these standard vocabulary items, which in other sentences are correctly detected by the algorithm, it is not clear why these sentences were not correctly detected.

False positives

As evident from the confusion matrix in Figure 4.2 (last column), as well as from the high precision scores discussed above, false positives are not a major issue for the model. The only two domains for which the precision is lower than the recall are BER and STM.

Table 4.11 shows a randomly sampled example of a false positive for each domain (except for ATT, for which there were no false positives). The sentence in the first row seems to be not an actual false positive, but rather an annotation mistake: it discusses the need for oxygen, which should be labeled as ADM, so the predicted label is correct. The STM example in the last row is somewhat borderline between a diagnosis (which should not be annotated, according to the guidelines) and a description of emotional functioning (which should be annotated); the algorithm probably assigns the STM label based on the words *angstig* (anxious) and *suicidaliteit* (suicidality), which in other contexts can be related to emotional functioning.

The rest of the examples are actual false positives. Most of them contain words that are strongly associated with a specific domain; it is therefore clear why the algorithm assigned the label that it did. Specifically, the false BER positive contains the verb *werken* (work), the false ENR positive contains the word *moe* (tired), the false ETN positive contains the word *sondevoeding*

(feeding tube), the false FAC positive contains the verb *lopen* (walk; here in a different meaning), the false INS positive contains the word *conditie* (fitness). The MBW fasle positive is less clear, since it does not contain any weight-related words.

4.4 Levels: Regression Models

4.4.1 Method

For the task of assigning a functioning level to a sentence, a regression model for each domain was trained. A regression model outputs a continuous output, i.e. a decimal number; this decimal can also be outside the boundaries of the original scale (e.g. 4.1 for ADM).

The models are trained with the same method as the domain classification model: fine-tuning the 'from scratch' medical language model of Verkijk (2021). This is implemented with the Python library Simple Transformers; the default hyperparameters values are applied:

Optimizer: AdamW
Learning rate: 4e-5
Number train epochs: 1
Train batch size: 8

Data

The division into a training set, a development set and a test set is the same as for the domain classification model (Section 4.3.1). However, for the regression models, only the sentences that have a gold level label (for a specific domain) are used. For example, to train the ADM level model, we select all the sentences in the training set that have an ADM level label (adm0, adm1, adm2, etc.). The number of sentences used for training and evaluating each of the 9 models is shown in Table 4.12. For ADM there are over 5,000 training sentences available, but for the other domains the numbers are much lower: for most domains there are between 1,000 - 2,500 training sentences, for MBW there are about 750 training sentences, and for ATT and BER there are only about 200 training sentences for each. Table 4.13 shows the datasets aggregated to a note-level.

	ADM	ATT	BER	ENR	ETN	FAC	INS	MBW	STM
train	5,233	251	216	1,005	2,491	1,086	1,104	766	1,420
dev	440	23	29	107	236	124	132	98	148
test	421	32	26	100	183	139	136	60	155

Table 4.12: Levels classification: datasets, sentence-level

	ADM	ATT	BER	ENR	ETN	FAC	INS	MBW	STM
train	2,344	175	162	705	1,417	717	699	536	792
dev	189	17	25	71	128	74	77	71	83
test	200	21	22	70	123	79	74	41	84

Table 4.13: Levels classification: datasets, note-level

4.4.2 Results

The predictions of the regression models are evaluated on the test set against the gold labels using three standard metrics:

- Mean absolute error (MAE). This is the most intuitively clear metric; it is the average of all the absolute values of the errors. For example, if the first sentence in the ADM test set has the gold label 4 and the model predicted 3.6, the absolute error for this sentence is 0.4. After all the errors in the test set are calculated in this way, the average of all the errors is the MAE.

- Mean squared error (MSE). This metric is the average of all the squared values of the errors. For example, when the absolute error is 0.4 (like in the above example), the squared error is 0.4^2=0.16; the average of all the squared errors in the test set is the MSE. The disadvantage of this metric compared to MAE is that the MSE is sensitive to outliers, i.e. it gives them a big weight.

- Root mean squared error (RMSE). This metric is the root of the MSE; for example, if the MSE is 0.55, the RMSE is $\sqrt[2]{0}$.55=0.74. Similarly to the MSE, the RMSE is sensitive to outliers.

For all these metrics, a lower value is better; when there is no difference between the gold labels and the predictions the value of the metrics is 0.

