FINDING THE SMOKE SIGNAL: Smoking Status Classification with a Weakly Supervised Paradigm in Sparsely Labelled Dutch Free Text in Electronic Medical Records

Research Master's Thesis

ReMA Linguistics & Communication Sciences

Author: Myrthe Reuver

Supervisor: Dr. Iris Hendrickx

Co-Supervisor: Jeroen Kuijpers, MSc (Topicus)

Second Assessor: Dr. Nelleke Oostdijk

Centre for Language Studies

Department of Language & Communication

Radboud University Nijmegen

Date submitted to repository:

January 25, 2021

Date defense:

November 16, 2020

Acknowledgements

This thesis would not have seen the light of day without the help of my supervisors. Dr. Iris Hendrickx was always available when I needed her, even at odd times or under stressful circumstances, and was especially and helpfully supportive whenever I felt I could never finish this thesis, or only saw hurdles along the way. She trusted me and believed in me even when I did not. Jeroen Kuijpers from Topicus Healthcare, as my second supervisor, was also always supportive and interested in my ideas, and helped me navigate the complex data environment I found myself in. I count myself really lucky with these supervisors, and I hope my work shows my effort and their support.

I would also like to thank the LaMa research group (the Language Machines) for their support whenever I was once again complaining about my progress during our weekly meetings. Especially as the COVID-19 pandemic raged and I was working from my home office for months on two master's theses at the same time with barely seeing another soul, these weekly meetings were the highlights of my week and also provided much-needed support and discussions about research, research practices, and data. Some fellow-LaMas, aside from Iris, deserve special mention. Merijn Beeksma, who showed me by example how to be a conscientious, ambitious, and encouraging researcher and mentor (and how to become a ResMA with the specialization Language & Technology), and prof. dr. Martha Larson, who showed me how kindness, community, and precision are essential researcher characteristics, and how important teaching is.

And lastly, the quote that kept me going through the last few weeks of thesis writing:

"Cause I'm gonna be free and I'm gonna be fine
(But maybe not tonight)"
- Florence + The Machine, "Delilah"

Myrthe Reuver

Arnhem,
January 25, 2021

Summary (one page)

Smoking status is a clinical variable in (primary) healthcare, defining whether or not a patient smokes or has ever smoked cigarettes or cigars. However, it is currently under-reported by GP (General Practitioner) offices in the Netherlands. GPs experience a heavy documentation load, and often opt for describing complaints in the free text of the consultation (the 'SOEP' text) rather than formally documenting it in a variable - while a documented variable is more retrievable for clinical professionals needing this information later. This thesis attempts to use Natural Language Processing (NLP) and Machine Learning (ML) to automatically classify smoking statuses recorded in the free text of consultation reports in Dutch GPs. We found a specific problem: smoking status is under-documented and sparsely labelled in EMRs (Electronic Medical Records), while modern NLP approaches require large labelled datasets. We use a weak supervision as well as a Transfer learning approach to combat this "small dataset problem".

We attempt to answer the following question:*"How can we best automatically detect and classify the smoking status of primary care patients' EMR (Electronic Medical Record) on the basis of the free text in GP doctor's notes?"*

We worked with medical data storage company Topicus to obtain 17.873 EMRs from 6 GP offices in the Netherlands, of which only a sub-set is labelled for smoking status (4.978 training examples, 651 development examples and 628 test examples) into three classes: non-smoker, ex-smoker, and smoker.

Our results indicate Transfer learning is a potentially fruitful approach to smoking status classification. We found a fine-tuned pre-trained Transformer model BERTje model performs well (F1 (micro) = .79), and out-performs our rule-based baseline (F1 = .55). Our results however do not match earlier work's results, where rule-based methods already obtain high performance scores (F1 = .91) on similar smoking status tasks in English. We cannot replicate these high-performing rule-based methods, but our Transfer learning approach with BERTje is relatively effective at correctly detecting especially the non-smoker and ex-smoker class in EMRs. Increasing the training set size in a weak supervision approach with a generative labelling model does not increase performance of BERTje (F1 (micro) = .79), though does lead to a better classification of ex-smoker and non-smoker examples.

Thus, we find a Transfer learning approach with BERTje a potentially interesting approach for smoking status classification in Dutch EMRs even with small datasets. These now popular pre-trained models could be a step for research into smoking status classification away from rule-based methods.

Contents

1 Introduction

Smoking is a serious public health problem. The Dutch national bureau of statistics (CBS) records an individual's smoking status in three categories, based on self-report data: smoker, non-smoker, and ex-smoker. Their data indicates that 23% of Dutch adults above the age of 25 identify as smoker, with up to 50% identifying as smoker among those with little schooling (high school or less) (Centraal Bureau Statistiek, 2018).

Smoking status - whether or not a patient smokes or has ever smoked - is relevant for a number of treatment choices and health outcomes. These include (but are not limited to) medication use: recent research indicates smokers have a higher risk of opiate addiction when using painkillers (Young-Wolff et al., 2017), a higher risk of cardiovascular disease even when the patient's weight is controlled for (King et al., 2017), and a higher risk of lung problems including lung cancer (Islami et al., 2015). Thus, a physician's knowledge of a patient's smoking status can aid the patient's treatment and health, and such knowledge remains relevant when the patient's complaint is seemingly not directly related to respiratory issues. Smoking status also negatively affects the health outcomes of others, especially young children in a smoker's household, and even that of third parties in public spaces, as identified by the Care Standard and Regulation on Tobacco Addiction by the Dutch governmental organizations related to healthcare (Partnership Stop met Roken, 2019). This Care Standard document was designed in 2019 to improve and standardize care for tobacco addiction and reduce the number of smokers in Dutch primary healthcare.

However, smoking status is now often not reliably and uniformly recorded in EMRs. Instead, this information is often mentioned in the free, unstructured text written about medical consultations (Partnership Stop met Roken, 2019), which are known as SOEP texts in Dutch primary healthcare. Smoking status classification with Natural Language Processing (NLP) and Machine learning (ML) has been identified as a potential method to find smoking status information in such free text of consultations for decades (Uzuner et al., 2008; Palmer et al., 2019).

Such automatic methods depend on large-scale labelled datasets to train algorithms to recognize new, unseen cases. The healthcare domain usually does not have such large labelled datasets available for clinical problems, especially not publicly available. This is also the problem for smoking status classification: publicly available datasets for smoking status classification are small, or not (fully) labelled for smoking status. In response, research on the automatic classification of smoking status consists of solutions built with small data sets

and simple methods such as algorithms based on regular expressions (Palmer et al., 2019) or logistic regression (Weng et al., 2017). These methods are not the State of the Art (SOTA) for current NLP problems, and often under-perform compared to more advanced methods with machine learning, deep learning, and language modelling (Mikolov et al., 2013) on similar problems such as disease classification with text from EMRs (Zhao et al., 2019) and the classification of medical emergency levels in triage on the basis of text from the EMR (Horng et al., 2017).

We thus identify the under-utilization of modern methods and techniques on smoking status classification, and a major factor in this situation being the fact that there are only small datasets. We intend to overcome this problem with a weakly supervised paradigm: we label unlabelled data points with LFs (labelling functions) in a generative LabelModel trained with SNORKEL (Ratner et al., 2016, 2017), to then use a larger training set labelled by this model to train a machine learning classification algorithm. This allows for a model that is flexible enough to pick up on elements not seen by human developers of rules by the machine learning model, while it also allows for more training data than in the original labelled dataset. Additionally, we do not train a model from scratch, but use Transfer learning, where a large Transformer is pre-trained on language understanding (Devlin et al., 2019). We use a Dutch pre-trained language model (Vries et al., 2019) and then fine-tune this model in the final layer with a relatively small number of labelled smoking status examples. This also allows us to work with only a small number of labelled datapoints, since the model already has knowledge of the Dutch language from its pre-training phase. We thus intend to overcome the 'small dataset' problem for smoking status classification in two ways. First, by leveraging pre-training with Transformer-based models (Devlin et al., 2019) on large (language) datasets because we do not have such large datasets for this problem. Secondly, increasing the number of labelled datapoints for smoking status classification in a manner that is scalable (with a generative model).

Out of this analysis of the research gap and specific problem follows the following research question:

> *"How can we best automatically detect and classify the smoking status of primary care patients' EMR (Electronic Medical Record) on the basis of the free text in GP doctor's notes?"*

We intend to answer this question with three sub-questions:

- Can we obtain similar performance with a rule-based baseline on smoking

status classification as earlier literature (Palmer et al., 2019; Weng et al., 2017; Palmer et al., 2019)?

- Does Transfer learning, by fine-tuning a pre-trained BERT Transformer model (Vries et al., 2019), improve performance over the rule-based baseline in classifying smoking status?
- Does a weak supervision programming paradigm (Ratner et al., 2016), by labelling more training data points with a generative labelling model (Ratner et al., 2017), improve performance over the rule-based baseline and the earlier classification model in classifying smoking status?

This thesis will attempt to answer this question and its sub question with the following chapters:

Chapter 2 provides an extensive background to smoking status classification and smoking status detection, and to weakly supervised learning. It will identify the gaps in previous literature, as well as analyze some phenomena prominent in earlier work.

Chapter 3 describes the data and the extensive data pre-processing that leaves us with a dataset use-able for answering our research question.

Chapter 4 discusses our methods for smoking status classification, with Section 4.1 discussing our method for supervised learning with with a Transfer learning framework by fine-tuning of a pre-trained Transformer language model (Vries et al., 2019) as well as with our rule-based baseline. Section 4.2 describes our method for weakly supervised learning using a SNORKEL generative model (Ratner et al., 2017).

Chapter 5 explores the results of our smoking status classification experiments. Section 5.1 describes the performance of our rule-based and supervised models on our test set, while section 5.2 describes the performance of our weakly supervised model on the same test set.

Chapter 6 discusses the results in a broader context, explores the results' implications and factors involved, and also discuss choices made in previous chapters and the influence these might have had on the results.

Chapter 7 concludes the thesis by answering the research question and its partial questions.

Several **Appendices** provide additional calculations or background information. The text will refer to these appendices where they might be relevant for the reader. Right before the appendices, there is a **list of Figures** and a **list of Tables** provided, with descriptions and page numbers for reference.

2 Background

The following chapter will outline earlier work related to smoking status detection and classification, and also discuss related work that sort EMRs into groups based on free text in the EMR (in primary healthcare). It will also discuss the current state-of-the-art in (clinical) NLP as well as some problems related to clinical NLP. This chapter also introduces two useful paradigms for this study: weakly supervised learning, and transfer learning.

2.1 Smoking Status in Current Dutch Primary Healthcare

Medical information can nowadays be stored, processed and accessed more efficiently than ever before. **EMR** (Electronic Medical Records) document large volumes of medical information connected to one specific patient. EMRs consist of long-term, patient-specific data on consultations, medicine use, and health. Medical professionals such as GPs, medical specialists, and dentists record the details of a specific consultation, treatment, or illness, which aids the next medical professional treating the patient. EMRs are thus meant to be a reliable document of the patient's medical history. In the Netherlands, a version of the EMR has been in use since 2012. It is currently (in 2020) widely used by medical professionals in the Netherlands (Medisch Contact, n.d.).

However, not all information is documented in the EMR, or documented in a systematic and comprehensive way. Medical professionals often feel a tension between coding information systematically (for instance recording a patient's complaint or illness as a medical code, or recording information in a specific field of the EMR), and simply describing it in the text they write of the consultation or procedure in in the EMR (Ford et al., 2016). The latter option takes less time, as the medical professional does not need to navigate a codebook or an electronic interface to find the right medical code or sub-field. This reduces the documentation load, which leaves more time for actual medical work such as consultation and treatment. In the Netherlands, some documentation load is required by the government or health insurers, but additional, not required documentation usually does not happen (Partnership Stop met Roken, 2019). However, it greatly helps other medical professionals working with the patient if information is recorded in a structured and systematic manner, as this allows users of the EMR to quickly and reliably find and use the information.

One specific case where this tension is prominent is the **smoking status** of patients. A patient's smoking status is whether or not the patient smokes cigarettes, and usually also includes information on whether a patient has ever

smoked, and possibly also for how long and how many packs a year (Marston et al., 2014; Partnership Stop met Roken, 2019). Primary care facilities such as GPs can record smoking status, but this is not legally required in the Netherlands (Partnership Stop met Roken, 2019). Additionally, recent research into primary care in the UK has indicated that clearly defined definitions of terms such as "ex-smoker" and "non-smoker" are essential for reliable documentation of this information (Marston et al., 2014), as otherwise individual GPs record this information in different manners.

The 2019 Care Standard guideline for Tobacco Addiction was written together with healthcare organizations and the Trimbos addiction research institute, and requires all healthcare providers to advise smokers to quit smoking at least once a year (Partnership Stop met Roken, 2019). The Care Standard acknowledges that, in order for this to happen, documentation on smoking status needs to improve.

One reason for this lack of uniformity in smoking status recording in primary healthcare in the Netherlands is health professionals not consistently using the specific medical codes related to tobacco usage, claim a group of research, healthcare, and governmental institutions under the supervision of the Dutch Ministry of Health in the 2019 Care Standard for Tobacco Addiction: Partnership Stop met Roken (2019). One example mentioned in the care standard is that the DSM definition of "tobacco addiction" has different aspects and ways of diagnosis than the ICPC code for "tobacco abuse", meaning these categories have slightly different meanings. This leads to not every smoker being recorded in them (as the DSM notes, not every person smoking cigarettes is necessarily addicted, while of course a non-addict can also harm themselves by abusing tobacco). This means a "tobacco addict" is differently coded from a "tobacco abuser" in their EMR, while they both would likely be classified as a "smoker" by any medical professional. Such confusion of category and terminology is at the heart of the problems with smoking status classification.

The current codebook for primary case in the Netherlands, (*the "Tabel Diagnostische Bepalingen", version 33*), indicates there are 39 different ways of recording tabacco use in the EMR. These include binary variables, text field variables, and numeric variables, from "amount of sigarettes smoked a day", "how often patient tried to stop smoking", "consultation on smoking", "reasons not to stop smoking", and "given advice to stop smoking". Another documentation for smoking behaviour is the International Classification of Primary Care (ICPC) code for "tabacco abuse", P17, while the DSM-5 and ICD-10 also have a code for "tabacco addiction" (Partnership Stop met Roken, 2019).

Another factor is of influence in the documentation and the (lack of) record-

ing of smoking status in EMRs is that the Dutch primary healthcare system makes a principal distinction in their patient population, which also influences how well smoking status is documented. Around 10% of all patients have a chronic illness that requires long-term monitoring, such as heart disease, diabetes or COPD (Zorginstituut Nederland, n.d.). These patients are put in a special track within the healthcare system, called **Ketenzorg**, in which these patients have yearly contact moments with primary care, and have standardized documentation of their illness - including their smoking status. This more intense contact with primary healthcare has consequences for our research project, because it means there is a sub-group of patients that has their smoking status systematically and reliably documented, while for the majority of patients (the other 90%) this is not the case. However, this also offers possibilities: perhaps the 10% in Ketenzorg programs can be used to help more reliably organize and extract the information from the patients not in this specialized program. We further explore this option in our methods section and our further study.

The specific problem is thus that EMRs are not often labelled for smoking status, while such information is needed in primary care. In order to solve this problem, this study turned to Natural Language Processing (NLP).

2.2 Natural Language Processing and Machine Learning as a Solution

Natural language processing (NLP) is the analysis and understanding of language with computers and computational methods. It is an interdisciplinary field finding its roots in computer science, linguistics, mathematics, electrical engineering and even psychology (Jurasky & Martin, 2000). The advantages of using NLP techniques, especially for problems requiring the analysis of large quantities of text (also known as Text Mining), is that a computational approach can quickly and consistently process information where humans would not be able to do so. NLP is thus useful for solving a problem requiring the analysis of large and complex corpora of text, such as the smoking status problem in EMRs.

NLP, especially in combination with **Machine Learning**, has recently seen a drastic increase in use and performance. Machine learning is a technique where algorithms learn to independently do tasks that generally require intelligence, without being specifically programmed to do these tasks (Ng, 2020). Machine learning algorithms are usually trained, by seeing examples of data, to then perform their task on data not seen in training. The use of machine

learning has attracted increasing interest in the medical domain over the use of "expert systems" (systems built out of rules made by the help of professionals in a certain domain) (Obermeyer & Emanuel, 2016). A system based on rules designed by experts, such as "if the level of value X is above Y", does not learn from the data - making it fundamentally different from a machine learning approach, and also less flexible and adaptable to new data.

Recent uses of NLP and machine learning on medical text include the prediction of the palliative phase on the basis of EMRs (Beeksma et al., 2019), detection of medical concepts in unstructured text (Tulkens et al., 2019), or predicting the risk of type 2 diabetes based on information in the EMR (Mani et al., 2012). The most common recent techniques used in such research are word embeddings, a manner of capturing semantic information. Word embeddings do this by placing words in a high-dimensional vectorspace while placing words with similar contexts near one-another (Mikolov et al., 2013). Word embeddings are trained on large datasets of texts by predicting the context from the word, or the word from the context. Words in similar context thus get a similar vector representation, and in practice these words usually have similar meaning. In this way, implicit semantic information is encoded in the vector. Another technique that has seen recent widespread use in NLP and text classification is artificial neural networks (ANN): a highly flexible class of machine learning algorithms using vectorized matrix operations, able to also learn complex semantic and linguistic information from text. A neural network can be defined as "a parallel, distributed information processing structure consisting of processing elements (which can possess a local memory and can carry out localized information processing operations) interconnected together with unidirectional signal channels called connections" (Hecht-Nielsen, 1992). Such an architecture can learn to map any mathematical input-output pair in training, and then after training be used for new analysis on unseen examples. Backpropagation is able to learn this efficiently by computing the gradient of the loss by the weights of different computations in the network, and then adjust these weights to better predict the outcome. The algorithm thus optimizes itself to classify unseen examples.

A machine learning approach is thus more flexible (as it learns from the data and can adapt to new data) and more suited to large, complex data with complex relationships than a system with pre-programmed rules. This project will thus focus on developing a Machine Learning solution to solve the problem of smoking status classification.

As noted in the systematic overview by Kreimeyer et al. (2017), 46% of NLP projects aiming to identify and extract elements from unstructured text in EMRs

still use rule-based systems, while the other large group uses a hybrid approach of Machine learning and expert rules (26%). Only a minority (4%) uses solely Machine Learning and does not use expert rules at all. This is remarkable, since more general NLP has moved on to more advanced methods than heuristic rules. Another systematic review Ford et al. (2016) also shows that research in clinical NLP often does not use machine learning but rules made by experts and keywords to tackle complex text mining problems such automatically identifying certain illnesses.

Most recent developments in the NLP field have moved away from rule-based systems to other techniques. The most recent developments in NLP and machine learning are the adoption of large-scale pre-trained Transformer networks. The central idea in these models is transfer learning: the transfer of knowledge from one task to another. The famous first uses of such models include BERT (Devlin et al., 2019) and ELMO (Peters et al., 2018), trained on millions of texts in performing a specific task, such as predicting masked words or predicting the next sentence. The previously attained linguistic knowledge in these networks allows the network to more easily solve new, language-related problems. However, this knowledge is language-specific, so recently researcher have made Dutch BERT models (Text Mining Research Leiden, n.d.; Vries et al., 2019), and also domain-specific, explaining the rise of models like medBERT for the medical domain (Alsentzer et al., 2019).