	ADM	ATT	BER	ENR	ETN	FAC*	INS*	MBW	STM
MAE	0.37	1.03	1.49	0.43	0.50	0.66	0.61	0.60	0.68
MSE	0.34	1.47	2.85	0.42	0.47	0.93	0.64	0.56	0.87
RMSE	0.58	1.21	1.69	0.65	0.68	0.96	0.80	0.75	0.93
support	200	21	22	70	123	79	74	41	84

Table 4.14: Levels classification: evaluation results, note-level

	ADM	ATT	BER	ENR	ETN	FAC*	INS*	MBW	STM
MAE	0.48	0.99	1.56	0.48	0.59	0.70	0.69	0.81	0.76
MSE	0.55	1.35	3.06	0.49	0.65	0.91	0.80	0.83	1.03
RMSE	0.74	1.16	1.75	0.70	0.81	0.95	0.89	0.91	1.01
support	421	32	26	100	183	139	136	60	155

Table 4.15: Levels classification: evaluation results, sentence-level

Table 4.14 shows the evaluation results on a note-level, which is the meaningful unit for the healthcare professionals. For each note, the note-level gold label is calculated by averaging all the sentence-level gold labels in the note; e.g., if there are 3 ADM sentences in the note and their levels are adm1, adm2 and adm3, then the note-level label is adm2. Similarly, the note-level predicted label is calculated by averaging all the sentence-level predicted labels in the note. The metrics are then calculated by comparing the gold labels and the predicted labels, as explained above. For simplicity, only the MAE metric is discussed in detail.

For all domains except for ATT and BER, the MAE is below 0.7. This means that, on average, the predicted functioning level for a note does not deviate more than 0.7 points from the annotated functioning level. In the annotation guidelines, the functioning levels are defined in intervals of 1; a MAE of less than 1 can be therefore considered good, since it means that the prediction is likely to be inside the range of the correct level. For example, if the human-annotated functioning level is 1 and the model's prediction is 1.7, this prediction is still lower than the next defined level (which is 2). For the ADM and ENR domains, the MAE is especially low (less than 0.5). For the ATT and BER domains, on the other hand, the MAE is above 1. This lower performance is expected based on the low number of training examples

	ADM	ATT	BER	ENR	ETN	FAC	INS	MBW	STM
IAA	0.25	0.32	0.38	0.39	0.28	0.17	0.30	0.32	0.31
model	0.48	0.99	1.56	0.48	0.59	0.70	0.69	0.81	0.76

Table 4.16: MAE: inter-annotator agreement vs. model performance

available for these domains.

Table 4.15 shows the metrics on a sentence-level, which is the model's original classification unit. The MAE values are higher compared to the note-level MAE, but the trends are similar. The ATT and BER domains show the lowest performance, with MAE's of around 1-1.5; the rest of the domains have MAE's under 0.8 (with the exception of MBW, which stands on 0.81). Even though this can be viewed as a good performance, as discussed above, it should be noted that the model's performance is significantly lower than the human performance on this task. As shown in Table 4.16, the IAA MAE is lower than 0.4 for all domains, while the model's MAE is usually at least twice as high. This suggests that this task could benefit from more training data; since the gold labels are quite consistent and reliable (based on the good IAA scores), the models could probably reach more human-like results if they had seen more examples. Specifically, it would probably be useful to collect enough training examples for all possible levels; as mentioned in Section 3.3, for some domains a specific level is very dominant in the current dataset (e.g. for ENR level 2 is very dominant) and this probably affects the performance of the models on under-represented levels.

4.5 Conclusion

This chapter described the final output of (this phase of) the project: a machine learning pipeline that reads a clinical note in Dutch and assigns one or more ICF functioning levels to it. The pipeline includes a domain classification model, which detects the domain(s) mentioned in a sentence, and 9 regression models, which assign a functioning level to sentences in which a specific domain was detected.

Although the models generate predictions on a sentence-level, the unit of interest for the healthcare professionals is the note. Therefore, the classifiers are evaluated both on a sentence-level and on an aggregated note-level.

For human annotators, the domain classification task is the more challenging part of the pipeline, as evident from the relatively low IAA scores discussed in Section 2.3. The performance of the multi-label classification model that was trained for this task is comparable to the human performance, in terms of the F1-score. The main problem of the model is low recall (on a sentence-level), which is observed for all 9 domains. This means that the classifier does not manage to detect all the relevant sentences, i.e. there are many false negatives.

However, for 6 out of the 9 domains, the recall on a note-level is quite a bit higher (above 0.7); together with a high precision (above 0.8), this results in F1-scores above 0.8. The only domains that have an F1-score below 0.8 on a note-level are ATT, BER and INS. For ATT and BER the low performance is likely related to the fact that there are not enough examples for these domains in the dataset. For INS, one reason behind the low performance is probably that the gold labels are inconsistent, as suggested by the very low IAA for this domain. Another possible reason is that the vocabulary used to discuss this domain is very heterogeneous, since it includes a wide range of activities (walking, household, work, groceries, cycling, football, etc.).

The task of assigning a functioning level to a sentence in which a specific domain was detected is the less confusing part for the human annotators, as discussed in Section 2.3. The 9 regression models that were trained for this task do not manage to achieve human-like performance, in terms of the MAE. However, their performance is quite good: for 7 out of the 9 domains, the MAE is below 0.8 on a sentence-level and below 0.7 on a note-level. The performance can be probably improved by adding more training data, especially examples for levels that are under-represented in the current dataset.

To conclude, the pipeline as a whole performs in-line with the pre-defined goals for 6 out of the 9 domains. For ATT and BER, more training data is needed to reach the desired performance levels; for INS, the annotation guidelines need to be reconsidered, since the current definitions are not clear enough to generate consistent gold labels.