These large-scale language models are able to capture semantic and linguistic content from text by training on very large datasets of text. The task they are trained on is predicting masked words, or predicting the next sentence, based on input. This enables such models to learn which words are likely to occur in similar contexts, and capture such semantic similarity in the word vectors. In a large vector-space, words with similar contexts are thus placed close together. This technique builds on earlier language modelling approaches, such as predicting a word from its context and context from its word in the `WORD2VEC` method (Mikolov et al., 2013).

Especially BERT and BERT-based models have seen widespread recent use. BERT's innovation over earlier (large-scale) masked language models is that it is a Transformer-based neural model which uses bidirectional prediction: a word is predicted both from the left-to-right and the right-to-left word context, in a large neural model able to learn from feedback, as all neural models with backpropagation. The use of BERT or BERT-based language models has led to greatly improved improvement all kinds of NLP tasks, from classification to question answering.

These modern methods are not usually utilized in earlier research on smok-

ing status classification, which we will explore in the next section.

2.3 Smoking Status Classification and Information Extraction in Earlier Literature

The first work related to automatic smoking detection and classification is a shared task from 2006, described in (Uzuner et al., 2008). It had five smoking status labels to predict (never, past, current, smoker temporality unknown, and unknown smoking status), and the shared task consisted of consultation notes in 502 patient dossiers. The most successful approach, in terms of performance measured in F1 score, was designed with Weka and "Nuanced Medical Extraction" rule-based system that identified markers such as smoking-related medication to get features, together with unigrams and bigrams, and then used SVM as a model to train on these features. This approach already pre-filtered the training material for phrase-mention of smoking, as they classified anything not mentioning "smok" as of unknown status in their first step. Another successful approach in this study first searched for the stemmed tokens "smok", "tobac", "cigar", classifying the rest as "unknown", and then used linear SVMs for further classification, again with tri and bigrams. This solution fits neatly in the trend (Ford et al., 2016) see for clinical NLP: relatively simple, rule-based solutions are widely used.

Currently, the most cited paper related to smoking status in NLP is a paper on automatic classification of patient groups related to the risk of cardiovascular disease Weng et al. (2017). One of the features used is smoking status. They tried to identify patients that will ever have cardiovascular issues, and in their study compare their system to a currently used baseline (clinical guidelines changed into if/then rules). The machine learning approach led to significant increase in precision, highest scoring is a neural model with .67 accuracy, but most models are underspecified in the work, making it unclear which architecture or design choices were used (a simple description such as "neural networks" does not fully explain which technique or architecture is used). Additionally, Weng et al. (2017) do not provide much attention to how this experimental outcome is tested and validated, the authors only mention the use of a random 25% validation set. This also signifies a common trend in clinical NLP: the research methodology and research term clash between clinical researchers and machine learning experts. We will discuss this further below.

Later approaches on smoking detection seem to follow the earlier proven rule-based regular expression approach, with systems working to simply detect the words "smoking" and "tobac". Interesting is a recent bioinformatics paper

by Palmer et al. (2019). This group worked with a small dataset (758 instances), with as goal finding more detailed information about smoking behaviour such as packs a year and cessation date (when patient stopped smoking). Methods tested where a rule-based system with regular expressions. They also tested SVM with the stemwords used in the best solution from the shared task from 2006 (!): "smok", "cig", "tobac", "nicoti", with +/- 5 words surrounding the words. They received F1 = 0.90 with simple regex rule-based system, but also have some validation issues. They have self-labelled data, and also artificially upsampled the amount of smokers to 50% of all training data by using a search for stem "smok" as detector, while this is not realistic distribution of smoking data, nor a good representation of all "smoking" data. Notable is that a rule-based classification method from 2006 is still used in a paper from 2019.

2.4 Challenges in Clinical NLP in General, and Smoking Status Classification in Particular

We thus identified a main problem: work concerning the identification of systematic information in free clinical text (such as smoking status) often does not use state-of-the-art methods. The question remains why. There are several substantial challenges to implementing state-of-the-art practices and techniques in clinical NLP.

2.4.1 Explainability

One possible reason are possibly the low explainability of more state-of-the-art approaches to NLP, such as word embeddings and neural networks. These methods are highly effective at capturing and extracting complex semantic and linguistic information in language, but are often not able to provide explanations of their decisions. An example is Beeksma et al. (2019)'s use of Long Short Term Memory networks to identify the palliative phase for individual patients. The algorithm used performs better than medical specialists at identifying palliative patients, but cannot fully explain how it does so. This lack of explainability can be problematic in an age where patients' medical information is increasingly their own, and where patients have a right to fully understand why and how their data was used to reach a certain medical decision. A medical or clinical application is a high-stakes domain as identified by Rudin (2019) in her analysis of the use of uninterpretable models, meaning such an algorithm or model will greatly influence the individual's life and choices, and thus the patient has the right to understand and interpret the algorithm's decisions.

2.4.2 A Babylonic Misunderstanding

A related hurdle is the lack of understanding between the medical discipline on the one hand, and the computer science and AI field on the other. This is especially visible in methodology. While clinical research is rightly and carefully built upon rigorous inferential statistics (most commonly null hypothesis significance testing), the AI field uses prediction and performance metrics related to a held-out test set of examples as their main approach to determining the validity of an approach or study. These approaches can be complementary, but are often seen as contrasting and incompatible. Together with a lack of understanding what machine learning exactly is, and similar or the same terms being used in both fields, this can lead to confusion. For instance, "logistic regression" can be used for inferential statistics and is often used in this manner by clinical researchers, while it is essentially a machine learning model that learns from the data, and thus also used by the computer science field. This Babylonian language confusion can be seen in a fairly popular paper (Jie et al., 2019), where the authors (clinical research specialists) claim "Machine Learning" is not better than "logistic regression" for clinical prediction, without clearly defining the boundaries between the two concepts and creating a false contrast. Additionally, the authors seem to not understand the fine-grained differences in machine learning architectures, especially different neural networks, or use terms underspecified machine learning specialists (such as "classification trees", which can refer to several very different learning algorithms). This same problem is visible in Weng et al. (2017), where the authors describe their algorithm as "neural network" without further specifying the architecture of their model. Neural models are a class of models, and knowing exactly which "flavour" is used is essential in understanding and reproducing the results.

To further illustrate the Babylonic confusion in using terminology in this field, see Table 1 below, designed by Maarten van Smeden in a Twitter post on 5 February 2020 (van Smeeden, 2020)). Dr. van Smeden is a Dutch clinical specialist. The post, while not grounded in rigorous scientific study, illustrates the confusion the two disciplines and paradigms lead to when a researcher trained in one of them attempts to use methods from the other field, with terms such as "model" and "noise" having different or conflicting definitions.

These terms are, furthermore, often not 1:1 'translatable'. A feature is not always a variable, and the term "confound" as used in hypothesis testing with traditional statistics does not fully translate to "noise", as noise can also be a "measurement error". This makes for a confusing landscape for researchers on the crossover area between especially clinical prediction studies and clinical

NLP/Machine Learning, and could explain why we see such under-specified or simply wrong definitions in these studies. In order to correctly label and identify clinical classes or variables, one needs clinical expertise, but for machine learning research one often needs expertise in the data science and machine learning field as well.

Table 1: Babylonic Misunderstandings between traditional medical statistics and the NLP/Machine Learning field, as based on Maarten van Smeden's table. These terms

Traditional / Inferential Statistics	**Machine Learning/NLP**
Fitting	Learning/Training
Measurement Error	Noise
Predictor Variables	Features
Outcome Variable	Target/Label
Model for Discrete Var. (often: Logistic Regression)	(Supervised) Classifier
Model for Continuous Var. (often: Regression / Linear Model)	(Supervised) Learning Model
Covariate/Confound	Noise
Sensitivity	Recall
Positive Prediction Error	Precision
Derivation-Validation	Training-Test

2.4.3 The Hungry, Hungry Neural Network

A third problem is related to the data size, as we discussed briefly above in Chapter 1 as well. First of all, due to very reasonable privacy and practical concerns, clinical data is not easily accessible to researchers, while modern machine learning research benefits from large quantities of open-access, shared datasets. This ethical conundrum is well described by Suster et al. (2017). The process to get permission to use clinical data, especially clinical data with personal information on individual patients, is long and complicated while research and progress in the Machine Learning and especially NLP field happens fast. Methods of only three to five years ago, such as Support Vector Machines, seem outdated compared to word embeddings and even, especially, pre-trained Transformer models. This means such 'data delay' can cause slower progress in clinical NLP, which might be another factor contributing to the use of older methods in NLP recent as identified by Ford et al. (2016) and Kreimeyer et al. (2017).

A related data problem is sparsity of labels. For supervised machine learning, where a machine learning algorithm learns to categorize or predict new cases based on data seen in training, the training-data needs to have a label, for instance the medicine name or group membership. However, due to the ear-

lier mentioned lack of systematic recording in EMRs and medical data, labels are often missing. This is especially challenging because state-of-the-art neural models require large quantities of data. Solutions have been the adoption of unsupervised methods Tulkens et al. (2019), or attempts to work with carefully selected smaller datasets as mentioned by Suster et al. (2017). However, these smaller datasets have their own problems concerning generalizability across situations, hospitals, or regions.

Especially the last problem is also a problem for smoking status classification and detection in (Dutch) EMRs. As stated before, even positive cases (where the patient is a smoker) are not always consistently labelled. This makes training any supervised model quite difficult, as there is limited training data. This is especially the case for Dutch-language corpora and data, as the majority if not all publicly available smoking status classifiers are based on English-language EMRs and pre-trained models.

We can thus conclude that NLP is a suitable method for approaching the detection and classification of smoking status in Dutch EMRs, but that several hurdles remain in the use of advanced NLP techniques on medical data, and especially concerning smoking detection and classification.

2.5 (Weak) Supervision

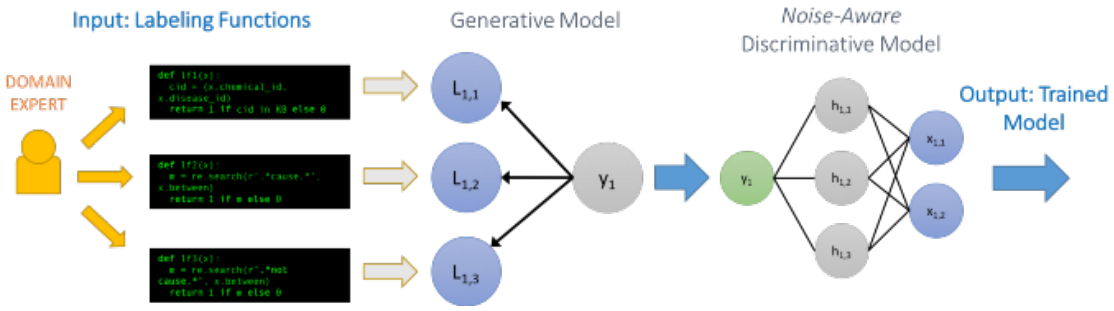
In Machine Learning, the most common approach is to provide a training algorithm with labelled examples. **Supervised learning** is a machine learning paradigm where a model learns to label, classify, or group new datapoints based on earlier seen data in training. For instance, seeing many examples of patient EMRs with the label "smoker" versus "ex-smoker", the model will be able to find the parameters distinguishing the two classes, and apply these parameters to new data. In order for this to work, the key ingredient is correctly labelled training data. Especially when the labels require specialist knowledge, experts in the specific (sub)fields are needed for careful and correct labelling.

As Ratner et al. (2017) explain, a common bottleneck in current machine learning research is lack of such labelled training data. Wang et al. (2019) and Suster et al. (2017) noted, this problem is even more pronounced in clinical NLP: datasets are often hard to find due to understandable privacy and ethics constraints, and the data that is available often lacks reliable, systematic labelling. Ratner et al. (2017) discuss several ways to solve this problem. One of them is the recently popular **transfer learning** framework. In this context, transfer learning would mean using a pre-trained model, trained before with other data. However, in order for this to work, the pre-trained model *also* needs

enough access to training data. We would be transferring the 'data scarcity' problem simply one step ahead in the pipeline.

Another approach would be to obtain more labelled datapoints, without having to hire new expert labellers while retaining the high-quality of the labels. This could be done with a so-called **weak supervision** paradigm. In this form of supervised learning, the labels are of lower quality because they are either made by non-experts, based on other values or heuristics, or even based on expected distributions and statistics of the data (Ratner et al., 2017). In terms of weak supervision, there is a recent paradigm possibly interesting: the approach by software package SNORKEL developed by Stanford, as described in Ratner et al. (2017). It does not only use regular expressions and keyword-based Labelling Functions (LFs) tested on a hand-labelled training set, but also checks whether a label assigned by these rules is probable. It does this by comparing the different LFs to determine inter-LF agreement but also the most accurate LFs. All LFs get a vote, which allows for fuzzy labelling (e.g. between 0 and 1) to denote certainty, with LFs also having the option to abstain from judgment when a certain probability threshold is not reached for an instance. This trained model can then be used to label millions of datapoints with high accuracy, and this larger training set can in turn be used to train models. SNORKEL is currently used by models of companies such as IBM and Intel. In Figure 1, we see the entire pipeline explained of a SNORKEL model as explained by Ghelani (2019). Labelling Functions are used to determine a label with a generative model, after which the LabelModel is trained to distinguish noise from signal by weighting the different LFs and exploiting their (dis)agreement. The output is then a trained LabelModel that can be used to label unlabeled datapoints reliably.

Figure 1: Visualization of SNORKEL from Ghelani (2019) , explaining the Data Programming paradigm as implemented in SNORKEL. The idea is that experts make Labelling Functions (first step on the left), which are tested on a subset of the data - which allows researcher to test their accuracy. Then, Labelling Functions are used to predict a label with a generative model (second step on the left), after which the Noise-Aware LabelModel (third step from the left) is trained to distinguish noise from signal by weighting the different LFs and exploiting their (dis)agreement. The output is then a trained LabelModel that can be used to label unlabeled datapoints reliably. After this, a trained LabelModel can label more examples, which can then be used for training a ML-model.



A useful usecase for our problem is Wang et al. (2019), who also noted lack of ground-truth labelled data in clinical NLP due to privacy and time constraints. They use the weakly supervised paradigm on three clinical NLP cases, one of them being smoker status detection and classification. They used 23.336 instances from the Mayo Clinic, and used a regular expression algorithm tested on 475 hand-labelled dataset to label the other datapoints into the five smoking status classes used in the shared task as seen above.[1] This weakly labelled dataset was then used in testing several algorithms, with CNNs performing the best in terms of F1-score, but a simple RegEx-based rule based system also performed well, better than SVMs. Another interesting find was that any training size increased above 5,000 did not lead to increased performance for the CNN model, which was 20% of the dataset. Word embeddings beat any other form of featurization, including tf-idf.

Earlier literature thus shows us two key findings:

- Regular Expression or simple rule-based systems, sometimes already work fairly well for smoker detection or smoker status classification, and have been in use for decades;

- The problem of sparse labelling can potentially be helped with a weakly

[1]However, it must be said that the paper explained the weakly supervised paradigm with the SNORKEL (Ratner et al., 2017) software, but did not provide information on their SNORKEL labelling algorithm, instead only providing a simple rule-based regular expression labeller.

supervised paradigm, which as been applied before to smoker status classification, or the popular transfer learning approach.

2.6 Ethics: Humans in Categories

Several ethical questions and problems arise with clinical NLP research like the current project. One specific issue we would like to address comes from the fact "smokers" can be seen as a classification of *people*, while we are working with representations of patient *documents*. These two things are not the same. An Electronic Medical Record is a representation of data on a person, and not representative of a person. They are a reduction of all complex factors in an individual patient, and also likely have missing or incorrect information.

This becomes relevant in the use of such models to *detect* smokers, or classify humans into smoking status categories. The fact is: a model trained on EMRs is unable to do this. It is trained on an by itself rudimentary representation of a human's medical history, with in its recording of data also biases: positive results are more likely to be recorded than negative ones for the general population of patients, while chronically ill patients have been required by law and insurance companies to have a recorded smoking status. In other words: for some patient groups, smoking status is more likely to be recorded than for others, leading to bias in any trained model.

Another important point is that machine learning algorithms can be wrong. Mathematical models lead ultimately to reduction and error when actively applied in the categorization of humans, as O'Neil (2016) has shown. This is a high-stakes domain as identified by Rudin (2019), which means these algorithms could have real impact on an individual's life, which is also problematic because such models are indeed very fallible and far from 100% accurate.

In the medical setting, we find Bowker and Star (2000) have described the problems in classifications and abstractions made by doctors, nurses, and hospital administrators. They acknowledge classification is useful as a record-keeping practice. Classification of cases and patients have the aim to give access to the past and prepare for the future. However, classes are also often not able to deal with ambiguity and uncertainty. Note that, for usefulness in clinical practice, classification forms and schemes from death certificates to ICD diagnostic codes need to accept some inherent fuzziness in the categories, as not every case neatly fits one cause of death or into a diagnosis. Bowker and Star (2000) also mention the important concept of *duration*: membership to a certain class does not have to be fixed or eternal, and individual patients or cases can move from class to class. This concept is especially important for smoking

status classification, as it is inherent in this concept that patients will leave certain classes (smoker) when they quit smoking. Computational models especially do not work well with classes with limited duration, as the assumption is that examples are examples of a class and only of this class. This underlying assumption of our method thus does not fit the real-life data practice.

Results of this project would be potentially harmful if medical professionals, or potentially researchers analyzing large-scale data, believe an algorithm's categorization into smokers and non-smokers applies to *people* instead of EMRs, and also believe it is (nearly) always accurate in classification of smoking status. In many situations, this is not the case and can lead to cascading errors in research and analyses based on data labelled with our model(s). This thesis project aims to prevent these problems by clearly stating this in this thesis, as well as advising the company we worked with (Topicus) to not use this model to predict smoking status of individual patients.

This study also aims to prevent the problem of bias for chronically ill patients by training separate models for patients in and outside of chronically ill programme (Ketenzorg). Additionally, we deal with the *duration* problem by looking only at one point in time (the latest identified label in the last four consultations), and identify whether EMRs which in that time frame switch from class membership are classified differently.

In the following chapter, we will describe our data and the methods and processes used to preprocess this data for smoking status classification.

3 Data and Preprocessing

3.1 Data Description

For this project, we worked together with software company Topicus. Topicus is a software provider for the public domain, and one of their services in the medical domain is the storage and management of data for GPs (General Practicioners), care centers, and other medical professionals. Topicus was able to provide us with data on patient consultations from 6 large GP offices in the Netherlands. These GP offices gave their permission for us to use their data for the development of a novel method for smoking status classification. We accessed the data only in a secure server environment and were not privy to any personally identifiable information such as name, birthyear, birthdate, year, place of birth, address, or even the city of the GP office.

A query on these six GP databases provided us with data available in storage from these GP offices. This data consists of EMRs of all patients currently registered to these GPs, and especially their documented doctor's consultations, which encompasses the past 9 years (2011 to 2020) - but with most GPs only storing the previous five years (2015-2020) in this system, leading to 75% of data being from 2015 to 2020. [2]

The query retrieved patient ID, age, and sex (M/F) information on patients from these GP offices, text records of their consultations with the GP, and also retrieved measurements on the Ketenzorg chronic illness programs as well as measurements related to smoking. These smoking-related measurements include 'P17', a long-term illness ("episode") icpc code for tabacco addiction, and '1739', a smoking status measurement which categorizes smoking behaviour with more detail into non-smoker, past smoker, and current smoker, much like previous literature on smoking status.

This query on the 6 GP offices provides 4 distinct data files for each GP office: a `PATIENTS` file with information about patient's age and sex; a `CONSULTATIONS` table with text on patient's GP consultations; a `MEASUREMENTS` table with patient's icpc codes and diagnoses (in this case, 42 related to smoking as well as identification whether a patient belongs to the long-term care program Ketenzorg); and an `EPISODES` table where long-term illnesses are registered with icpc codes.

The 24 datafiles in total contain information on 46.064 patients, of which 21.260 have at least one recorded consultation. Of these patients, 1.459 have a

[2] We decided not to restrict the data retrieval to the past five years, however, since the labelled data we needed for supervised classification of smoking status is rare and our neural Machine Learning (ML) models need large quantities of data.

P17 tabacco notification (one, but not the only, documentation of current smokers). Consultations are from 2011 to 2020, but with most (75%) falling between 2016 and 2020. We see all GPs have between 5.000 and 10.000 patients and are comparable to one-another in terms of the ratio of patients that have recorded consultations and recorded long-term tabacco addiction. This latest diagnostic is not often used, with only a small percentage of patients in a GP office receiving a P17 label. We will further describe some general characteristics of these different data files below, and also provide some general statistics in table 1 below.

Table 2: The content of each of the four tables for each GP office before preprocessing procedures. We see all GP offices are roughly comparable in size and number of consultations, though some (e.g. GP1) have more consultations per patients than others (e.g. GP3). The dataset has 21.260 unique patients who have had doctor's consultations, of which 1.459 have a P17 tobacco notification.

NOTE: this is before under-age patients are deleted, so shows more EMRs than will eventually be used in this study.

		PATIENTS	CONSULTATIONS	MEASUREMENTS	EPISODES
GP1 (0000)		5.022	241.220	25.377	286
	N patients with consultations	3.868			
GP2 (0594)		8.214	127.104	12.170	219
	N patients with consultations	3.601			
GP3 (1438)		5.804	119.709	10.144	210
	N patients with consultations	2.900			
GP4 (4970)		11.475	213.032	10.519	181
	N patients with consultations	3.885			
GP5 (6179)		4.781	110.517	13.018	308
	N patients with consultations	3.454			
GP6 (9048)		10.768	132.175	13.209	255
	N patients with consultations	3.552			
All		46.064	943.757	84.437	1.459
	N patients with consultations	21.260			

We first investigated whether our data could actually answer our research question, since our research question demanded very specific data. We report on this in the next section, after which we will describe our preprocessing steps.

3.2 Can This Dataset Answer Our Research Question?

One of our questions is related to finding whether increasing the training data set size with a programming paradigm and SNORKEL increases performance of smoking status classification for clinical notes. This research question demands a very specific dataset: one in which a large group of patient files is labelled for smoking status, while a portion of patients have smoking status recorded in a SOEP descriptive text on a GP consultation without this being formally labelled.

These specifics are needed because we want to classify smoking status, and also test whether we can use weak supervision in order to enlarge the training set. For weak supervision to be reliable, we need unlabelled datapoints that do contain smoking status information a weak labelling function (such as regular expression-based rules in SNORKEL models) would be able to identify.

We thus need to ask ourselves: does our dataset meet these specifications? In order to answer this question, we take two random samples from our un-preprocessed dataset, one small sample of 1.000 doctors consultations and one larger sample of 10.000 doctors consultations. Note that these are individual consultations, and not the representations used for smoking status classification. Both samples are only a fraction of the full un-preprocessed dataset, which consists of 943.757 consultations.

One assumption, strengthened by domain experts and especially data experts who worked on this database before, was that ketenzorg patients (patients in programs for chronic illnesses such as COPD and heart disease) are a distinct population from the general patient population, with especially a better recording of smoking status because of government and health insurance regulation of these programs. More attention to the recording of smoking status for these patients would lead to better recorded smoking statuses. We thus investigated the prevalence of the smoking status classification variable 'P1739' within and outside of the patients labelled as ketenzorg patients. There might also be some inherent relationship between Ketenzorg patients and smoking status: some chronic illnesses, such as COPD, are related to smoking, while chronic illness is more prevalent in lower socio-economic classes (Centraal Bureau Statistiek, 2018), which also see more patients smoking (Centraal Bureau Statistiek, 2018).

An important note on these analyses is that we do not know whether all ketenzorg patients are reliable labelled, and whether this dataset contains many false negatives that are not labelled 'ketenzorg' while in fact they are in the ketenzorg program. In fact, some domain specialists have noted that it is diffi-

cult to write a query returning all Ketenzorg patients. Usually, membership is documented with a binary variable. We used three possible queries with these binary variables: one where any Ketenzorg variable (whether heart disease, COPD, or any other chronic illness program) was positive, one where the 'GP as primary health professional' variable was positive, and one where both variables were positive. Best results were obtained with the first option - combining the ketenzorg variables to one binary variable, and leaving out the healthcare provider variable. Requiring both to be positive actually reduced the number of ketenzorg patients to 0, even though database experts considered this combination one that was more reliable for identifying these patients than either one.

3.2.1 Random Sample of 1.000 consultations

In a random sample of a 1.000 consultations, there are 941 unique patients. In these 1.000 consultations, 22 mention smoking when testing with the regular expression "([Rr]ookt niet|[Rr]oken|[Rr]ookte|[Rr]ook|[Rr]oker)".

These consultations also belong to 22 unique patients. Of these consultations, 4 are labelled with 1739, also 4 unique patients. Which means 18 patients do mention smoking in their consultations, but are not labelled with 1739. This means smoking status is indeed under-reported in these EMRs, with more EMRs mentioning smoking status in the free text than in a recorded variable.

Of the labelled consultations that do mention smoking in the free text, only 1 patient is in ketenzorg. However, this means that percentually 7.69% of the 13 ketenzorg patients mention smoking and are labelled. This counts for only 3 out of the 901 non-ketenzorg patients, or 0.33% of these patients. This suggests ketenzorg patients indeed have a higher prevalence of recorded smoking status.

3.2.2 Random Sample of 10.000 consultations

In a random sample of a 10.000 consultations, there are 6134 unique patients. In these 10.000 consultations, 121 mention smoking with regular expression "([Rr]ookt niet|[Rr]oken|[Rr]ookte|[Rr]ook|[Rr]oker)". These consultations also belong to 121 unique patients.

Of these consultations, 35 are labelled with 1739, also 35 unique patients. Which means 86 patients do mention smoking in their consultations, but are not labelled with variable 1739. Of these labelled consultations, 9 patients are in ketenzorg. Of all labelled consultations, 25% (9 out 35) is thus KetenZorg.

Of all KetenZorg patients (103), 9 mention smoking and are labelled - which

means 8.7% of all Ketenzorg patients have labelled training examples. Over all EMRs (6134) this is only 35 EMRs, or 0.57%. This suggests the KetenZorg EMRs are indeed better labelled on smoking status compared to the general population.

Table 3: Two random samples of 1.000 and 10.000 consultations, and showing the statistics (how many labelled for smoking status, how many mentioning smoking in the textdata). This provides evidence this dataset is sufficient for answering the research question on weak supervision, because a subset is labelled and a larger subset mentions smoking status in the textdata without being labelled.

	1.000 random	**10.000 random**
unique consultations	941	6134
ketenzorg consultations	13 (1.38%)	103 (1.68%)
RegEx mention smoking	22	121
mention smoking & are labelled	4	35
mention smoking, & labelled, and KetenZorg	1	9

3.2.3 Two populations: Ketenzorg and non-Ketenzorg

The fact that KetenZorg EMRs are more often labelled might not only be because they more often see their GP and because of more attention to smoking status for these patients, but because of related aspects of the ketenzorg population.

Ketenzorg population are patients who have COPD, heart disease, or diabetes. These patients, more often than the general population, likely smoke, since there is a causal effect between smoking and these chronic illnesses. And positive smoking status are more likely to be registered, especially if this smoking status has an influence on health status - which is likely the case with COPD and heart disease. Another factor of bias is that smoking behaviour is more common in people of lower social-economic status, who also more frequently deal with chronic illnesses. These factors can all lead to the higher prevalence of smoking status labels in the ketenzorg population.

There are thus both inherent biases (with a smoking status being prevalent in a certain population, the chronically ill) and recording biases (with the GPs more often recording it for a the same sub-set of the population due to requirements.

Our research entails training models on this dataset. This might lead to bias towards consultation descriptions of patient who are chronically ill, have

lower socio-economic status, or both, while not detecting smoking status for patients who are higher-class or not chronically ill. This bias in the dataset, where Ketenzorg patients are more often labelled, can thus lead to a biased model. Additionally, a model trained on ketenzorg data might fail to detect non-ketenzorg smokers, since the specific descriptions of chronically ill, lower-class patients could differ from other patients in the general population.

Testing whether Ketenzorg patients are indeed a different population than non-ketenzorg patients, we performed a t-test that showed the difference in age was significant (t = 45.7, p=0.0), with the ketenzorg (chronically ill) patients significantly older, though the SD measure shows the age is spread widely within both groups. We also found the ketenzorg group had relatively more male patients (50.1%) than the percentage of males in the non-ketenzorg population (42.6%). The ketenzorg group is thus significantly older than the general population, and consists of more male patients than the non-ketenzorg patients.

In order to prevent a bias, or at least study its effects, we also separate ketenzorg patients from non-ketenzorg patients when training our models to see what the effect is of this variable.

These datafiles were not immediately useable for our research question. The following section discusses steps we took to preprocess the data for our research question and method. We worked in a step-wise fashion to preprocess the dataset for our use, also shown in figure 2 below. Step 1 to 3 are performed iteratively on data from one GP office (4 datafiles) at the time, while from step 4 we are working with one combined dataset with data from all GP offices.

3.3 Preprocessing

3.3.1 Step 1: `CONSULTATIONS` and unique ID

Within the `CONSULTATIONS` table for each GP office, every row consisted of a text field. The texts on consultations are stored in a different variable for different text entry fields, related to the SOEP system of GP administration: Subjectief (Subjective), Objectief (Objective), Evaluatie (Evaluation) and Plan (Plan), which allows the GP or medical professional to document different aspect of a single consultation. A subjective aspect is for instance how the patient feels or describes symptoms, while the GP notes heartrate or medicine use under "objective". "Evaluation" is where the GP evaluates the situation, and "Plan" is where the GP or medical professional documents what steps are taken to take care of the situation. From a cursory glance at our dataset, we see that sometimes this distinction between objective and subjective observations and measurements is dilligently kept, while at other times all information is sim-

ply entered in only one of the four text entry fields, with the others remaining empty.

We decided to merge all these textfields to obtain one description of one GP consultation, but also retain the separate fields for possible future research interests. We first made each date for individual patients consistent with one row. We were then able to combine all text written on this one date in one variable.

For each table in each GP office, we also changed the ID variable to one where the patient id was combined with the GP office ID, in order to ensure each patient had a unique identifier that allowed us to link the different tables to one-another.

3.3.2 Step 2: adding PATIENT information and age

We were then able to connect the patient's information in the PATIENT table, such as sex and age at query extraction time, to the texts on consultations by the use of this new unique ID.

By combining the CONSULTATIONS table with the PATIENTS table, we were able to now also calculate a patient's age at the time of a consultation: by extracting the year from the consultation date, subtracting this year from the year 2020 for the time passed since the consultation, and then subtracting this number from the patient's current age as registered at extraction time.

The outcome of this step is thus one table with patient and consultation information for each patient, within one GP office.

3.3.3 Step 3: adding the smoking status and chronic illness indicators

For our classification experiments, we need labelled examples. The MEASUREMENTS table as well as the EPISODES table provided us with 43 possible measurements on smoking status, from number of cigarettes per day to "tabacco addiction" as a long-term illness. We decided to take P1739, because this variable is in-use (unlike some other, old-fashioned variables such as "pipe smoking"), and also has three classes (ex-smoker, non-smoker, and smoker) much like the distinctions between classes made in earlier NLP work. This allows us to easier compare our approach and results to these earlier approaches as well.

In the MEASUREMENTS table, variables are coded in three columns: a column indicated the name of the variable, a binary column indicating whether or not the variable is filled, and a third column with the value, as well as the date when this measurement was recorded. We wrote a filter that kept each value

in a row where the P1739 column was indicated as filled-in, and connected this row to the correct patient ID.

We combined data from all 3 files (PATIENTS, CONSULTATIONS and MEASUREMENTS) for each GP office. [3] We combine all patients in the COPD, diabetes, heart disease, or other Ketenzorg programmes from the MEASUREMENTS file in one binary variable called 'Ketenzorg', and then also add a variable which added whether the GP was the main healthcare provider, since data experts told us some GPs did not use the Ketenzorg variable but the 'GP as main healthcare provider' variable to identify Ketenzorg patients. We thus use both variables in our final dataset to identify ketenzorg patients.

All merging of these tables and combining of variables is done based on date and patient id, ensuring the variables are connected to the correct patient and consultation. We connected this table to the earlier file that combined the patient and consultation data.

This means the output of this step is one file per GP office (meaning: 6 data files in total), with each patient having information on consultations consisting of dates, doctor's notes, smoking status, and long-term indicators of chronic illness or tabacco addiction.

3.3.4 Step 4: combining GPs and filtering out minors

After having done these transformations for all GP tables, we then combine all the six GP tables into one large datafile.

Our next step then consists of filtering out all minors, in two ways: every patient who was less than 18 years old on the date of query extraction, and any consultation where the patient was less than 18 years old. This had two reasons: we consider it ethically dubious to further process sensitive textdata on children, and minors usually do not usually smoke - so their data is less relevant to our research question.

The output of this step is one file, consisting of all 6 GPs offices, with per unique EMR all information found in the previous steps.

3.3.5 Step 5: Train/Test/Dev split

We then split the dataset in a 80% train, 20% test/dev set. We use the training set to later train the machine learning models, while we use the development set to fine-tune and develop the rule-based model and the SNORKEL LabelModel.

[3] We exclude the EPISODES file, as this file only had the P17 long-term smoking status we decided not to use as labels for our classification task.

The test set is used to compare the performance of all smoking status classification algorithms.

We ensured there were different and unique patients in the train, test, and develop set, in order to ensure there was no contamination that would lead to an algorithm having already encountered a testing example in its training phase.

3.3.6 Step 6: split the subset in ketenzorg/non-ketenzorg and labelled/non-labelled

We then split the train, test, and development set into a subset that is labelled with P1739, and a subset that is not. We also split these sets into a subset that is ketenzorg, and a subset that is not. We do this so we can easily use the labelled examples for supervised training, and can compare the performance of ketenzorg versus non-ketenzorg.

3.3.7 Step 7: only keep last labelled smoking status conversation from each unique patient

Lastly, we were required to normalize the labelled smoking status EMRs. The mean number of consultations per patient in our final preprocessed dataset is around 4 (M = 3.7), but with large spread (SD = 2)

Some patients had several consultations labelled with a smoking status, and some only one. However, the latest labelled consultation is the most relevant one: we thus decide to normalize data length for all EMRs by only allowing one consultation with smoking status per EMR as example, and this the latest one for current smoking status. One consultation has a mean of 36 words, but with large spread (SD = 42), and a lower median (23 words).

We also filter out some datapoints at this point from the labelled dataset. This is the around 1% of datapoints having a nonsensical or unofficial P1739 label not mentioned in the codebook, such as "0" or "x". These are not values used in P1739, and cannot be used in further classification experiments.

Our final dataset after this procedure consisted of EMR representations from 17.236 unique patients. The preprocessing procedure is visually displayed in figure 2 below, where the "episodes" table - also one of the four tables in the dataset - is not further used in the preprocessing procedure and is thus shown as grey and without connection to the final dataset. It consists of all previously described steps, such as combining the GP ID with the patient ID to obtain a unique identifier to the connection of the SOEP consultation text to one patient and date.

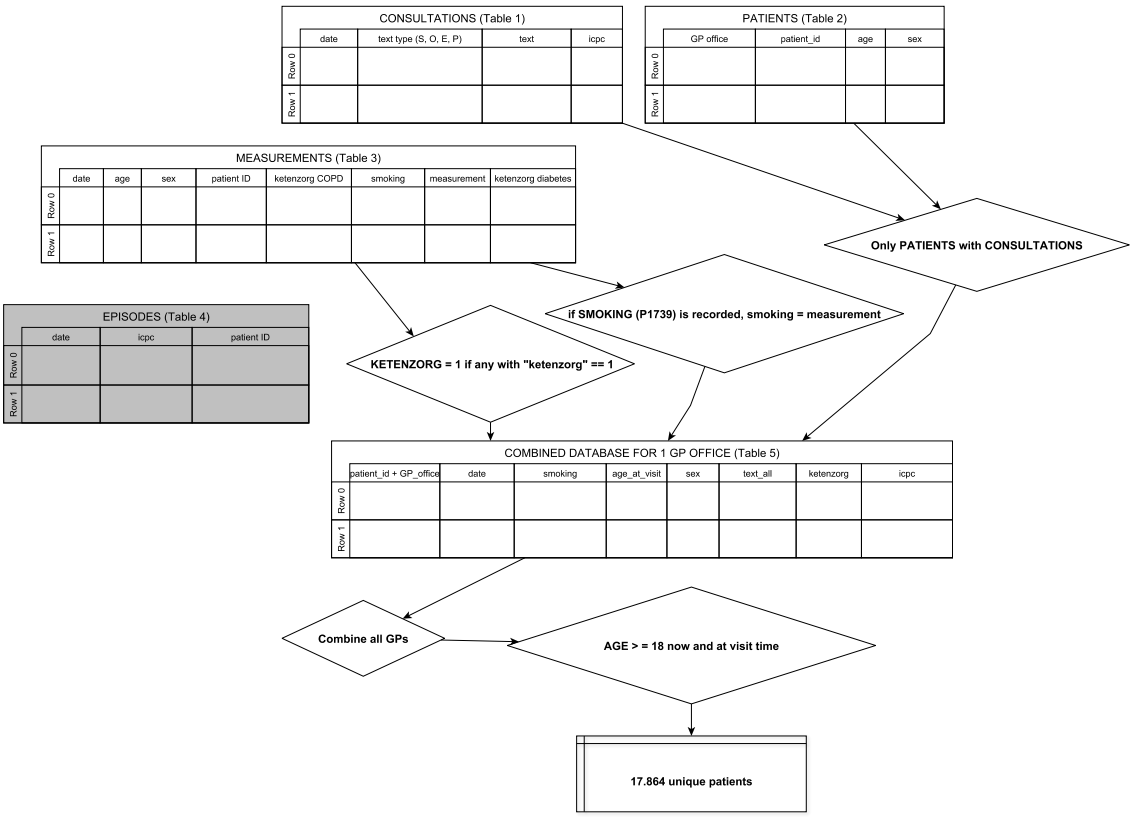
Table 4 shows two fictional examples of the final data representation. The

first column shows the unique identifier made from the patient ID and GP ID, while other variables are connected to this variable. The columns used in the further experiments are the SOEP text, used for the classification experiments as examples to classify, and the smoking variable 1739, which is used as labels for training models to classify smoking status.