Chapter 5

Intermediate Experiments

5.1 Introduction

The setup for the final models described in Chapter 4 was decided based on a round of experiments. The goal of these experiments was to determine (a) the optimal configuration of the training data, (b) the optimal pre-trained language model to fine-tune, (c) the optimal classification unit, and (d) whether it is necessary to train an additional classifier that would filter out background/target sentences from the notes.

The experiments were performed on the data from annotation weeks 14-26; the annotated notes were split into a training set (80%), a development set (10%) and a test set (10%). All experiments were evaluated on the development set; based on this evaluation, the setup for the final models was decided. Additional data from annotation weeks 31-34 was then added to the training set, and the final models described in Chapter 4 were trained.

This chapter shortly describes the experiments and their results. Section 5.2 focuses on the experiments done on the domains classification model; Section 5.3 focuses on the experiments performed on the levels regression models.

5.2 Experiments with the Domains Model

Figure 5.1 shows an overview of the experiments performed on the domain classifier. Four different models were trained:

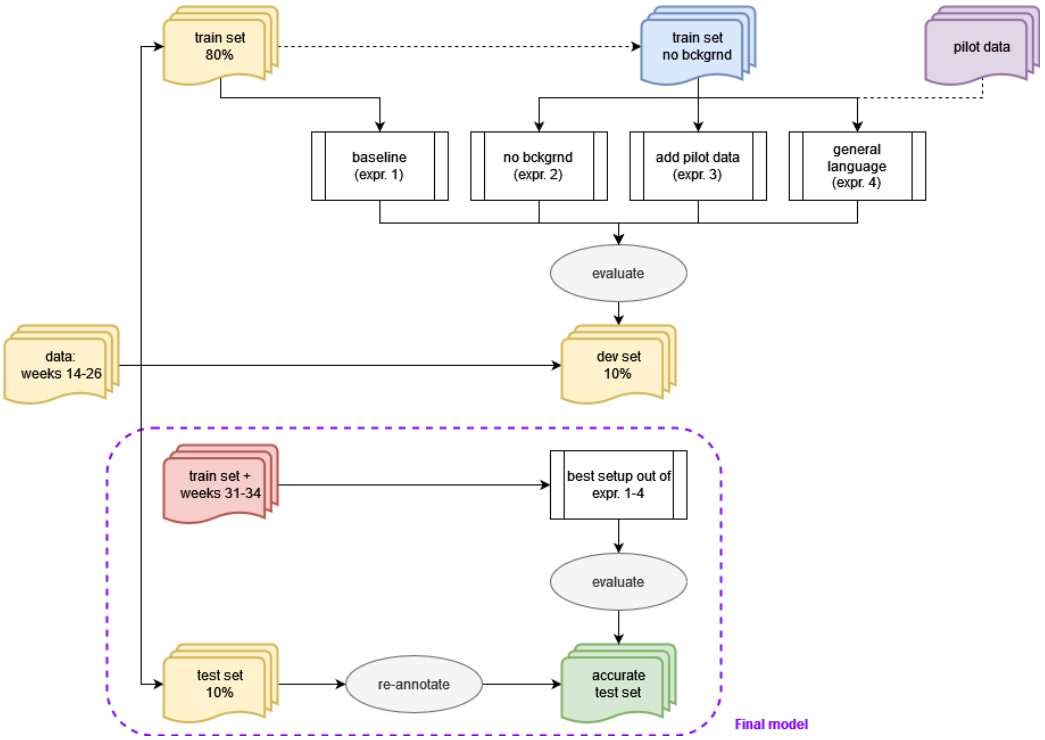

Figure 5.1: Domain classification: overview of experiments

- The 'baseline' model (expr.1 in the figure) is a fine-tuned 'from scratch' medical language model of Verkijk (2021), trained on all the sentences of the training set.
- The 'no bckgrnd' model (expr.2) has the same setup as the 'baseline' (a fine-tuned 'from scratch' medical language model), but all the background/target sentences are removed from the training set.
- The 'add pilot' model (expr.3) has the same setup as the 'no bckgrnd', but positive examples from the pilot project are added to the training set.
- The 'general language' model (expr.4) is a fine-tuned general language model: RobBERT (Delobelle et al. 2020).

The models in expr.1 - expr.3 were trained with the default hyperparameters:

Optimizer: AdamW
Learning rate: 4e-5
Number train epochs: 1
Train batch size: 8

The model in expr.4 was trained with different hyperparameters, as discussed in Section 5.2.3.

The four models were evaluated on the development set; the results of the evaluation are discussed in the below sections. The performance on the ATT and BER domains is not considered when the results are compared and discussed; it varies greatly from model to model because of the small number of examples (both in the training set and in the development set), so the differences are not really informative.

5.2.1 Exclude Background/Target

Some sentences in clinical notes are relevant to one of the 9 ICF domains but do not discuss the current level of functioning. For example, it can be stated that the patient needed tube feeding a week ago (background information), or it can be stated that the patient should be able to walk independently in the next 4 weeks (setting a target for the treatment). In such cases, the annotators were instructed to mark the sentence as 'background' or 'target', and not to label it with either domain or level labels.

	ADM	ATT	BER	ENR	ETN	FAC	INS	MBW	STM
precision	0.74	0.69	1.0	0.74	0.58	0.54	0.60	0.73	0.63
recall	0.65	0.50	0.1	0.83	0.45	0.50	0.21	0.75	0.63
F1-score	0.70	0.58	0.19	0.78	0.51	0.52	0.31	0.74	0.63
support	411	22	29	105	225	119	127	96	147

Table 5.1: Baseline: evaluation on development set, sentence-level

	ADM	ATT	BER	ENR	ETN	FAC	INS	MBW	STM
precision	0.72	0.69	0	0.71	0.53	0.50	0.51	0.73	0.67
recall	0.64	0.41	0	0.84	0.61	0.61	0.24	0.73	0.56
F1-score	0.68	0.51	0	0.77	0.57	0.55	0.33	0.73	0.61
support	411	22	29	105	225	119	127	96	147

Table 5.2: Exclude background: evaluation on development set, sentence-level

When the algorithm is learning from the labeled examples in the training set, these background/target sentences are seen as negative examples, i.e. sentences that do not discuss any relevant domain, because they don't have domain labels. This might have a detrimental effect on the learning; if the algorithm encounters sentences about a feeding tube (*sondevoeding*) that are labeled ETN but also sentences about a feeding tube that are not labeled as ETN, this inconsistent input makes it harder to learn the correct pattern. One possible way to address this issue is to remove the background/target sentences from the training set. The experiment reported in this section examines the effect that this has on the performance of the classifier.

Table 5.1 shows the performance on the development set (sentence-level) of a model trained on all the sentences in the training set. Table 5.2 shows the performance when background/target sentences are excluded from the training. It should be noted that some differences in performance are always expected; as mentioned in Chapter 4, even models that are trained on exactly the same data and with the same settings differ from each other because of the random initialization of deep learning algorithms.

With this caveat in mind, we can compare the F1-score of the 'baseline' model (Table 5.1) to the the F1-score of the 'no bckgrnd' model (Table 5.2): 'no bckgrnd' performs slightly worse on ADM (-0.02 points), ENR (-0.01

point), MBW (-0.01 point) and STM (-0.02 points); however, it performs quite a lot better on ETN (+0.06 points) and slightly better on FAC (+0.03 ponts) and INS (+0.02 points).

The domains on which the performance of 'no bckgrnd' is better are the ones that overall have the lowest performance, i.e. excluding the background/target sentences helps the most "problematic" domains. Most significantly, it helps to improve recall on the ETN domain (from 0.45 to 0.61) and recall on the FAC domain (from 0.5 to 0.61). This suggests that background/target sentences related to these domains (like the feeding tube and the walking examples given above) were indeed "confusing" for the classifier. Therefore, it was decided to adopt the 'no bckgrnd' setup, i.e. to exclude background/target sentences from the training set. Although this results in somewhat lower performance on the "stronger" domains, it seems more important to give an extra push to the "weaker" domains that benefit from this adjustment.

Afternote: In hindsight, the annotation of the background/target sentences should have been done differently. It would have been better if the domain and level labels had been assigned in such sentences as well, in addition to the 'background' / 'target' label. This way, these sentences could have been used as positive examples for the classifiers, which is reasonable since they do in fact discuss a relevant domain and a level of functioning. The only part of the pipeline where these sentences might create an issue is when assigning a functioning level to a whole note; this is discussed in detail in Section 5.3.2 below.

5.2.2 Add Pilot Data

The "pilot" phase of the project focused on 4 ICF domains: BER, FAC, INS and STM. The annotation round in this phase resulted in about 4,000 sentences with labels (see Section 3.4 for details). In this experiment, we examine whether it is beneficial to add these positive examples from the pilot phase to the training set. On the one hand, this would provide the classifier with more examples for the 4 domains in question, which is expected to have a positive effect. On the other hand, this might have a detrimental effect on the other domains, which were not labeled in the pilot. For example, the sentence *Loopt, eet, drinkt, geen dyspnoe* (Walks, eats, drinks, no dyspnea) is relevant for 3 domains in the current scheme: ADM, ETN, FAC. However, in the pilot it is only assigned the label FAC (since ADM and ETN were not

	ADM	ATT	BER	ENR	ETN	FAC	INS	MBW	STM
precision	0.70	0.92	0.44	0.73	0.55	0.43	0.38	0.74	0.47
recall	0.64	0.50	0.66	0.68	0.67	0.75	0.46	0.76	0.80
F1-score	0.67	0.65	0.53	0.70	0.60	0.55	0.42	0.75	0.60
support	411	22	29	105	225	119	127	96	147

Table 5.3: Add pilot: evaluation on development set, sentence-level

annotated in the pilot). Adding this sentence to the training data, means that it would be incorrectly viewed as a negative example for ADM and ETN, which might be "confusing" for the the classifier.