Table 4: Two rows showing fictional examples of our final dataset, visualizing how all information was structured per row. As shown here, each unique consultation for each unique patient was one row in our dataset.

patient ID_GP ID	Sex	Age at consult	Age in 2020	SOEP text	date	smoking (1739)	Ketenzorg
9999_777	F	40	43	Mevrouw heeft buikpijn Translation: Mrs. has stomach pain	23-04-2017	4	0
8888_666	M	63	62	Is gestopt met pasta eten, is afgevallen. Has stopped eating pasta, has lost weight	05-07-2019	1	1

Figure 2: This figure displays the full preprocessing pipeline in order to get to the final dataset used for further experiments. For each of the 6 GP offices, we have 4 data files (CONSULTATIONS (Table 1), PATIENTS (Table 2), MEASUREMENTS (Table 3),and EPISODES (Table 4). First, we combine all documents coming from one GP (Table 5), making a new identifying variable consisting of patient ID and GP office id. We then combined these 4 GP databases into one database, and then filtered out any underage patients and any patient underage at the time of their GP consultation.



3.3.8 Text Transformation: BERT and BERT finetuning

One of the latest great development in current NLP is the use of large-scale, pre-trained language models, such as BERT (Devlin et al., 2019). The BERT-based language model we use is the Dutch BERTje (Vries et al., 2019). It is trained on 12GB of high-quality, formal text data (capturing around 12.4 billion tokens) from wikipedia and linguistic corpora like SoNaR. Their pre-training procedure further incorporated the SentencePiece tool, allowing for sub-word embedding as well. We use their cased model (meaning uppercase words retained their upper- and lower casing in training) for text transformation in our classification experiments. A possible downside to this choice is that BERTje is trained on general-purpose language, not on specific medical language. With finetuning, we can teach BERTje specific vocabulary for our task, in our case: Dutch medical data, and smoking status classification.

One particular feature of BERT-based models is that these can only process short texts (up to 152 tokens). Such a tensor-based model also needs all of its input to be equal length, we decide to only take the 152 first tokens of each set of each smoking status labelled consultation. If our text is shorter than 152 tokens, we add empty [PAD] padding tokens. [4]

Our representation of an EMR for smoking status classification is thus the latest labelled smoking status consultations of an individual patient, a binary variable indicating whether this EMR consists of data from a patient belonging to the chronic illness group (KetenZorg), as well as a GP-annotated variable P1739 with 3 possible values: smoker, non-smoker, and ex-smoker.

All splits are stratified by GP office, ensuring that one large GP office does not dominate the training set and thus leads to a biased model, or GP-specific writing harms the generalizability of our models.

The development set is used to fine-tune a baseline, which is a rule-based model based on earlier successful rule-based models for smoking status classification using regular expressions (Wang et al., 2019). We also use the development set to fine-tune the BERT-based models and the SNORKEL LabelModel. The test set is held out until we have built our final models, after which we compare several models in performance on prediction of this test set.

The distribution across labels is visible in Table 6. Several things are notable. First of all, we see that the value "never smoked" is most used. In Figure 3, we also see that there are an (admittedly very small) number of incorrect val-

[4] We also looked at Sun et al. (2019), who found that for text classification, it worked best to take a combination of the first 128 and last 382 tokens. We also explore this option, but find the consultation actually often already reaches the maximum number of 152 tokens, with a median length of 40 tokens per consultation.

Table 5: Table showing the size of training, test, and development sets. Note that this table does not show the distribution for only the labelled datapoints, but shows all datapoints: both labelled and unlabelled EMR examples.

	Training set	**Development**	**Test**
EMR representations	14.298	1.788	1.787

ues. That is, values the GP or assistant put into the system that are impossible in the dictionary of values for the smoking variable 1739 as presented by the National GP Society (NHG). In total, this is around 1% of the training and test set, meaning a handful of examples (up to 88 in the training set) are labelled with such a value, and need to be deleted. Most often the value '0' is used, while this is not a value represented in the dictionary of values. These examples are thus deleted from the labelled examples.

Table 6: The **labelled** datapoints in values used for the smoking status variable '1739' ("smoking"), as defined by the NHG (National GP Association)

	Training	**Dev**	**Test**
"smoker"	794	115	103
"never smoked"	2081	268	274
"ex-smoker"	2103	268	251
total labelled EMR representations	4.978	651	628

3.4 Research Process

Our experimental set-up is as follows: we start with supervised learning, where we only consider the labelled data. We then explore SNORKEL as a means to increase the training size with weak supervision. We compare these approaches to a rule-based baseline. We visually describe the research process and set-up of our experiments in Figure 4. More information on the exact method of each of these approaches can be found in Chapter 4 below.

Figure 3: Our train/development/test split of the pre-processed dataset, also showing how we split these sets in labelled and unlabelled EMRs, and ketenzorg and non-ketenzorg EMRs.

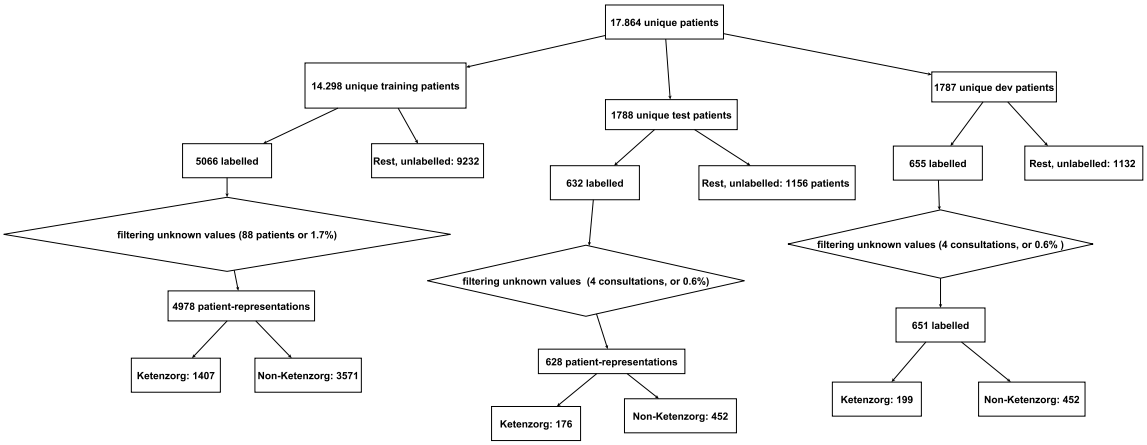
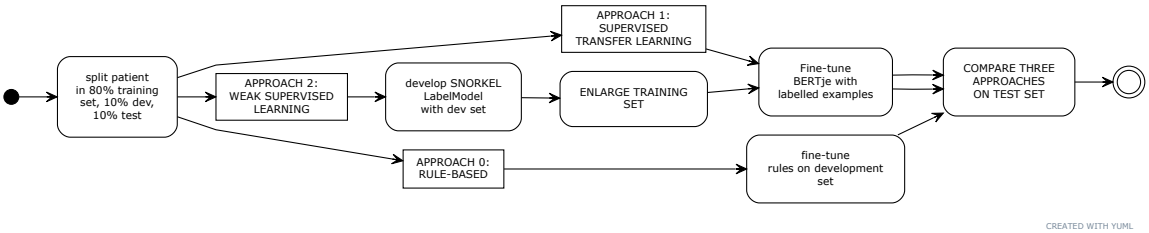


Figure 4: Our research process, described in a process chart. We start with the above-mentioned train/dev/test split, and then explore three approaches to smoking status classification: rule-based classification, supervised learning, and weakly supervised learning.



4 Method

Our experiments come in two distinct parts with two different methods: **supervised learning** and **weakly supervised learning**. We describe each of these approaches in their own subsections in the methods and results chapters. Subsections 4.1 and 5.1 are about the methods and the results of supervised learning with BERTje. Subsections 4.2 and 5.2 describe the methods and the results of the weakly supervised paradigm with SNORKEL and BERTje.

4.1 Method Supervised Learning

Our first experiments are **supervised** smoking status classification, thus we are limited to the previously labelled datapoints. We are dealing with N = 4.978 labelled EMR consultations in the training set (out of 14.298 training examples in total). In the development set of 1.788 EMR representations, there are 651 labelled EMR representations. The test set has 628 labelled EMR consultations.

4.1.1 Baseline: Keyword/Regex-Based Algorithms

Our baseline consisted of a rule-based classifier based on RegExes from Wang et al. (2019) and Palmer et al. (2019), which obtained F-scores over > .90 in their study. Other studies also obtained similarly high performance with similar approaches on smoking status classification (Weng et al., 2017).

We develop and test three different versions of such a rule-based baseline: two directly taken from Wang et al. (2019), and one designed ourselves from earlier literature as well as our own additions. This last algorithm was fine-tuned and developed with the labelled development set. The keywords were identified with data experts on the dataset, and were also obtained from the Zorgstandaard (Partnership Stop met Roken, 2019) - in this latter category were for instance medicine names for smoking cessation as used in the Netherlands by GPs.

We translated the regular expressions from Wang et al. (2019) from English into Dutch. There were two datasets mentioned in this paper, with distinct regular expression algorithms: one developed on the Mayo Clinic dataset, and one developed on the ib2b shared task dataset. Their smoking status classification on the Mayo Clinic dataset was a binary classification, with only smoker and never-smoker as classes, while their classification on the ib2b shared task dataset had three classes: current smoker, ex-smoker, and non-smoker. We use both to test on our test set. These regular expression rules for the classification

of smoking status from Wang et al. (2019) are given in Appendix A. Some of the regular expressions lose their meaning in translation (e.g. ENG 'cig' -> Dutch 'sig'), so are removed, while others (different words for "quitting smoking", e.g. "ceasing" or "discontinuing") are not Dutch collocations so are also removed.

Our own rule-based algorithm works as follows, and is depicted in pseudo-code in Table 7. It first sets the label to -1 (abstain, or unknown), and then identifies whether there is evidence of a positive smoking status, by matching words such as "roken" (smoking) and "roker" (*smoker*). Then, it starts looking for evidence of ex-smoking status, since ex-smoking status is usually identified by smoking status signifiers (e.g. the keyword "roken" (*smoking*)) with an added negation ("niet" or "geen" (*not*)) or added information ("gestopt" (*quit*)). If such a construction is found, the label is updated from "smoker" to "non-smoker". Then, the algorithm checks whether constructions that signify cessation of smoking ("rookt niet meer" (*does not smoke anymore*), "gestopt met roken" (*quit smoking*)) are in the free text, and if these are there, updates the label to "ex-smoker". In this way, the label is iteratively updated based on the information in the text.

Table 7: The regular expression algorithms used to detect different smoking status classes, combining both the regular expressions by Wang et. al. 2019, and terms from the Care Standard on Tobacco Addiction 2019.

smokeRegex = '((e(-)?sig(aret)?) | pakje | pakjes | nicotineverslaving | tabaksverslaving | (stoppen.*?met.*?roken)
| (heeft | doet | blijft).*?((ge)?rookt | roken) | (gebruikt.*?tabak) | (huidig | momenteel | nu | nog) .*?(roker | roken) | tabak(s)?
gebruik*(ja | nog steeds | rookt.*?nog | rookt) |
(roker | rookt? | rookte | roken | rokers?) | tabak | sigaret(ten)? | sig? | pijp?
| nicotine | siga(a)?r(en)?)'

nonSmokeRegex = '((heeft.*?nooit.*?gerookt | rookt.*?niet | geen.*?roker | is geen roker) | ((niet | non | geen | nooit | negatief).*?(roker | roken | rookte | tabak))
| niet(-)?roker | ontkent.*?roken |
(tabak | rook | roken | nicotine)*?
(nooit | niet) | rookt.*?niet | (0 | nul).*?rokers?)'

exSmokeRegex = '((was.*?roker | gestopt | is.*?gestopt.*?met.*?roken | rookt.*?niet.*?meer)
| ((is gestopt | stopte | hield.*?op.*?met | hield.*?op).*?
(tabak | roken)) | (vorige | vroeger | verleden | ooit | voormalig |
(ex- | ex).*?(tabak | roker | stop(te)?.*?(met)?).*?(roken | tabak(s)?))
.*?gebruik.*?*?(rookte | stopte) | roken*(gebruikte | vroeger))'

```
for EMR in EMRs:
    label = -1
    if smokeRegex in EMR:
        label = SMOKER
    if nonSmokeRegex in EMR:
        label = NEVER
    if exSmokeRegex in EMR:
        label = EX
```

4.1.2 Classification Model: Fine-tuned BERTje

Fine-tuning a pre-trained Transformer effectively means adding layers on top of the previously trained layers, and adjusting the weights in the layers to the new task or domain (Devlin et al., 2019). We use BERTje (Vries et al., 2019), trained on Dutch text and consisting of 12 layers previously trained in language understanding by next sentence prediction and word/context prediction.

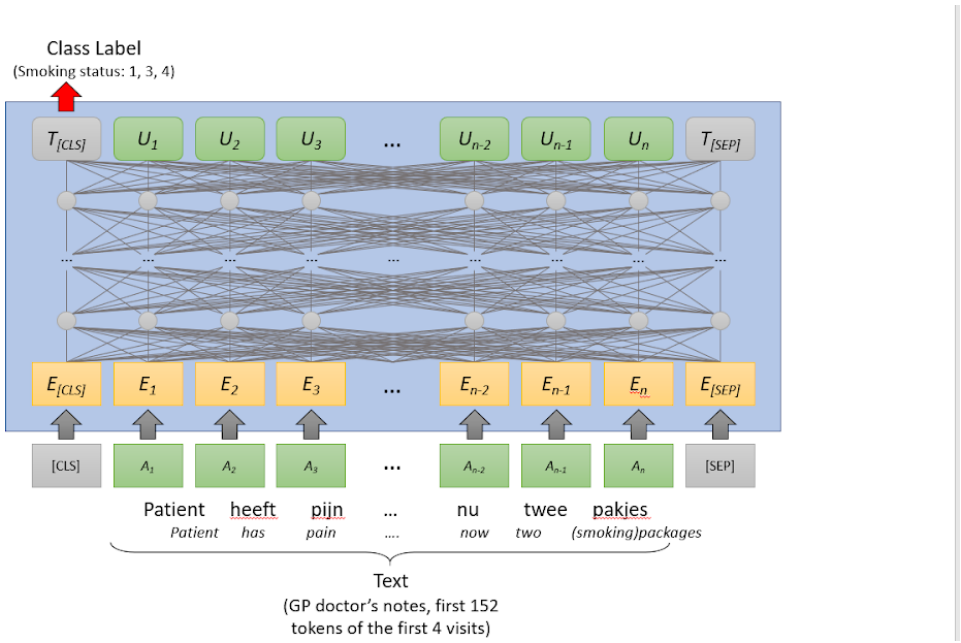
We train BERTje on the task of smoking class prediction by adding an additional untrained classification layer, and training with our own dataset in several additional epochs. We follow most of the code in a tutorial by Chris McCormick for our fine-tuning (McCormick & Ryan, 2019). In Figure 5 as adapted from the basic BERT schema made by Lin (2020), we see schematically how BERTje works, with an input sequence going through twelve Transformer layers and outputting a class. [5] The Transformer model consists of an input of

[5]The full architecture, with specification of all layers, is shown in Appendix B.

tokenized texts separated by specific tokens signifying the start ([CLS]) and the end ([SEP]) of a sequence. These token representations are then fed through BERTje's 12 pre-trained layers, after which a 13th added linear layer is tasked with providing one of three smoking status labels: SMOKER, EX-SMOKER, or NON-SMOKER.

We trained in small batches of 5 instances due to memory constraints. We also tested batches of 3, but this led to worse performance on the validation set. Our best performing settings on the development set are a learning rate of 0.00005 with 2 epochs (more led to overfitting on the training set, with higher training loss than validation loss) and a batch size of 5.

Figure 5: A depiction of BERTje's architecture as well as the input and output for the smoking classification task. Image is based on figures in the BERT paper (Devlin et. al., 2018) and adapted by Jimmy Lin for public use (Lin, 2020)



4.2 Method Weak Supervision

We are interested in seeing whether a larger training set, with additional examples labelled with a weakly supervised labelling model, is able to improve performance over a model only trained with hand-labelled datapoints.

We increase our training set size with SNORKEL (Ratner et al., 2017) and the Data Programming Paradigm (Ratner et al., 2016). This means we train a LabelModel to learn the optimal weighting of heuristic rules designed to label the unlabelled training examples. This generative labelling of examples works with the idea that each LF is an unique Labeller, which allows for more reliable labelling than with a simple rule-based labelling method. The SNORKEL method is also able to deal with noise by learning which rules are reliable. With the additionally labelled examples by the SNORKEL LabelModel, we hope to fine-tune a better performing BERTje model.

4.2.1 Labelling Model: SNORKEL

We start our experiments with a pre-defined set of 32 labelling functions (LFs) or heuristics. Of the 32 LFs, 4 are regular expressions and 28 are keyword scanners. 8 of these keyword scanners predict the ex-smoker class, 19 predict the smoker class, and the other 5 predict non-smokers. Rules predicting smoking (the positive class) are more common because positive cases have more signifiers. For our SNORKEL development, we initially found 6 of our 32 rules lacked coverage. Among these were 2 of the 3 mentioned medicines for quitting smoking in the 2019 Care Standard for the Tobacco Addiction (Partnership Stop met Roken, 2019). A full list of the 32 LabelFunctions can be seen in Appendix D. Many of these LFs are similar to the keywords and regular expressions used in the rule-based algorithm described above, with the important distinction that SNORKEL's LabelModel can weigh different rules in order to give highest weighing to the most reliable LabelFunctions - which is something our simple baseline cannot do.

Our process is one of iteratively tuning and improving these labelling functions to determine (1) how effective they are at capturing the three smoking status classes (EX, SMOKER, NON-SMOKER), and (2) how they perform in a generative SNORKEL (Ratner et al., 2017) model.

Our optimization process is shown in figure 6 below. Optimization shows these steps, where iteratively the LFs performing below a certain accuracy threshold are removed, after which the LabelModel is re-trained and again tested on the development set.

We iteratively remove LFs with low performance, and then train a SNORKEL

Labelling Model with a learning rate of 0.005 learning rate and 500 epochs. The specific results of our process, with the accuracy thresholds and the LFs left after each iteration, are shown in Figure 7. We first train a Label Model on the labelled training set with all Labelling Functions with coverage, then train a Label Model with all LFs with an accuracy above .20 (21 LFs) on the 1407 labelled training examples, then all LFs with an accuracy above .40 on the labelled training set, then all LFs with an accuracy above .70 and later an accuracy above .80 on the labelled training set. After each step, we validate the model on the held-out development set. We then choose our best performing model on the development set to label the unlabelled datapoints in the test set, which turns out to be the latest LabelModel with only the 6 LFs with > .80 accuracy.

Figure 6: A step-wise depiction of the process of our fine-tuning of SNORKEL, showing how the process is iterative with the arrow showing the loop from step 7 to step 4 and back.