Table 5.3 shows the performance on the development set (sentence-level) of a model trained on data that includes the pilot sentences; this should be compared to the 'no bckgrnd' model in Table 5.2. The effect on the "pilot" domains (excluding BER, as mentioned above) is that their recall increases significantly: +0.14 points for FAC, +0.22 points for INS, +0.24 points for STM. At the same time, their precision drops: -0.07 points for FAC, -0.13 points for INS, -0.20 points for STM. This means that the 'add pilot' model assigns these domains to more sentences, in some cases correctly (increased recall), and in other cases incorrectly (decreased precision). For all three domains, the beneficial effect on the recall is bigger than the detrimental effect on the precision.

The effect on the domains that were not in the pilot is minimal, with one notable exception: for ENR there is a significant drop in the recall, from 0.84 to 0.68. This is probably related to the fact that ENR is often mentioned in sentences that also discuss INS and ADM; for example *Beperkende factor: Vermoeidheid, verminderde inspanningstolerantie en kortademigheid* (Limiting factor: Fatigue, decreased exercise tolerance and shortness of breath), or *Weinig conditie, sporten was nog te veel, kortademigheid en vermoeidheid bij ADL* (Reduced fitness, exercise is still too much, shortness of breath and fatigue when performing activities of daily living (ADL)). In the pilot data, such sentences are annotated as INS only and this interferes with the model's learning, since it now encounters many examples where words like *vermoeidheid* (fatigue) are not labeled as ENR.[1]

[1]This is relevant for ADM as well, but because there are many ADM examples in the training set (five times more than ENR examples), the additional "confusing" examples

	ADM	ATT	BER	ENR	ETN	FAC	INS	MBW	STM
precision	0.67	0	0	0.69	0	0.49	0	0.61	0.52
recall	0.63	0	0	0.75	0	0.42	0	0.43	0.52
F1-score	0.65	0	0	0.72	0	0.45	0	0.50	0.52
support	411	22	29	105	225	119	127	96	147

Table 5.4: Fine-tuned RobBERT: evaluation on development set, sentence-level

Based on these results, it was decided that the benefits of adding the pilot data are bigger than the disadvantages. Since the recall is generally more problematic for the model than the precision, the significant boost in recall that is observed for 3 domains is very desirable, even at the cost of a reduced recall for one other domain.

5.2.3 General Language Model

The three models discussed above are created by fine-tuning the 'from scratch' medical language model of Verkijk (2021); this language model is pre-trained on clinical notes from the Amsterdam UMC. In the current experiment, we check how using a different pre-trained language model affects the performance. Specifically, we fine-tune a general (i.e. not medical domain) Dutch language model called RobBERT (Delobelle et al. 2020). This model was pre-trained using the same training architecture as the medical language model (RoBERTa architecture); the difference is that it learned its language representation from a very large corpus of Dutch text obtained by web crawling.

RobBERT has shown state-of-the-art performance on various NLP tasks (Delobelle et al. 2020), which indicates that it is a good representation of the Dutch language. Verkijk (2021) shows that when comparing the performance of RobBERT and of the medical language model on a general NLP task, like named entity recognition, RobBERT performs significantly better (F1-score 0.84 vs. 0.66). The question is whether it also performs better on a task that involves a very domain-specific language, like our domain classification task.

Table 5.4 shows the performance on the development set (sentence-level) of a fine-tuned RobBERT model trained on the 'no bckgrnd' training set.

do not have a big effect on the learning.

This model was trained with the following hyperparameters:[2]

Optimizer: AdamW
Learning rate: 3e-5
Number train epochs: 2
Train batch size: 8

The results in Table 5.4 should be compared to the 'no bckgrnd' model in Table 5.2. The RobBERT-based classifier performs significantly worse than the medical language classifier. For ETN and INS, RobBERT does not manage to detect any sentences at all; for FAC and STM the F1-score of RobBERT is about 0.1 points lower, and for MBW it is about 0.2 points lower. The only two domains for which the performance of the two models is comparable are ADM and ENR.

These results clearly show the importance of domain-specific language in the pre-trained language model. Even though RobBERT is pre-trained on a much larger dataset, it cannot compete with a smaller, but domain-specific, language representation when it comes to domain-specific downstream tasks.

5.2.4 Conclusion

Based on the evaluation results presented in the above sections, the following setup was chosen for training the final domain classification model:

- Pre-trained language model: the 'from scratch' medical language model of Verkijk (2021).
- Background/target sentences: excluded from the training set.
- Positive examples from the pilot: included in the the training set.

5.3 Experiments with the Levels Models

Figure 5.2 shows an overview of the experiments performed on the levels classifier:

[2]We also trained a RobBERT model with the default hyperparameters that were used for training all the other models; however, this model did not predict any labels for any sentence in the development set.