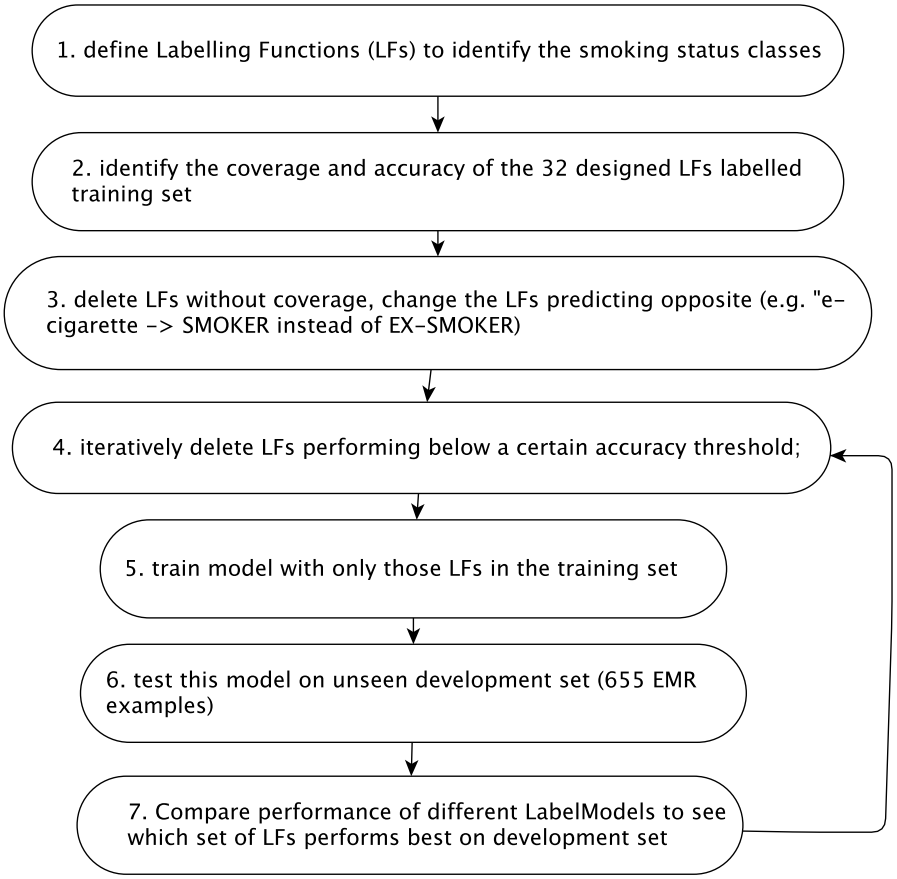


Figure 7: Our label model fine-tuning process with specific accuracy thresholds and LFs during the process, visualizing how we eventually arrived at a LabelModel with 6 LFs.

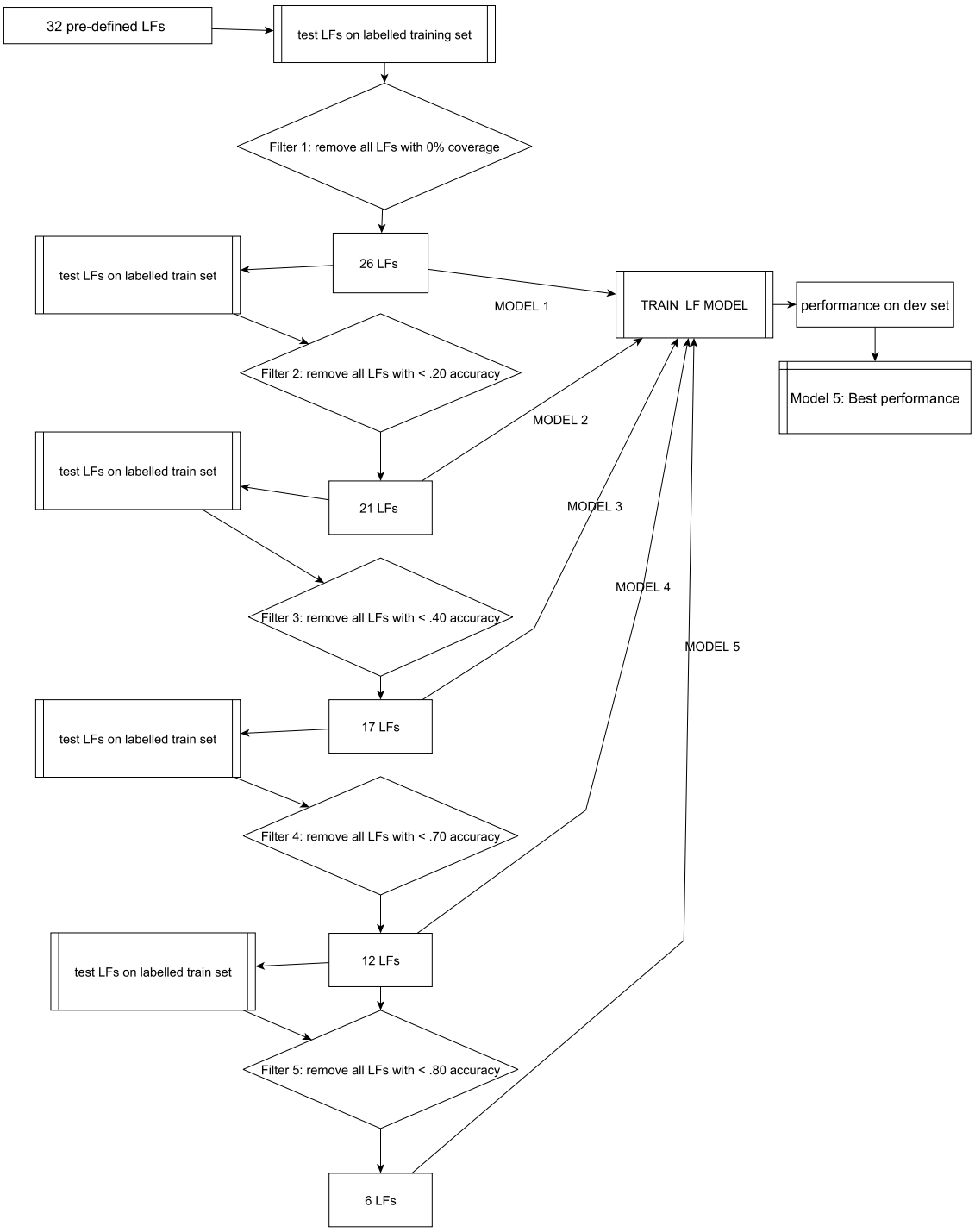


Table 8: Performance of trained LabelModels on development set

LabelModel	nr LFs	nr examples labelled in unlabelled training set	accuracy	F1 micro	F1 macro
all LFs with coverage	26	5.061	.33	0.33	.31
> .20 accurate LFs	21	5.061	.33	.33	.31
> .40 accurate LFs	17	1.572	.35	.35	.35
> .70 accurate LFs	12	527	.37	.37	.34
> .80 accurate LFs	**6**	**512**	**.38**	**.38**	**.37**

This final trained LabelModel can moderately accurately label an additional 512 examples. We attempt to label as much of the the 9.232 unlabeled EMRs in the training set as possible, without losing accuracy. We observed that relative noisy LFs do not contribute to the overall performance.

We saw some peculiar behaviour of the LFs: some of them seem to not behave like the Care Standard (Partnership Stop met Roken, 2019) requires. For instance, the keyword scanning rule "e-sigaret" (e-cigarette) is more often connected to the SMOKING label than to the NON SMOKER label, while according to regulations someone using an e-cigarette is a non-smoker. Perhaps e-cigarettes are being used by current smokers. The word "roken" (smoking) also seems to occur more with ex-smokers than smokers.

We found model 5 the best performing - the one with only 6 LFs, but 6 that are highly accurate. Even with 6 highly accurate LFs, the model fails accurately labelling many examples from the development set.

The results in accuracy are seen below in Table 21. The accuracy is below .50, even with many individual rules above .60, and the best LabelModel is not greatly accurate at labelling examples (an accuracy of .38), and also cannot label many additional examples (512). However, such a small increase already means 1/3 more data, as the originally labelled training set only had 1.407 examples.

Our second phase consisted of fine-tuning the hyperparameters of the LabelModel (learning rate and epochs) to see if we could improve over this latest score. We see a very slight increase with a lower learning rate, from an accuracy of .38 to an accuracy of .39. The number of additionally labelled examples does not increase, and remains 512.

When we increase epochs, we effectively give the model more opportunity to fine-tune on the training set because epochs are how often the full dataset passes through the entire network, allowing for more opportunity to optimize the weights. As seen in Table 10, we see absolutely no difference in performance with a higher number of epochs, unlike the (slight) difference of performance with different learning rates. However, the number of additionally labelled

Table 9: Performance in accuracy (right-most column) and number of additional examples labelled (middle column) by the LabelModel on the development set when increasing the learning rate (left-most column)

learning rate	examples labelled	accuracy
0.005	512	.392
0.0005	512	.393
0.00005	512	.393

Table 10: Performance of different numbers of epochs on the LabelModel, with a development set with learning rate = 0.00005.

epochs	examples labelled	accuracy
250	512	.393
500	512	.393
1.000	512	393

examples also does not increase when increasing the number of epochs. Neither does it improve when reducing the number of epochs, which would effectively prevent overfitting if this would have been applicable.

4.2.2 Classification Model: BERTje

We then fine-tune BERTje (Vries et al., 2019) with the labelled data combined with the data additionally labelled by the best performing Labelling Model, in the same manner as we did with the supervised BERTje model in Section 4.1.2, but in this case use the labelled training data together with additionally labelled datapoints as training set.

4.3 Evaluation

We evaluate each classification algorithm's performance on the 628 labelled unseen test examples with the evaluation metrics precision, recall, and F1-score. Precision shows the proportion of predictions of one class are actually a member of that class. The recall score shows how many existing members of a class are found by the algorithm. The F1 score is the harmonic mean of precision and recall, as seen in Equation 1 below.

$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{1}$$

Where $0 \leq \text{F1} \leq 1$

and F1 = 1 $\Leftrightarrow precision = recall = 1$.

Precision and recall are defined as below, in Equation 2 and 3:

$$precision = \frac{TP}{TP + FP} \tag{2}$$

$$recall = \frac{TP}{TP + FN} \tag{3}$$

Where TP = True Positives (predicted the correct class)
FP = False Positives (incorrectly predicted member of a class)
FN = False Negative (incorrectly predicted not member of a class)

We use micro-F1 score to calculate the F1-statistic over all classes. That is, we do not average the F1 statistic over the three distinct classes (smoker, ex-smoker, non-smoker) but rather take all false positives, true positives, and false negatives of all classes together to calculate the overall score on the test set. This ensures classes with less examples, such as ex-smokers of which there are simply less than non-smokers, get weighted according to the number of examples and the classes are not weighted equally - which would lead to more influence for the less populous classes than their size warrants. [6]

We compare the performance of several algorithms: the rule-based models, the supervised model, and the weakly supervised model, on the same labelled test set of 628 labelled EMR conversations.

4.4 Software and Packages

The preprocessing as well as the data analysis was done with Python 3.6 with the PyCharm Professional development environment. For our preprocessing we use the pandas package 0.25.3 (The Pandas Development Team, 2020) as well as Scikit Learn (Pedregosa et al., 2011) for evaluation.

Loading, training and testing our BERTje classification models is done with the Transformers python package (Wolf et al., 2019) version 3.0.2 with PyToch 1.5.1 and NumPy 1.18.4. For our classification experiments, we mostly use the code shared in a BERT tutorial (McCormick & Ryan, 2019). For our weak supervision experiments, we use the SNORKEL python package version 0.9.5 (Ratner et al., 2017).

[6] In Appendix C are the results also included macro and weighted performance measures on the test set.

All software was installed on a remote linux server where the data was also stored, and our scripts were run there without us downloading the data to a local environment or directly accessing the data in any way.

5 Results

5.1 Supervised Learning

5.1.1 Baseline: Keyword/Regex-Based Algorithm

We compare the performance of the several rule-based models we developed: the translated one from Wang et al. (2019) as well as an algorithm we developed with our development set consisting of both phrases in these earlier works as well as rules and words from the the Care Standard (Partnership Stop met Roken, 2019). We test each algorithm on the unseen test set of 628 labelled EMR examples.

All results can be seen in table 20 below, which shows the performance of each algorithm (columns) on precision, recall, and F1-score (rows). In our baseline performance, we found that the translated Mayo Clinic algorithm performed best on overall recall (.44) and F1 (.65), however, this algorithm only detected 2 classes (smoker and never smoker), and did not take into account the EX-SMOKER class. [7]

When we compare our algorithm to the other algorithm classifying three classes, the ib2b algorithm, we see a marked improved of over .10 points in precision, recall, and F1 score. The ib2b algorithm has .43 on precision, 0.41 on recall, and .30 on F1, while our developed algorigm has a precision of .54, a recall of .43, and an F1 of .49.

When looking at in-class performance in Table 12, we see our own algorithm performs best at classifying EMRs that are ex-smoking and non-smoking, but is compared to the mayo algorithm (which, once again, is a binary algorithm - only detecting non-smoking versus smoking - performing badly on especially recall and F1. Our algorithm however scores higher on precision (0.24) than the Mayo clinic one (0.23) for the SMOKING class. All in all, our own algorithm - consisting of a combination of the two others and our own keywords such as 'tabaksverslaving' (tabacco addiction) performed especially well for detecting non-smokers (F1 = 0.63), but also drastically out-performed the translated ib2b algorithm for ex-smokers (F1 = 0.24 versus F1 = 0.02 for the ib2b algorithm).

[7] Additional results also showing weighted and macro precision, recall, and F1 score can be found in Appendix C.

Table 11: Performance on the test set for all of the smoking status classes for models trained with the labelled development set.
NOTE; the mayo clinic algorithm only has 2 classes.

	algorithm ib2b	algorithm Mayo clinic	own algorithm
precision (micro)	0.39	0.44	**0.49**
recall (micro)	0.36	**0.65**	0.43
F1 (micro)	0.38	0.52	**0.55**

Table 12: In-class performance on the test set for all of the smoking status classes for models trained with the labelled development set.

	algorithm ib2b			algorithm Mayo clinic			own algorithm		
	precision	recall	F1	precision	recall	F1	precision	recall	F1
SMOKING	0.20	0.23	0.22	0.23	**0.32**	**0.27**	**0.24**	0.30	0.26
NON-SMOKING	0.46	**0.85**	0.59	0.51	0.77	0.62	**0.55**	0.74	**0.63**
EX-SMOKING	0.50	0.01	0.02	-	-	-	**0.65**	0.15	**0.24**

5.1.2 Classification Model: Fine-tuned BERTje

We then looked at the performance of the BERTje fine-tuned classification model on the test set, we see an improved performance over the rule-based baseline, with a weighted F1-score, recall, and precision all at .79 or .80.

When we look at in-class results, we see BERTje is significantly better at precision for the smoking class, while recall for the non-smoking class is higher. [8]

5.1.3 Analyzing Confusion matrices

In order to fully understand how the classification models work, we also analyze confusion matrices. These enable us to see when the models confused one class for the other.

In Figure 8, we see the absolute numbers of this confusion matrix. The test set had 628 EMRs to predict, of which 103 had a SMOKER smoking status, 251 had an EX-SMOKER status and 274 a NEVER smoking status. Of the 103 SMOKING status, 69 were correctly predicted - an in-class accuracy of 66%.

[8] In Appendix E we have also described the classification experiments with BERTje where we only train on chronically ill patients in the Ketenzorg program. We do see that especially detecting smokers is easier with a BERTje only fine-tuned on KetenZorg EMRs.

Table 13: In-class performance on the test set for all of the smoking status classes for models trained with the labelled development set.

		Rule-Based	BERTje
precision	micro	0.49	**0.79**
recall	micro	0.43	**0.79**
F1	micro	0.55	**0.79**

Table 14: In-class performance on the test set (628 EMRs) for all of the smoking status classes for models trained with the labelled training dataset

	baseline: own algorithm			**BERTje**		
	precision	recall	F1	precision	recall	F1
SMOKING	0.24	0.30	0.26	0.92	0.67	0.78
NON-SMOKING	0.55	0.74	0.63	0.77	0.83	0.80
EX-SMOKING	0.65	0.15	0.24	0.78	0.81	0.79

The other classes, EX-SMOKER and NON-SMOKER, had more correctly classified elements, and less confusion. In Figure 9, we see the in-class **accuracy** of this confusion matrix. Of the 251 EMRs that were EX-SMOKER, 203 were correctly predicted, or an in-class accuracy of 81%. And of the 274 that were NEVER-SMOKER, 227 were correctly predicted, or 83% in-class accuracy. This makes the SMOKER status relatively the least accurately predicted class. However, the most confusion - both absolutely and relatively - does not happen with the SMOKER class, but with the EX-SMOKER and NEVER SMOKER class, where up to 18% of EX-SMOKERS get classified as NEVER SMOKERS, and 16% of NEVER SMOKERS getting classified as EX-SMOKERS.

In Figure 10, we see the relative distribution over the predicted smoking status classes, also known as the in-class **precision**. We see that the majority of predicted labels is predicted correctly: 92% of predicted smokers are actually smokers, 77% of predicted never smokers are actually never smokers and 78% of predicted ex-smokers are truly ex-smokers. The most confusion happens between ex-smokers and smokers: 15% of predicted never-smokers are actually ex-smokers, and 17% of predicted ex-smokers were actually never-smokers.

Figure 8: Confusion Matrix of the **supervised model**'s performance on the test set, showing the distribution of predicted labels (in the vertical columns) and the true labels (horizontal columns) in the test set. There are a total of 103 EMRs with a 'smoker' smoking status, 251 with an 'ex-smoking' status and 274 'non-smoker' status.

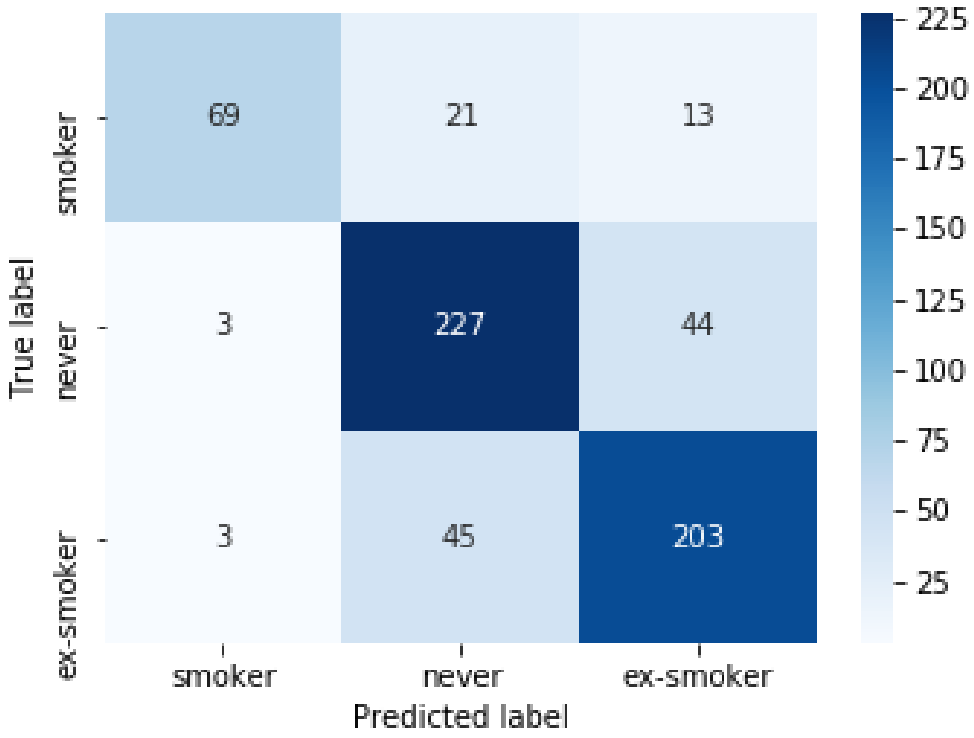


Figure 9: Confusion Matrix of the **supervised model**'s performance on the test set, showing the **relative** (in percentage) distribution over True labels (in the vertical columns) and the true labels (horizontal columns) for the test set (628 EMRs in total). This matrix shows that 67% of all SMOKER smoking status EMRs were correctly predicted, versus 83% and 81% of all NEVER and EX-SMOKING items respectively. We do see that 16% of true NEVER SMOKED items gets classified as an EX-SMOKER, as well as 18% of EX-SMOKER getting classified as NEVER SMOKER.

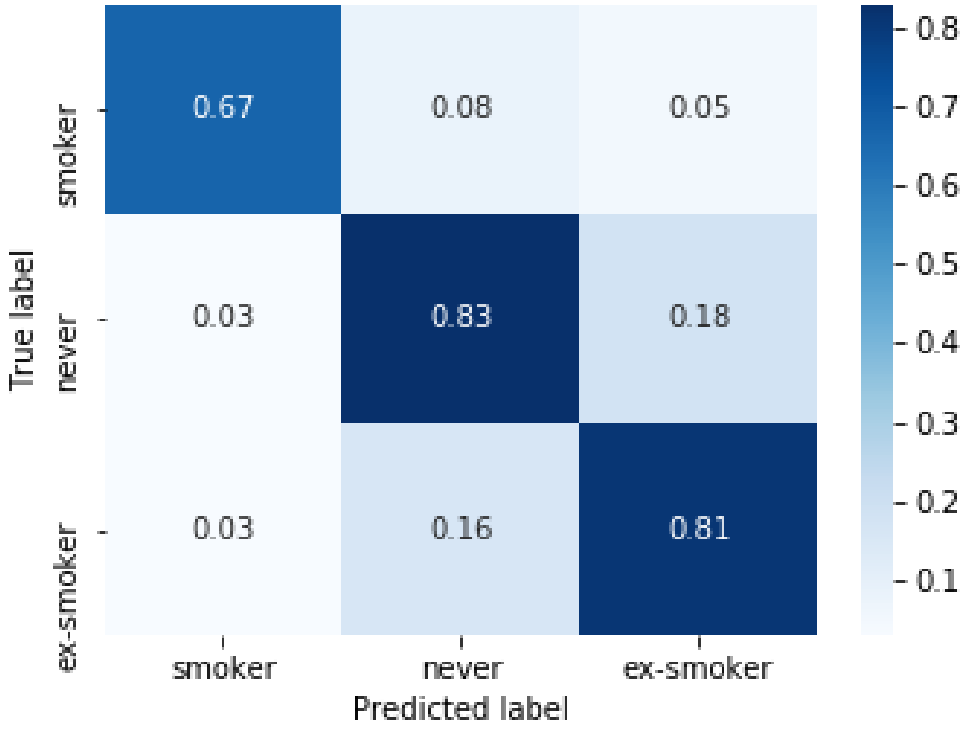
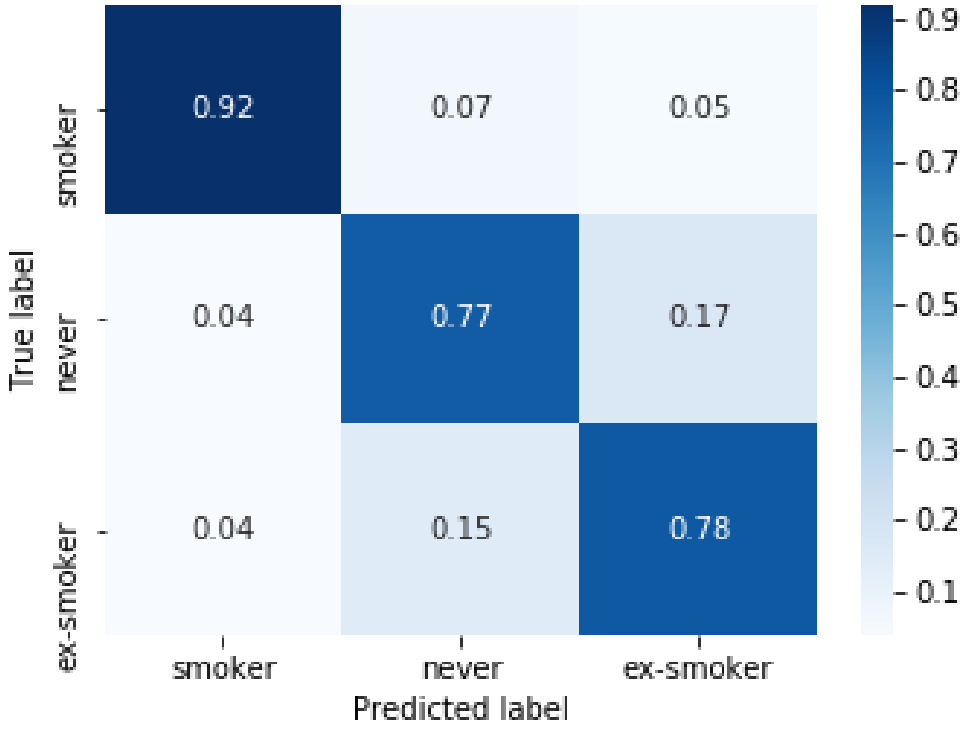


Figure 10: Confusion Matrix of the **supervised model**'s performance on the test set, percentage of predicted labels. In the most left column we see the percentages for smoking prediction: 92% of predicted smokers are actually smokers.



5.2 Weak Supervision

5.2.1 Classification model: BERTje trained with SNORKEL-labelled data.

We then use the LabelModel to label an additional 512 EMRs with smoking status, and then fine-tune BERTje with this new dataset. Then, we evaluate this model's performance against the BERTje fine-tued with the new dataset.

In Table 15, we see that this leads to no discernable performance increase over only BERTje: precision, recall, and F1 is the same at respectively 0.79, 0.79 and .79 for both methods.[9]

However, looking at in-class performance in 16 we see something interesting. The model trained with the 512 additionally labelled SNORKEL examples performance significantly better especially on detection of the NON-SMOKING class, with F1 = 0.81 compared to F1 = .75 with the BERTje trained model on only the hand-labelled examples. There is also some improvement in the SMOKING class (F1 = 0.73 & prec = 0.86 compared to the older F1 = .72 and prec = .86). The EX-SMOKING class, in contrast, does not improve the model with the new labelled trained examples, and even performs worse (F1 = 0.79) than in the earlier trained BERTje (F1 = 0.82).

Table 15: Comparison of performance by BERTje and SNORKEL-trained BERTje on the test set (628 examples)

	BERTje	SNORKEL + BERTje
precision (micro)	0.79	0.79
recall (micro)	0.79	0.79
F1 (micro)	0.79	0.79

[9] In Appendix C, we provide results with also macro and weighted F1, precision, and recall of smoking status classification with this method.

Table 16: In-class performance of predicting the test set (628 examples) by the BERTje model (left side) and the BERTje model with SNORKEL.

	BERTje			SNORKEL+BERTje		
	4.978 training examples			*5.490 training examples*		
	precision	recall	F1	precision	recall	F1
SMOKING	0.82	0.64	0.72	**0.86**	0.64	**0.73**
NON-SMOKING	0.74	0.76	0.75	**0.79**	**0.84**	**0.81**
EX-SMOKING	**0.82**	**0.83**	**0.82**	0.78	0.80	0.79

5.2.2 Analyzing Confusion matrices

We then analyze the confusion matrices to see whether the model confuses the classes for one another. Compared to the confusion matrix of the supervised BERTje model in Figure 8, the model as shown in Figure 11 below identifies less items correctly as SMOKER and EX-SMOKER, but 2 more correctly as NEVER-SMOKER.

In Figure 13, we see a relatively large decrease of accurate prediction, with 83% of all predicted smokers being correctly smokers (versus 92% in the model trained without BERTje), and more confusion as well: 21% of the predicted NEVER smokers are actually SMOKER, compared to only 15% by the model without SNORKEL data as seen in Figure 13

We also see a decline of recall of the individual classes, as seen in Figure 12 compared to the earlier model performance displayed in Figure 9. Now, 64% of all SMOKERs are correctly identified (compared to 67% in the BERTje without SNORKEL data). The only improvement we see is .01% more correctly labelled on the NEVER class (84% correct compared to 83% in Figure 9. We also see a percentage point doubling of confusion between SMOKER and NEVER SMOKER (from 3% of smokers being identified as NEVER smokers to 6%), and also more confusion between EX-SMOKER and SMOKER (8% of EX-SMOKER predictions are actually SMOKERS, compared to 5% in the BERTje model without SNORKEL).

Figure 11: Confusion Matrix of the **weakly supervised model**'s performance on the test set, showing the distribution of predicted labels (in the vertical columns) and the true labels (horizontal columns) in the test set. There are a total of 103 EMRs with a 'smoker' smoking status, 251 with an 'ex-smoking' status and 274 'non-smoker' status.

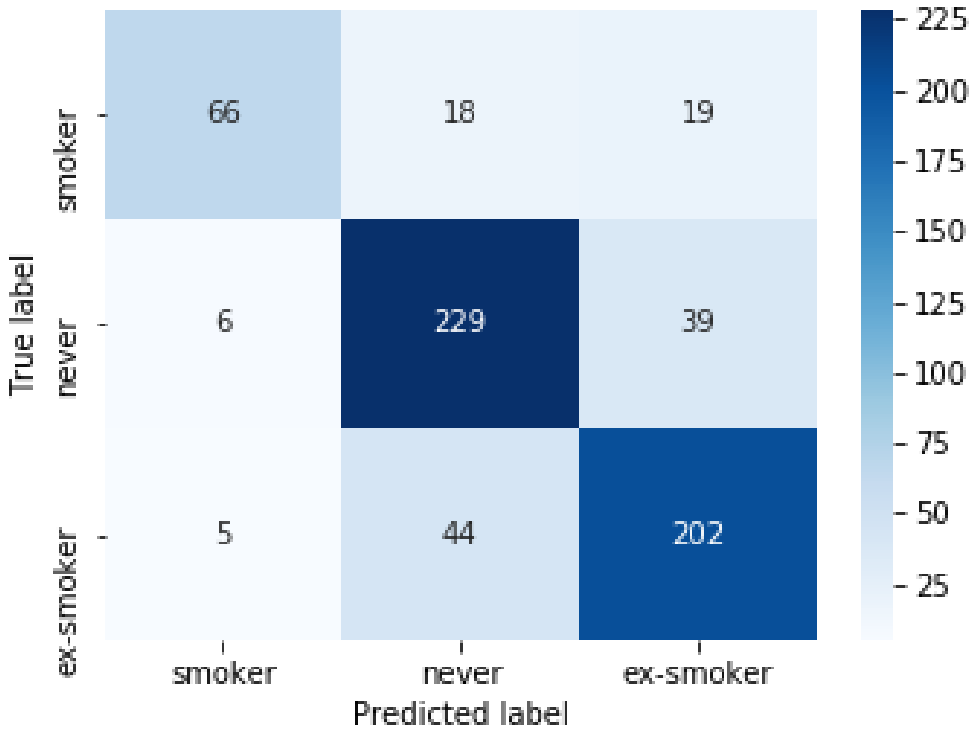


Figure 12: Confusion Matrix of the **weakly supervised model**'s performance on the test set, showing the **relative** (in percentage) distribution over True labels (in the vertical columns) and the true labels (horizontal columns) for the test set (628 EMRs in total). This matrix shows that 64% of all SMOKER smoking status EMRs were correctly predicted, versus 84% and 80% of all NEVER and EX-SMOKING items respectively. We do see that 16% of true NEVER SMOKED items gets classified as an EX-SMOKER, as well as 16% of EX-SMOKER getting classified as NEVER SMOKER.

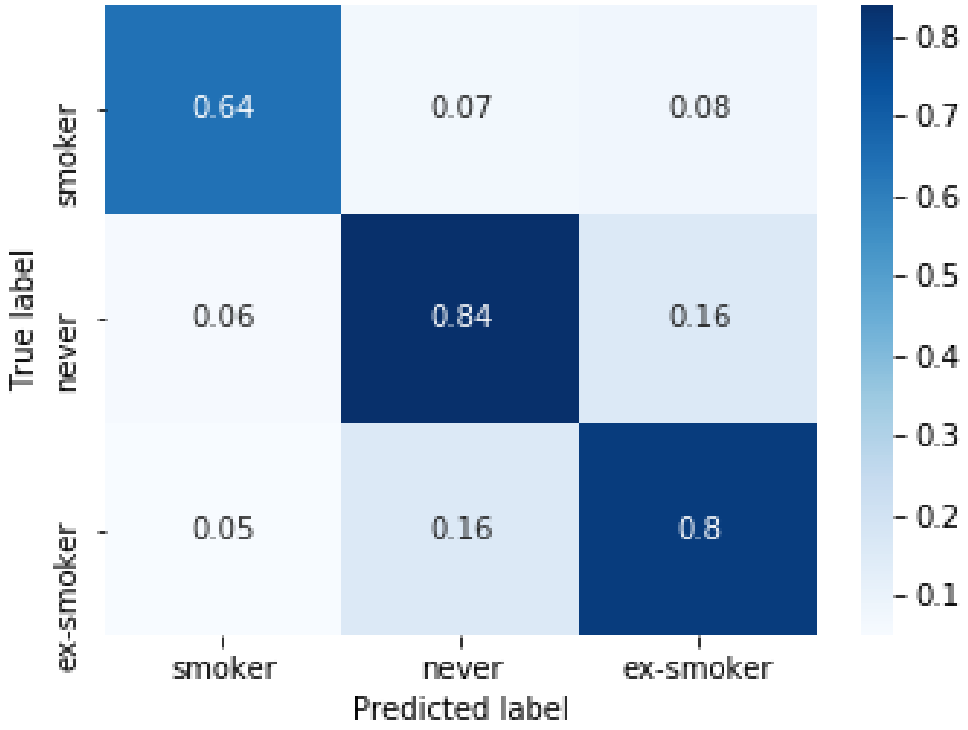
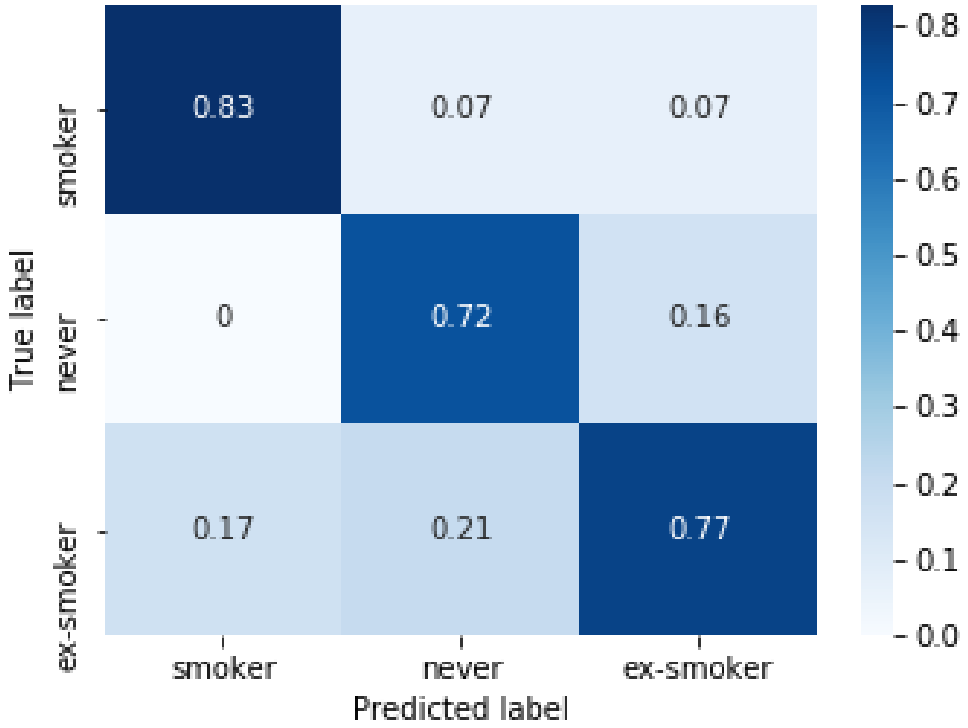


Figure 13: Confusion Matrix of the **weakly supervised model**'s performance on the test set, percentage of predicted labels. In the most left column we see the percentages for smoking prediction: 83% of predicted smokers are actually smokers.



5.2.3 Summary of Results

We found the pre-trained Transformer model BERTje was able to reliably predict smoking status from EMR consultations in the dataset, with an F1 of .79. This is a better performance than our best-performing rule-based method (F1 = 0.49). Working with SNORKEL to label more datapoints did not lead to an improved performance in BERTje (F1 = 0.79), though we did see in-class performance improving especially for the non-smoking class with SNORKEL (from F1 = 0.75 to F1 = 0.81).

In order to allow an easy comparison between models, we show a table here showing all of the model's performance on the same test set of 628 EMRs.

Table 17: Comparison of all models and their performance on the test set (628 EMRs) in precision (top row), recall (middle row) and F1-score (bottom row). We see training with additionally labelled examples by the SNORKEL LabelModel makes little to no difference in performance, while both improve performance over the rule-based method.

	Rule-Based	BERTje	SNORKEL + BERTje
precision (micro)	0.49	**0.79**	**0.79**
recall (micro)	0.43	**0.79**	**0.79**
F1 (micro)	0.55	**0.79**	**0.79**

6 Discussion

6.1 The Fuzziness of Classes

Notable in the preprocessing phase was the simple rarity of smoking status documentation. Less than 1% of all EMRs had a registered smoking status. Also notable was that there were values in the smoking status P1739 that were not in the official GP codebook, *the "Tabel Diagnostische Bepalingen", version 33*. This shows not only such a real-life dataset is more noisy and messy than the datasets provided in the shared tasks on smoking status classification in for instance (Wang et al., 2019; Weng et al., 2017), but also shows the inherent fuzziness and ambiguity in clinical or medical categorization identified by Bowker and Star (2000). One such an invalid entry used for smoking status P1739 was "0", which is not one of the documented and official values of the 1739 smoking variable. One could speculate about what 0 means - does it mean the smoking status is unknown? That it is negative, i.e. a non-smoker? Or that it is not relevant? We simply do not know. Fact is that such labels cannot be used in training for the classification model, because the machine learning approach is only based on examples from our established classes - not for these un-defined cases.

Another potential issue to this classification problem is that even the categorizations made within the validated codebook are not always clearly defined, and can be fuzzy and unclear in their definitions and boundaries. This seems counter-intuitive, as "smoker" seems such an easily definitive class ("someone who smokes cigarettes"), and yet this is not always clearly the case. In the smoking Care Regulations, for instance, someone is still a smoker for one year after someone has smoked his or her last cigarette (Partnership Stop met Roken, 2019), based on scientific and clinical literature also claiming a six months to one year wait before a patient's status has reliably changed from "SMOKER" to "EX-SMOKER" due to the risk of relapses. The distinction between what is a "smoker" and an "ex-smoker" thus becomes less clear, and in practical use less easy to define. Additionally, someone who smokes with an e-cigarette is not a smoker according to the Care Regulations (Partnership Stop met Roken, 2019). The Care Regulations also mention that the several official definitions of "smoker" also lead to confusion. For instance, the DSM-5 classification of a "smoker" mentions addiction is neccessary to be classified a SMOKER (Partnership Stop met Roken, 2019), while a positive smoking status in the 1739 smoking variable used in our work does not imply an addiction to tobacco.