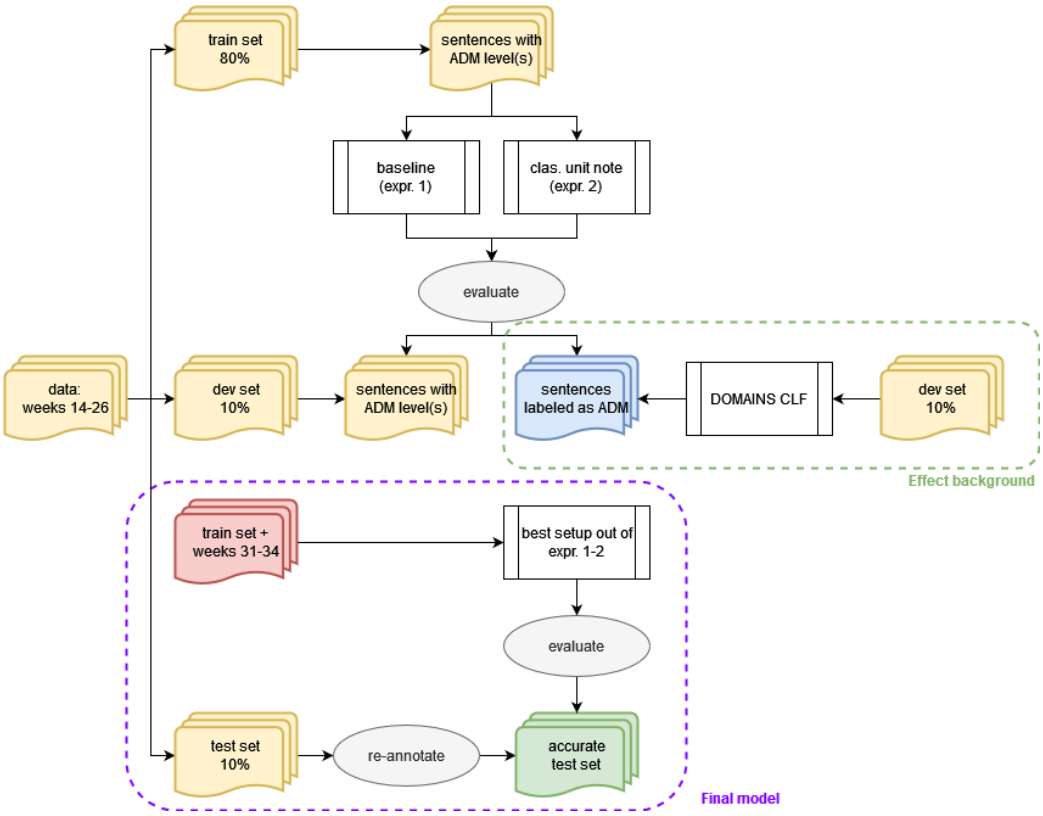

Figure 5.2: Levels classification: overview of experiments

	sent-level clf	note-level clf
MAE	0.39	2.53
MSE	0.28	8.22
RMSE	0.53	2.87

Table 5.5: Classification unit: evaluation on development set, note-level

- The first experiment compared an ADM levels model trained on a sentence-level (expr.1) to an ADM levels model trained on a note-level (expr.2).
- The second experiment (the green box in the figure) assessed the effect of the background/target sentences on the final note-level score by performing two types of evaluation: evaluation on gold labels vs. evaluation on the outputs of the domains classifier.

The two experiments are explained and discussed in the below sections.

5.3.1 Classification Unit

For the healthcare professionals, the meaningful unit for determining the level of functioning at a specific point in time is the note. This experiment checks whether it makes sense to train a model that "reads" a whole note and assigns a functioning score on a note-level. This is compared to an alternative approach: training a model that assigns a functioning score on a sentence-level and aggregating the scores to a note-level by averaging all the sentence scores that belong to the same note.

The experiment was performed on the ADM domain, for which we have the largest number of examples. The training set contains 1,655 notes in which an ADM functioning level is discussed. The note-level model was trained on the full text of each note[3] presented together with an aggregated gold note-level ADM score. For the sentence-level model, only the sentences that have a gold sentence-level ADM score were used for training (3,760 sentences). The two models were then evaluated on 189 notes with ADM level from the development set.

[3]Since some notes are longer than the maximal sequence length accepted by the algorithm, a sliding window method was applied.

Table 5.5 shows the performance of the two models on a note-level. When the classification is done on a sentence-level, the performance is significantly better; the mean absolute error (MAE) for a note score is 0.39 for the sentence-level model and 2.53 for the note-level model.

These results can probably be explained by the fact that a full note contains a lot more information than what is relevant for assigning a functioning level for a specific domain. This makes it difficult to identify the patterns that are relevant for the level score.

5.3.2 Effect of Background/Target

In the final machine learning pipeline, all sentences that are labeled with a specific domain label, e.g. ADM, are sent further to the ADM level model and are assigned a functioning level. Background/target sentences that discuss ADM are likely to get the ADM label as well, since they are likely to have a similar vocabulary to "regular" ADM sentences.

A potential issue that this might create is that the note-level functioning score is skewed by the presence of sentences that do not discuss the current level of functioning. For example, if there were two ADM sentences in a note, one stating that the patient needed mechanical ventilation two months ago (adm0) and the other stating that there are currently no respiratory problems (adm4), the note-level score would be the mean of the two sentences, i.e. adm2. This score does not reflect the current functioning level of the patient, which is adm4.