We found in our development of LFs based on this care standard that these

definitions also show some fuzzy, unclear categorizations even when there is only one smoking status variable (1739). For instance, we attempted to use a Labelling Function in our LabelModel to label ex-smokers based on the keyword "e-cigarette", but found this keyword was more related to SMOKERS than EX-SMOKERS, despite the formal definitions in (Partnership Stop met Roken, 2019) showing the opposite. Additionally, we found that medicine used for quitting smoking (such as champix) was not related to SMOKERS (who would take such medication while they were within the one-year period), but to EX-SMOKERS.[10]

This fuzziness of definitions and boundaries between smoking status classes leads to obvious problems in assessing any trained model in performance. Our labellers are the GPs using perhaps their own internal models of what a "SMOKER" is, despite the regulations indicating this differently. Some values as used by GPs that this model is trained and evaluated on are not the same as ones defined in the Zorgstandaard or other formal definitions, with care providers either using a different category than formally defined in the care standard or even using a category not used in the care standard at all. This phenomenon reiterates the work by Bowker and Star (2000) that identifies the inherent fuzziness in medical classification, especially the classification of patients and medical conditions, where formal guidelines sometimes do not capture the needs and categories in everyday use. This also calls into question the practical use of these methods for clinical practice. While established NLP research shows that extensive labelling guidelines for classification should be used to ensure multiple labellers adhere to the same definitions and standards, this is in stark contrast with clinical practice and real-life data, where the data is classified based on the individual GP's preferences.

However, our models do not regularly confuse "EX-SMOKERS" and "SMOKERS", as might be expected based on the basis of these observations of guidelines, LFs, and practical use showing some friction between these definitions. Instead, our models showed most confusion between the "ex-smoker" and "non-smoker" class, as seen in the confusion matrices in the previous chapters. This shows the training data actually shows less conflicting information when it concerns "ex-smokers" and "smokers" than perhaps expected from the conflicting guidelines and literature.

The problem of the real-life data not adhering to formal clinical standards and definitions comes on top of the earlier mentioned problem of the model being trained on EMR representations which are not even representative of the

[10]The medicine-based LF were relatively ineffective in any case because these keywords were fairly rare.

entire EMR (only consisting of the last labelled consultation text per unique patient), let alone representative of an entire patient's medical history. EMRs themselves are a limited ad one-sided representation of an individual patient's situation or history, as they cannot possibly contain all medical or clinical information of a patient - and especially where it concerns smoking behavior can have omissions or lies, also influenced by confounds such as Ketenzorg patients being asked more often about smoking status (Partnership Stop met Roken, 2019). Data is always a reduction and proxy of real life concepts and elements. We cannot capture someone's entire medical history, or even the entire consultation, simply also because it is not recorded as such.

Thus, there is categorical fuzziness of smoking status classes in the data, where someone could be identified as a SMOKER and EX-SMOKER by different GPs, medical codes, and sources even within one smoking status variable such as P1739. Together with earlier-defined problems, such as that that the data representation we used is not the same as representing an entire individual patient's medical history, is why any of our models cannot be used to reliably predict or identify individual smokers, and should not be used as such.

6.2 Data Representation

Our experiments are with a machine learning method, which requires examples (the labelled consultations) to learn how to label new, unseen examples. However, we throw away many consultations that were never used in the final representation, only using the last labelled consultation for each patient. This was a major choice and possibly major flaw in our research: why, if data is already sparsely labelled, to only use a subset of the labelled data points?

This decision was based on several practical rather than conceptual grounds. First of all, the last smoking status for each patient was the datapoint clinically relevant. Since smoking status is a fluid concept and the duration of class membership is not lifelong (Bowker & Star, 2000), for each EMR only the latest registered smoking status is relevant.

Secondly, we only used one consultation per EMR. There are several other options to provide examples for the machine learning model from the EMRs, and make representations of one EMR. One of these is the combination of several consultations from one patient into one text. However, we chose to only take only the last labelled consultation per EMR because of our method. A machine learning-based approach cannot process more evidence for one patient than for the other, as this can lead to more evidence for one EMR rather than the other, and thus unequal evidence per instance. Combining several consultations into

one representation for each unique EMR then leads to some EMRs being represented by (much) more textual information than others, and that information is biased as well since chronically ill Ketenzorg patients have more consultations. Thus, we would indirectly build a bias into our model that would make the detection of smoking status in the EMR of a chronically ill patient more reliable. An added reason for only using one consultation was that BERTje (Vries et al., 2019), our Transformer model used for classification, can only classify short instances with 512 tokens. All text from an EMR could not be used as one example in this method, since this text would in some cases far exceed 512 tokens.

However, we could have used multiple one-consultation instances from one EMR instead of having one large text per EMR. Why did we not do that? We did not use multiple instances related to the same EMRbecause of similar method constraints. This can lead to data leakage - one consultation for a patient is used in the training set, while another in the test or validation set. This is problematic, because a GP will use some similar descriptions in both consultations on the same patient - especially mentioning patient names or identifiers, since this dataset was not anonymized. Since we wanted the model to absolutely not have any prior knowledge at test time - it needed to identify entirely new, unseen information - we could not use these multiple consultation examples related to one patient. Otherwise, the model might learn that "ms. Jansen" EMRs have a P1739 smoking status that is SMOKING simply because the name "ms. Jansen" is mentioned, rather than learning anything about signifiers of actual smoking status. It would also lead to unfair advantages at test time, since the model would score higher on performance metrics simply because it had seen earlier information from ms. Jansen's EMR, rather than because the model is better at identifying smokers.

These are all methodological constraints: they do not fully reflect the medical or clinical practice. A non-machine learning method, such as a rule-based method, is not necessarily harmed by having multiple consultations of one patient, or having one large text file consisting of multiple EMRs instead of only one short consultation. That means the machine learning method we employed was limiting how and how much of the available data we could use, and not the other way around. Realizing this, we can further question whether the Machine Learning approach is a useful approach for practical use in primary care documentation, and actually the previous literature's use of rule-based method is not that strange.

6.3 Transfer learning with Pre-Trained Transformer

Our results indicate that it is possible to classify Dutch smoking status classification better than a rule-based baseline with a machine learning approach (with F1 = .79) by fine-tuning a pre-trained Transformer, BERTje (Vries et al., 2019). While this performance does perform better than the rule-based baseline, it performs notably lower than the reported F1 scores in earlier work by rule-based methods (Weng et al., 2017; Palmer et al., 2019) and other machine learning methods (Wang et al., 2019) on smoking status classification, who report F1 > .90 for similar smoking status tasks on English-language datasets.

We can explain this result in several ways. First of all, these earlier papers worked with pre-processed and especially selected shared task datasets, which likely leads to easier smoking status classification than in our real-life, realistic dataset. It is very well possible word use in our EMRs is less uniform than in these publicly released datasets. This likely means it is simply easier to get a higher score on such shared task data, such as from the ib2b task and the Mayo Clinic used in Wang et al. (2019). The latter dataset also only allows for binary smoking status classification, rather than multi-class smoking status classification, which also allows for higher performance simply because the choice is only between two classes rather than three or more.

Secondly, we are working in a different language than these earlier papers: Dutch. This affects several aspects of our research, including the Transformer model we use. We use a Transformer pre-trained on 12GB of Dutch text data (Vries et al., 2019), but there is simply much more text data available in English, with the original BERT (Devlin et al., 2019) trained on 16GB of language textdata. Additional pre-training data might improve performance on downstream tasks such as smoking class classification, as the Transformer model is then able to learn more semantic information about the language and is thus possibly able to better identify words.

Another reason we could explain our low performance compared to earlier work is that these largely rule-based methods perform so high because they are over-fitted towards their specific datasets. There is some evidence for that in Wang et al. (2019), where the authors describe very specific rule-based algorithms for their two distinct datasets. Such over-specification to the training data can lead to bad generalizability to other datasets or even other datapoints. This would make the reported score of F1 = .91 on this task far less useful for a comparison across studies using different datasets, since the performance only relates to that very specific small dataset and not to the task of smoking status classification in general.

Lastly, we wonder whether our Transformer under-performed over the reported scores in earlier literature because we used a Transformer model trained on general language. The semantic space of such a language model might be significantly different than that in the clinical or medical sector, which our texts are in. There are previously reported pre-trained language models on clinical text, such as Alsentzer et al. (2019), but these are trained on English language data rather than Dutch. Furthemore, many of such pre-trained models are not publicly available because of its training on sensitive material.

All in all, we can conclude that smoking status classification is possible with a pre-trained Transformer model, but that it cannot seem to perform ss well as earlier state-of-the-art performances on similar smoking classification tasks in the literature. However, this was with other datasets, and in English instead of Dutch texts.

6.4 Weak Supervision and Data Programming Paradigm

We worked with SNORKEL (Ratner et al., 2017), a software package allowing us to use the Data Programming approach, to see whether this increases performance on smoking status classification and also whether it is a viable approach to label the sparsely labelled data. The LabelModel function in SNORKEL is able to learn from a set of heuristics and rules (Labelling Functions) the optimal weighting for these individual Labelling Functions, in order to optimally label new examples. The LabelModel exploits the conflicts between LFs to find the optimal label for an example.

We first of all found that, despite SNORKEL's weighting, even a model with all individual labelling fuctions of > .80 accuracy performed on the labelled development set with only .39 accuracy. This is surprising: we expected the model to be able to generalize above the accuracy and performance of the individual rules in the LFs, but it was not.

Why could this be? One reason could be the limited coverage of many of the LFs, meaning they simply do not apply to many items. Low coverage does not only lead to fewer items labelled, but it can also lead to a worse performance due to the LabelModel's ability to learn from especially conflicts between labelling functions. Nearly all LFs had a coverage below 5%, as is visible in Appendix D. This low coverage is also a feature of the data: not only is smoking status variable P1739 rarely documented, words related to smoking are also rarely mentioned in the consultation texts. This makes it quite difficult to find the overlap and conflict the LabelModel needs to accurately adjust the weighting of different LFs (Ratner et al., 2017). Our data therefore might be less suitable

for such a weak supervision paradigm than earlier assumed.

The second step of the Data Programming Paradigm consisted of extended the training data size with the LabelModel, and label a fraction of the 9.232 unlabelled EMR examples in the training set. With our most accurate LabelModel on the labelled dataset, this led to 512 more examples. This is 10% of the already labelled training set of 4.978 items, thus leading to a 10% increase of training data size. This is not a large newly labelled dataset, but training with a less accurate LabelModel might actually lead to more noise in the training set instead of more training evidence and better performance. We found the small training set increase did not lead to better performance. We did find an improvement of in-class performance of the NEVER smoker class, while the added examples added more confusion for the fine-tuned BERTje model between the NEVER and SMOKER class. Future work might opt to see if labelling more examples with SNORKEL might improve performance even if such a larger labelled dataset is more noisy due to a less accurate LabelModel.

6.5 Future Work and Related Tasks

The set-up with a pre-trained model and a Data Programming paradigm makes sense for tasks where there is little (labelled) data, but sufficient evidence in the texts to identify the classes with Labelling Functions. This might be a fruitful research methodology for several related research problems where a high documentation load is resulting in a low number of documents labelled with ICD codes or other clinical variables in clinical practice, such as sleep disorders (Filip et al., 2017) and obesity (Hossain et al., 2018), both described as under-documented and under-reported in EMRs.

Our approach is also especially useful for tasks with fuzzy or ever-changing labels. This is because the LFs allow domain experts to have control over the machine learning algorithm by controlling the training data, which could be used for similar tasks in different domains. For instance, in the financial domain the usage of Labelling Functions to weakly supervise and increase training data for potential classification of fraudulent transactions or scams (Ngai et al., 2011). Such classes and especially signifiers and heuristics of these classes can change rapidly when scammers and fraudsters change their tactics and the words they use to describe the tactic. Then it is very useful to be able to immediately influence the training data systematically by adjusting the LFs, which in turn would change the model's predictions. This is similar to our find that the real-world category SMOKER had e-cigarette as keyword, rather than the on the basis of the guidelines expected NON or EX-SMOKER. We could eas-

ily adjust our LFs based on real-world knowledge about the data not in official guidelines

Lastly, future work might explore improvements on our weak supervision approach. The low coverage of our LFs could possibly have been prevented if we used a data mining technique rather than a literature search, manual inspection of development set cases, and expert guidance to design the LFs. For instance, we could have attempted some data mining technique to extract meaningful features from the labelled EMRs that could be turned into Labelling Functions. This is what (Kunde et al., 2020) did, who used Principal Component Analysis (PCA) to identify common features and words in e-commerce and financial classification, which would then be used to expand the dataset. Such data-based rather than expert-based heuristics are not what SNORKEL (Ratner et al., 2017) is designed to do. In fact, it goes against SNORKEL's main principle of allowing high-quality labels related to the expert knowledge of humans, because such an approach would not neccesarilty give high-quality labels based on quality guidelines and expert knowledge. However, a data mining approach could be very effective in finding LFs that are accurate *and* have high coverage. Future research might explore this tension between SNORKEL's intended use and possible data mining applications to improve a LabelModel.

Additionally, we used only one smoking status variable: P1739. Future work might want to combine several smoking status variables, including for instance P17 (a long-term smoking status), in order to see whether this allows for more training and test data for smoking status classification of EMRs. It might also be a more realistic representation of the clinical practice of smoking status, as GPs use several of these variables. However, a potential problem then is concept drift: these different smoking status classes in different variables might not have the same meaning, either in official definitions or simply in how they are used by GPs.

6.6 Limitations and Reflections

This thesis has several limitations. First of all, the relatively large amount of data we worked with was positive for large-scale neural models, and also useful because the phenomenon we were interested in (smoking status) was fairly rare in the EMRs. Previous work used pre-processed Shared Task data, which was publicly available but a small and simplistic dataset. Our pipeline was more realistic, with real-life clinical data. However, this also meant we worked with tens of thousands of samples that needed preprocessing, and it was difficult to oversee all features, variables, and aspects of the data influencing the research

question. Thus, we would advice future work and especially future thesis students to not take on such a large, complex dataset and expect to preprocess and research it in within 4 to 6 months.

Another limitation of this study is the testing and training of our models on data from six different GP centers in the Netherlands. While six different GP offices is more diverse than only one or two, it still only provides data from a very limited subset of the roughly 5.000 GP centers in the Netherlands. Other GP offices might have different documentation practices concerning smoking status, leading to a model that is less universally applicable than ideal.

Additionally, another possible limitation related to the generalizability of our results is the earlier mentioned problem of bias. In our dataset, EMRs of patients in the chronic illness program were much more likely to have a documented smoking status, leading to their over-representation in the training data. This could lead to a model biased towards the detection of smoking status in EMRs of chronically ill patients, while other patient groups' EMRs are less likely to be correctly identified in terms of smoking status by our model. Furthermore, this bias intersects with social-economic status and wealth, as such chronic illnesses are often related to poverty. This would lead to the kinds of biases mentioned by O'Neil (2016) in harmful machine learning practices: underprivileged and poor groups are disproportionately identified by machine learning models identifying certain behaviours or risks simply because these groups are more prevalent in the training data. Such a situation is something we would want to avoid, so future work should carefully identify such biases, and attempt to minimize them.

Lastly, our set-up with results based on training data labelled with another algorithm (the SNORKEL model) can especially lead to cascading error: if the SNORKEL LabelModel is wrong, the classification model will learn the wrong information. Such "cascading bias" is something inherent in our set-up of the weak supervision approach, and a potential weak spot of it, and thus also something to be wary of when interpreting the results.

7 Conclusion

In this thesis, we set out to do automatically classify and extract smoking status from Dutch Electronic Medical Records (EMR) from General Practicioners (GPs). We found a gap in previous research on smoking status classification. Earlier work often worked with small, English-language dataset, and used simple methods. We also found a specific problem: smoking status is under-documented and sparsely labelled in EMRs. We explored using a weak supervision as well as an transfer learning approach to combat this small dataset problem.

We attempt to answer the following question:

> *"How can we best automatically detect and classify the smoking status of primary care patients' EMR (Electronic Medical Record) on the basis of the free text in GP doctor's notes?"*

from which we obtained three sub-questions:

- Can we obtain similar performance with a rule-based baseline on smoking status classification as earlier literature (Palmer et al., 2019; Weng et al., 2017; Palmer et al., 2019)?
- Does transfer learning, by fine-tuning a pre-trained BERT Transformer model (Vries et al., 2019), improve performance over the rule-based baseline in classifying smoking status?
- Does a weak supervision programming paradigm (Ratner et al., 2016), by labelling more training data points with a generative labelling model (Ratner et al., 2017), improve performance over the rule-based baseline and the earlier classification model in classifying smoking status?

We made representations of 17.873 EMRs from 6 GP offices in the Netherlands in order to answer this question. We obtained the data through working with software provider Topicus, who stores data for these GP offices. In order to develop our models, we split the dataset in 14.298 training examples, 1.788 development items, and 1.787 test items. We then use the labelled subset of the training data (4.978 examples) to fine-tune the pre-trained Transformer model BERTje (Vries et al., 2019), and use the labelled development set (655 labelled EMRs) to both develop the rule-based baseline algorithm for smoking status classification and develop a LabelModel with SNORKEL (Ratner et al., 2017). Such a LabelModel is able to label the unlabelled examples from the training set by training on the labelled training data, and setting weights to a set of Labelling Functions or heuristics. We then again fine-tune a BERTje model with the training data enlarged with the datapoints labelled by the LabelModel.

We test all our machine learning model on the same unseen test examples: the 628 labelled datapoints from the test set portion of our dataset.

7.1 Question 1: Rule-Based Baseline

We first attempted to develop a rule-based baseline such as in earlier literature. We test earlier developed algorithms in (Wang et al., 2019) as well as combine these with our own heuristics to make our own algorithm. For sub-question 1, we found that we could **not** replicate the high performance of F1 > .90 with a rule-based method. In earlier studies (Wang et al., 2019; Weng et al., 2017), rule-based methods seemed to perform very well even when these were baseline approaches in the study, while this was not the case for us. Our replications of the earlier algorithms performed worse (F1 = 0.52) than our own (F1 = 0.55), but none of these come close to the score reported in the literature (F1 > .90) on smoking status classification with rule-based methods on English-language datasets. We cannot determine whether this is because we did not fine-tune (and perhaps overfit) the rules enough to the dataset, or whether perhaps the language use on smoking status is more diverse than we anticipated. This latter hypothesis was reinforced by the low coverage of the LFs in the SNORKEL model, since this means some words only very rarely appear in these texts. More research is needed, but this result does provide some tentative evidence that rule-based methods are not always the best approach for smoking status classification in EMRs.

7.2 Question 2: Transfer Learning results with BERT

We found transfer learning with the pre-trained Transformer network BERTje trained on general Dutch texts (Vries et al., 2019) is able to, with a minimum of computing power and training time, safely beat simple rule-based baselines for the smoking status classification task. We add one linear classification layer on top of BERTje, and training of 2 epochs and a learning rate of 0.00005.

This fine-tuned model has a micro-F1 of .79 on the labelled test set. This out-performed the baseline (F1 = .55), but did not out-perform the reported high performance scores for especially rule-based systems in the literature on other datasets with similar smoking classification tasks (F1 = .90).

Transfer learning does not seem to improve over state-of-the-art smoking status classification reported in other studies. However, it does perform significantly better than a rule-based method on our own dataset, and that with minimal training time and computing power with a Transformer model trained on a different language domain. This shows promise for pre-trained language

models, and shows such models are versatile in the tasks they can perform on datasets distinct from their previous training data.