One possible solution to this issue is building an additional classifier that would filter out the background/target sentences from the note, so that they do not get to the level-assigning models. However, it first needs to be determined whether the problem is actually big enough to justify this step. This is what the current experiment does. To assess the extent of the problem, we perform two evaluations:

- Evaluation on gold labels: only sentences that have a gold label are sent through to the levels-assigning models. This means that in the above example, only the sentence that discusses the current functioning level will be put through the ADM levels model (background sentences are not annotated with domain or level labels). The note-level gold score in this case is adm4, and the note-level predicted score would be the score that the model assigns to the sentence (let's say that it is 3.6).

	ADM	ATT	BER	ENR	ETN	FAC*	INS*	MBW	STM
MAE	0.39	1.0	1.86	0.53	0.53	0.78	0.49	0.81	0.58
MSE	0.28	1.28	4.45	0.57	0.47	1.38	0.42	0.95	0.52
RMSE	0.53	1.13	2.11	0.76	0.69	1.17	0.65	0.98	0.72
support	189	17	25	71	128	74	77	71	83

Table 5.6: Evaluation on gold labels, development set

	ADM	ATT	BER	ENR	ETN	FAC*	INS*	MBW	STM
MAE	0.42	0.85	na	0.55	0.55	0.57	0.52	0.82	0.47
MSE	0.36	1.07	na	0.57	0.53	0.66	0.38	1.01	0.38
RMSE	0.60	1.03	na	0.76	0.73	0.81	0.62	1.0	0.62
support	147	8	0	65	90	48	28	43	55

Table 5.7: Evaluation on output of the domains classifier, development set

- Evaluation on the output of the domain classifier: all sentences that were given the ADM label by the domain classifier are sent through the ADM levels model. This means that in the above example, both sentences will be assigned an ADM functioning score, and the note-level predicted score would be their average (let's say (3.6+0.4)/2=2).

Comparing the note-level error of these two evaluations gives us an indication of how big the effect of the background/target sentences is. In the first case, the note-level absolute error is 0.4, while in the second case the note-level absolute error is 2.

Table 5.6 shows the overall note-level results for the first type of evaluation (gold labels). This should be compared to the results of the second type of evaluation (output of the domain classifier), which are shown in Table 5.7. The observed effect is not very big. For most domains, the MAE in Table 5.7 is only slightly higher: +0.03 points for ADM, +0.02 points for ENR, +0.02 points for ETN, +0.03 points for INS, +0.01 point for MBW. For the other domains, the MAE in Table 5.7 is even lower than the one in Table 5.6.[4]

The results of the experiment suggest that the effect of the background/target

[4]Note that the comparison between the two tables is not direct, since the number of notes is different (see support). This is because not all the notes that have a gold label were also detected by the domain classifier.

sentences on the overall note-level score is not that big. Therefore, it was decided that there is no need in an additional classifier to filter out these sentences.

5.3.3 Conclusion

Based on the results of the experiments presented in the above sections, it was decided that:

- The levels-assigning models should be trained on a sentence-level, similarly to the domain classification model.
- There is no need to train a classifier that filters out the background/target sentences, since the effect of these sentences on the note-level functioning score is not that big.

Chapter 6

Discussion

The current phase of the project generated a number of deliverables and results: annotation guidelines for labeling the functioning levels for 9 ICF domains, a dataset of clinical notes annotated according to these guidelines, and a pipeline of AI models that assigns these functioning levels automatically. These results are summarized and discussed below.

Annotation guidelines

The annotation guidelines that were created for the project aim to provide a concrete, actionable scheme for describing the levels of functioning in 9 ICF domains. The ICF framework (World Health Organization 2013) offers a system of definitions and generic scales to describe functioning; however, our work shows that this system is not detailed enough for operational purposes.

To define what should be included under each domain, we used the official definitions provided by the ICF. Our analysis shows that the inter-annotator agreement over the domain labels is not high. This means that the task of deciding whether a sentence describes a certain ICF function is not easy, even for (para)medical students who were explicitly trained for this. This might indicate that the task is inherently difficult, partially because the boundary between an explicit description of functioning and an inference made by the reader is often not clear. In addition, it might also be a sign that the definitions regarding what should be labeled under each domain should be further clarified. Specifically for the INS domain, both the very low IAA score and direct feedback from the annotators signal that there is a problem with the current guidelines; the definitions for this domain should be reviewed in

future phases of the project.

To make the qualifier scales actionable, we created our own detailed interpretation of what constitutes a 'mild problem' or a 'moderate problem' for each ICF category. This yielded good results: a high inter-annotator agreement over the level labels. This indicates that the definitions of the levels scales in the annotation guidelines are clear and that the levels are easily distinguishable from each other.

Annotated data

In this phase of the project, about 6,000 clinical notes (286,000 sentences) were manually annotated. The annotation effort yielded 15,000 sentences with ICF domain labels (positive examples). The annotated data consists of clinical notes (of all types) from an academic medical center (Amsterdam UMC). Our analysis shows that this data has the following characteristics:

- About 5% of the sentences are relevant for at least one of the 9 ICF domains. This is true both for notes that are selected randomly and for notes that are selected based on keywords.
- The data is imbalanced in terms of the domains that it contains. Specifically, the ADM domain is very dominant (especially in notes that are related to COVID-19 patients), while the ATT and BER domains are very rare.
- The distribution of the levels per domain is also not always balanced; for example, for the FAC domain level 4 is very dominant, while for the STM domain level 2 is very dominant.

Machine Learning pipeline

The primary output of this phase of the project is an open-source classification pipeline that reads a clinical note in Dutch and assigns one or more ICF functioning levels to it. The pipeline includes a multi-label classification model that detects the domains mentioned in a sentence, and 9 regression models that assign a functioning level to sentences in which a specific domain was detected.

The domain classification model achieves an F1-score above 0.8 for 6 out of the 9 domains (on a note-level). This performance level is considered a

success because it was defined as the desired target at the beginning of this phase of the project. In addition, the model's performance on a sentence-level is similar to or higher than the inter-annotator agreement score (for all domains), indicating that in terms of detecting the relevant sentences, the model performs at a human-like level.

The three domains that do not achieve the desired level of performance are ATT, BER and INS. For the first two, the problem is that we do not have enough annotated data. As mentioned above, the hospital notes do not contain a lot of relevant examples for these domains; therefore, data from other sources (e.g. general practitioners, social workers, occupational therapists, etc.) needs to be obtained and annotated in order to build well-performing classifiers for ATT and BER. For the INS domain, the problem is likely related to the current definitions in the annotation guidelines. As mentioned above, this domain should be reconsidered for future phases of the project, since the current definitions do not generate consistent gold labels.

The levels regression models have a mean absolute error (MAE) below 0.7 for 7 out of the 9 domains (on a note-level). This means that, on average, the predicted functioning level for a note does not deviate more than 0.7 points from the gold functioning level. This is a good result (the pre-defined target was a MAE below 1); however, it is not as good as the human performance on this task, which stands on below 0.4 for all domains (on a sentence-level). The two domains that do not achieve the desired level of performance are ATT and BER; again, the reason is the low number of training examples.

For the next steps of the project, an important priority is to evaluate how well does the pipeline perform on different types of data: clinical notes from other hospitals and clinical notes from non-hospital sources (e.g. general practitioners, geriatric institutions, physiotherapists, dietitians, etc.). All the models in the pipeline, including the pre-trained language model that is used as the base for the classifiers, are trained on notes from the Amsterdam UMC. There is therefore a risk that the pipeline is over-fitted to this specific dataset.

Conclusion

The results presented in this report demonstrate that it is feasible to train well-performing AI models that automatically assign ICF functioning levels to clinical notes in Dutch. Possible directions for the future phases of the project include: (a) evaluating the performance of the existing pipeline on

other types of clinical data, (b) showing how the outputs of the pipeline can be used to answer clinical research questions (e.g. analysis of recovery of functioning over time), (c) improving the performance on the existing 9 domains by annotating more data and/or refining the annotation guidelines, (d) extending the method to other ICF categories.

Appendix A

Links to Resources

Machine Learning Pipeline

- GitHub: https://github.com/cltl/aproof-icf-classifier
- DockerHub: https://hub.docker.com/r/piekvossen/a-proof-icf-classifier

Models

- Domain Classifier: https://huggingface.co/CLTL/icf-domains
- ADM levels: https://huggingface.co/CLTL/icf-levels-adm
- ATT levels: https://huggingface.co/CLTL/icf-levels-att
- BER levels: https://huggingface.co/CLTL/icf-levels-ber
- ENR levels: https://huggingface.co/CLTL/icf-levels-enr
- ETN levels: https://huggingface.co/CLTL/icf-levels-etn
- FAC levels: https://huggingface.co/CLTL/icf-levels-fac
- INS levels: https://huggingface.co/CLTL/icf-levels-ins
- MBW levels: https://huggingface.co/CLTL/icf-levels-mbw
- STM levels: https://huggingface.co/CLTL/icf-levels-stm

Other

- Annotation guidelines: click here
- INCEpTION config: click here
- Keywords: click here
- Code (incl. data processing, model training, model evaluation, data analysis, etc.): https://github.com/cltl/a-proof-zonmw

Bibliography

Delobelle, P., Winters, T., & Berendt, B. (2020). RobBERT: A Dutch RoBERTa-based language model. *arXiv preprint arXiv:2001.06286*.

Hripcsak, G., & Rothschild, A. S. (2005). Agreement, the f-measure, and reliability in information retrieval. *Journal of the American medical informatics association*, *12*(3), 296–298.

Klie, J.-C., Bugert, M., Boullosa, B., de Castilho, R. E., & Gurevych, I. (2018). The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, 5–9.

Verkijk, S. (2021). *The role of domain specific language when modeling Dutch hospital notes with transformers using limited computational power*. M.A. Thesis, Vrije Universiteit Amsterdam.

World Health Organization. (2013). *How to use the ICF: A practical manual for using the International Classification of Functioning, Disability and Health (ICF)* [https://cdn.who.int/media/docs/default-source/classification/icf/drafticfpracticalmanual2.pdf (visited 2021-09-28)]. Exposure draft for comment. Geneva: WHO.