The fact we do not come close to performance reported by earlier machine learning work such as Wang et al. (2019) on smoking status classification can possibly be explained by the fact we work on a different language (Dutch rather than English) and that we work with realistic clinical data rather than prepared shared task data sets.

7.3 Question 3: Weak Supervision with SNORKEL + BERT.

Our final experiments consisted of a weak supervision approach with the Data Programming paradigm of SNORKEL (Ratner et al., 2017), allowing us to train a model that can learn how to label new training data on the basis of rules (Labelling Functions, or LFs). We found that a LabelModel with fewer LFs but ones that were accurate was performing better on the development set than one with more, but less accurate, LFs. However, the best performing LabelModel still performed with an accuracy of only .39.

We were able to label an additional 512 examples with this LabelModel, leading to 5.490 training items instead of 4.978. Fine-tuning BERTje again with two additional epochs led to a micro F1 of .79 compared to a similar F1 of .79 for the model without this extra data. However, we do see some improvements in especially the in-class performance of the SMOKER and NON-SMOKER class as compared to the same in-class results for the model only trained with hand-labelled data.

The answer to our third partial research question is that a learning model SNORKEL can partially improve performance over a model without SNORKEL. While overall performance on the test set did not seem to increase, in-class performance for SMOKER and NON-SMOKER clearly gained from the extra labelled datapoints. This effect might be more pronounced with more datapoints labelled by SNORKEL, or with a stronger and more accurate LabelModel.

7.4 Overall Research Question

This thesis attempted to answer the following question: *"How can we best automatically detect and classify the smoking status of primary care patients' EMR (Electronic Medical Record) on the basis of the free text in GP doctor's notes?"* All in all, we can answer our research question as follows.

We found a machine-learning based Transfer learning approach works overall well, while a weak supervision approach to data labelling can increase in-class accuracy for some classes. In this study, despite earlier literature such as

(Palmer et al., 2019) stating otherwise, we found a rule-based method is not necessarily the best method to classify smoking status in EMRs. We found such a method performs worse than the pre-trained Transformer model BERTje (Vries et al., 2019), which obtained F1 = .79 on the same test set while the rule-based method only scored F1 = .55. While BERTje does not come close to results shown in earlier literature on English-language smoking status classification (Wang et al., 2019; Weng et al., 2017), it does perform rather adequately, classifying more than 80% of all NON-SMOKER and EX-SMOKER EMRs correctly and 64% of SMOKER EMRs. This result falls in line with previous literature, where BERTje and other Transfer learning models perform well on all kinds of language task with minimal fine-tuning.

We also explore a weakly supervised approach, where we attempt to enlargen the training set size by a trained LabelModel with SNORKEL (Ratner et al., 2017). We find no performance improvement with this method on the test set over the BERTje trained on only the hand-labelled datapoints, but we do see some improvement on the classification of EX and NEVER smokers. There also is the additional factor of Ketenzorg: for chronically ill patients in this program, training on only Ketenzorg EMRs helps the model distinguish especially Ketenzorg NEVER and EX SMOKING EMRs.

All in all, smoking status classification is a text mining problem in a complex domain (the clinical setting, specifically: primary care) with several societal and technical factors influencing the results. However, BERTje allows us to set a promising first step in performing smoking status classification more accurately than a rule-based approach, with minimal training time and computing resources. Future work can carefully step forward based on these results.

List of Figures

List of Tables

A Appendix: Details of Rule-Based Algorithms

Table 18: The regular expression algorithms used to detect different smoking status classes in different datasets by Wang et. a. 2019)

	dataset	**RegEx**
"smoker"	Mayo Clinic	Smoker smokes?, smoked, smoking, smokers?, tobaccos?, cigarettes?, cigs?, pipes?, nicotine, cigars?, tob
"never smoked"	Mayo Clinic	(no | non | not | never | negative)*(smoker | smoking | smoked | tobacco), nonsmoker, denies*smoking, (tobacco | smoke | smoking | nicotine)*(never | no), doesnt́ smoke, 0 | zero smokers?
"current smoker"	ib2b 2006	(does | has | continues to) smoked?, uses tobacco, active smoker, (current | currently) (smoker | smoking), current smoker, tobacco use*(yes | still using | still smoking | smokes)
"ex-smoker"	ib2b 2006	(stop | stopped | quit | quitted | discontinued) (tobacco | smoking), (previous | prior | remote | distant | former | ex- | ex) (tobacco | smoker), stop(ped)? smoking, tobacco use*(smoked | quit), smoking*(used | former)
"non-smoker"	ib2b 2006	default: other patients

Table 19: Our Dutch translated regular expression algorithms used to detect different smoking status classes in different dataset by Wang et. al. (2019)

	dataset	**RegEx**
"smoker"	Mayo Clinic	roker, rookt?, rookte, roken, rokers?, tabak?, sigaretten?, sig?, pijp?, nicotine, siga(a)?r(en)?
"never smoked"	Mayo Clinic	(niet | non | geen | nooit | negatief)*(roker | roken | rookte | tabak), niet-roker, ontkent*roken, (tabak | rook | roken | nicotine)*(nooit | niet), rookt niet, 0 | nul rokers?
"current smoker"	ib2b 2006	(heeft | doet | blijft) (ge)?rookt | roken, gebruikt tabak, huidig roker, (huidig | momenteel | nu | nog) (roker | roken), tabakgebruik*(ja | nog steeds | rookt nog | rookt)
"ex-roker"	ib2b 2006	(is gestopt | stopte | hield op met | hield op) (tabak | roken), (vorige | vroeger | verleden | ooit | voormalig | ex- | ex) (tabak | roker, stop(te)? roken, tabaks?( )?gebruik)?*(rookte | stopte), roken*(gebruikte | vroeger)
"non-smoker"	ib2b 2006	default: other patients

B Appendix: Architecture of BERTje

Structure of BERTje:

==== Embedding Layer ====

bert.embeddings.wordembeddings.weight (30000, 768)
bert.embeddings.positionembeddings.weight (512, 768)
bert.embeddings.tokentypeembeddings.weight (2, 768)
bert.embeddings.LayerNorm.weight (768,)
bert.embeddings.LayerNorm.bias (768,)

==== First Transformer ====

bert.encoder.layer.0.attention.self.query.weight (768, 768)
bert.encoder.layer.0.attention.self.query.bias (768,)
bert.encoder.layer.0.attention.self.key.weight (768, 768)
bert.encoder.layer.0.attention.self.key.bias (768,)
bert.encoder.layer.0.attention.self.value.weight (768, 768)
bert.encoder.layer.0.attention.self.value.bias (768,)
bert.encoder.layer.0.attention.output.dense.weight (768, 768)
bert.encoder.layer.0.attention.output.dense.bias (768,)
bert.encoder.layer.0.attention.output.LayerNorm.weight (768,)
bert.encoder.layer.0.attention.output.LayerNorm.bias (768,)
bert.encoder.layer.0.intermediate.dense.weight (3072, 768)
bert.encoder.layer.0.intermediate.dense.bias (3072,)
bert.encoder.layer.0.output.dense.weight (768, 3072)
bert.encoder.layer.0.output.dense.bias (768,)
bert.encoder.layer.0.output.LayerNorm.weight (768,)
bert.encoder.layer.0.output.LayerNorm.bias (768,)

==== Output Layer ====

bert.pooler.dense.weight (768, 768)
bert.pooler.dense.bias (768,)
classifier.weight (3, 768)
classifier.bias (3,)

C Appendix: Extensive Performance Metrics

C.1 Defining macro, micro, and weighted F1 score

- macro-average:
  score class 1 + score class 2 + score class 3 / number of classes = F1-macro
- micro-average:
  TP = TP of all classes together, FP = FP of all classes together, FN = FN of all classes together
- weighted average F1:
  *(weight * score1) + (weight * score2) + (weight * score3) / number of classes*

C.2 Results with macro, micro, and weighted F1 score

Table 20: Performance on the test set for all of the smoking status classes for models trained with the labelled development set. NOTE; the mayo clinic algorithm only has 2 classes

		algorithm ib2b	algorithm Mayo clinic	own algorithm
precision	micro	-	0.44	**0.49**
	macro	0.39	0.39	**0.48**
	weighted	0.43	0.44	**0.54**
recall	micro	-	**0.65**	0.43
	macro	0.36	**0.55**	0.40
	weighted	0.41	**0.65**	0.43
F1	micro	-	0.52	**0.55**
	macro	0.28	**0.44**	0.38
	weighted	0.30	**0.52**	0.49

Table 21: In-class performance on the test set for all of the smoking status classes for models trained with the labelled development set.

		Rule-Based	BERTje
precision	micro	0.49	**0.79**
	macro	0.48	**0.83**
	weighted	0.54	**0.80**
recall	micro	0.43	**0.79**
	macro	0.40	**0.77**
	weighted	0.43	**0.79**
F1	micro	0.55	**0.79**
	macro	0.38	**0.79**
	weighted	0.49	**0.79**

Table 22: Comparison of all models and their performance on the test set (628 EMRs)

		Rule-Based	BERTje	SNORKEL + BERTje
precision	micro	0.49	0.79	0.79
	macro	0.48	**0.83**	0.81
	weighted	0.54	**0.80**	0.79
recall	micro	0.43	0.79	0.79
	macro	0.40	**0.77**	0.76
	weighted	0.43	0.79	0.79
F1	micro	0.55	0.79	0.79
	macro	0.38	**0.79**	0.78
	weighted	0.49	0.79	0.79

D Appendix: Labelling Functions Performance

D.1 LabelFunctions Coverage

keyword_esig_coverage coverage: 0.0%
keyword_nvm_coverage coverage: 0.0%
keyword_pakje_coverage coverage: 0.7%
keyword_pakjes_coverage coverage: 0.3%
keyword_roken_coverage coverage: 31.4%
keyword_champix_coverage coverage: 0.6%
keyword_gerookt_coverage coverage: 2.9%
keyword_gestopt_coverage coverage: 7.7%
keyword_gestopt2_coverage coverage: 2.6%
keyword_heeftgerookt_coverage coverage: 0.1%
keyword_nicotine_coverage coverage: 0.2%
keyword_nicotineverslaving_coverage coverage: 0.0%
keyword_nooitgerookt_coverage coverage: 0.3%
keyword_roker_was_coverage coverage: 0.0%
keyword_roker_coverage coverage: 0.7%
keyword_rookt_coverage coverage: 12.0%
keyword_rookte_coverage coverage: 0.6%
keyword_rooktgeen_coverage coverage: 0.1%
keyword_rooktgeenis_coverage coverage: 0.0%
keyword_rooktniet_coverage coverage: 3.0%
keyword_sig_coverage coverage: 1.5%
keyword_sigs_coverage coverage: 1.0%
keyword_stoppen_coverage coverage: 2.6%
keyword_tabak_coverage coverage: 0.1%
keyword_tabaksverslaving_coverage coverage: 0.0%
regex_rookt_niet_coverage coverage: 7.0%
regex_rookt_niet_meer_coverage coverage: 1.1%
regex_heeft_gerookt_coverage coverage: 1.4%
regex_is_roker_coverage coverage: 0.3%

D.2 Conflicts Individual Labelling Functions

- Polarity -> Does the LF output all possible labels?
- Coverage -> How many datapoints does the LF cover?

- Overlaps -> Does this LF overlap with at least one more LF on datapoints?
- Conflicts -> Does this LF disagree with one other LF?

Based on Given Labels in Development Set:

- Correct -> How many of data points does this LF label correctly in the development set?
- Incorrect -> How many data points does this LF label incorrectly in the development set?
- Accuracy –> What is the accuracy of this LF on the development set?

	Polarity	Coverage	Overlaps	Conflicts
keyword_e-sigaret	0	0.000402	0.000402	0.000402
keyword_nvm	1	0.000201	0.000201	0.000201
keyword_pakje	2	0.007431	0.006829	0.005021
keyword_pakjes	3	0.002812	0.002812	0.001808
keyword_roken	4	0.313918	0.102832	0.102832
keyword_champix	5	0.005624	0.005222	0.005021
keyword_cytisine	6	0.000000	0.000000	0.000000
keyword_gerookt	7	0.028721	0.028721	0.022294
keyword_gestopt	8	0.077325	0.057240	0.057240
keyword_gestopt met roken	9	0.025909	0.025909	0.025909
keyword_heeft gerookt	10	0.000603	0.000603	0.000603
keyword_nicotine	11	0.002410	0.002209	0.002008
keyword_nicotineverslaving	12	0.000000	0.000000	0.000000
keyword_heeft nooit gerookt	13	0.002611	0.002611	0.002611
keyword_was roker	14	0.000000	0.000000	0.000000
keyword_roker	15	0.006829	0.005624	0.004820
keyword_rookt	16	0.119703	0.107451	0.098012
keyword_rookte	17	0.005824	0.005824	0.005824
keyword_geen roker	18	0.000803	0.000803	0.000803
keyword_is geen roker	19	0.000000	0.000000	0.000000
keyword_rookt niet	20	0.029725	0.029725	0.029725
keyword_sigaret	21	0.014862	0.014862	0.012251
keyword_sigaretten	22	0.010042	0.010042	0.007833
keyword_stoppen met roken	23	0.025507	0.025507	0.025507
keyword_tabak	24	0.001406	0.001205	0.001205
keyword_tabaksverslaving	25	0.000201	0.000201	0.000201
keyword_varencline	26	0.000000	0.000000	0.000000

regex_rookt_niet 27 [1] 0.070094 0.070094 0.070094
regex_rookt_niet_meer 28 [2] 0.011247 0.011247 0.011247
regex_heeft_gerookt 29 [0] 0.013657 0.013657 0.011649
regex_was_roker 30 [0] 0.000803 0.000803 0.000803
regex_is_roker 31 [0] 0.003013 0.003013 0.002209

D.3 Accuracy Individual Labelling Functions

	Correct	**Incorrect**	**Emp. Acc.**
keyword_e-sigaret	1	1	0.500000
keyword_nvm	1	0	1.000000
keyword_pakje	21	16	0.567568
keyword_pakjes	3	11	0.214286
keyword_roken	549	1013	0.351248
keyword_champix	13	15	0.464286
keyword_cytisine	0	0	0.000000
keyword_gerookt	15	128	0.104895
keyword_gestopt	277	108	0.719481
keyword_gestopt met roken	125	4	0.968992
keyword_heeft gerookt	1	2	0.333333
keyword_nicotine	5	7	0.416667
keyword_nicotineverslaving	0	0	0.000000
keyword_heeft nooit gerookt	13	0	1.000000
keyword_was roker	0	0	0.000000
keyword_roker	18	16	0.529412
keyword_rookt	268	328	0.449664
keyword_rookte	21	8	0.724138
keyword_geen roker	0	4	0.000000
keyword_is geen roker	0	0	0.000000
keyword_rookt niet	96	52	0.648649
keyword_sigaret	53	21	0.716216
keyword_sigaretten	41	9	0.820000
keyword_stoppen met roken	99	28	0.779528
keyword_tabak	6	1	0.857143
keyword_tabaksverslaving	1	0	1.000000

keyword_varencline 0 0 0.000000
regex_rookt_niet 128 221 0.366762
regex_rookt_niet_meer 37 19 0.660714
regex_heeft_gerookt 8 60 0.117647
regex_was_roker 3 1 0.750000
regex_is_roker 11 4 0.733333

E Appendix: Training on only Ketenzorg EMRs

Our second set of experiments consisted only those EMRs that are in the KetenZorg group. These are chronically ill patients - with diabetes, COPD, or heart failure - and these patients have yearly consultations with their GP, where the GP is also required to ask about smoking status (Partnership Stop met Roken, 2019).

There are 2.726 training EMR representations that are ketenzorg. There are 363 EMR representations in the development set that are ketenzorg. We train our models in this section with only these examples, and test them also with this sub-section.

E.1 Model I: Rule-Based Baseline

In Table 23, we see the results from the overall classification results. We see the Mayo Clinic algorithm working really well for the ketenzorg population on the general performance on the test set, with the highest F1 score both measured in micro (F1 = 0.41), macro (F1 = 0.29) and weighted (F1 = 0.43). Apparently, the Mayo clinic algorithm works well for smoking status classification in this specific EMR group.

Our in-class results in Table 26 show something else: there, the individual classes perform best within the ib2b algorithm. Interesting to note is that the EX-SMOKING class has optimal precision (prec = 1.00) on this test set, while very low recall (rec = 0.01), which is almost opposite for the NON-SMOKING class: there, the ib2b algorithm as developed by Wang et al. (2019) gives high recall (rec = 0.86), but not high precision (prec = 0.39).

E.2 Model II: BERTje FineTuned on Ketenzorg

The results of our fine-tuning of BERTje (Vries et al., 2019) is also displayed in Table 25. We compare three models on the same test set: the KetenZorg examples from the test set. We compare a rule-based model, a Ketenzorg-trained BERTje model, and the BERTje model trained on everything in their performance on the Ketenzorg examples in the test set, which are 176 EMR examples.

Despite being trained with smaller sample of examples (1.407 training examples are Ketenzorg, versus the 4.978 in the entire training set), the Ketenzorg-trained Bertje appears to, however slightly, improve over the BERTje trained on all the dataset. The same effect is not seen when only training on non-ketenzorg examples (3.571), which harms performance on the BERTje test set. This can

Table 23: Performance on the **Ketenzorg** labelled test set for all of the smoking status classes. NOTE; the mayo clinic algorithm only has 2 classes

		algorithm ib2b	algorithm Mayo clinic	own algorithm
precision	micro	0.31	**0.44**	0.35
	macro	**0.50**	0.22	0.36
	weighted	**0.71**	0.33	0.49
recall	micro	0.34	**0.61**	0.31
	macro	0.38	**0.43**	0.30
	weighted	**0.71**	0.61	0.31
F1	micro	0.34	**0.41**	0.33
	macro	0.24	**0.29**	0.26
	weighted	0.22	**0.43**	0.29

Table 24: In-class performance of the rule-based algorithms on the test set of only **Ketenzorg** examples

	algorithm ib2b			algorithm Mayo clinic			own algorithm		
	precision	recall	F1	precision	recall	F1	precision	recall	F1
SMOKING	0.11	**0.29**	0.15	0.05	0.14	0.08	0.06	0.14	0.08
NON-SMOKING	0.39	**0.86**	0.54	0.39	0.71	0.51	0.41	0.65	0.50
EX-SMOKING	**1.00**	0.01	0.02	-	-	-	0.61	0.11	0.19

be seen in Table 25, where we see especially the weighted and micro performance scores (where size of the predicted classes is taken into account) is outperforming a BERTje model trained on all classes.

Looking at in-class performances, we see this difference in performance becomes more pronounced for especially the ex-smoking and non-smoking class. Interestingly, identifying the SMOKER class is considerably easier with the BERTje trained on the entire dataset, while the other classes become easier to detect with a non

This strengthens two assumptions we had: that the Ketenzorg population is a separate population from the general population, and that within ketenzorg there are different descriptions used of smoking status than in the general population.

A short glance on some of our development texts show indeed this: the ketenzorg texts on a positive smoking status often describe longer, more narrative texts, while "never smokers" in the non-ketenzorg groep often describe cough-

ing or astma.

Table 25: Performance of predicting only the **Ketenzorg** examples of the test set (176 EMR examples)

		Rule-Based	Ketenzorg-trained BERTje	all-trained BERTje	non-Ketenzorg-trained BERTje
precision	micro	0.35	**0.81**	0.79	0.75
	macro	0.36	0.76	**0.79**	0.73
	weighted	0.49	**0.80**	0.79	0.75
recall	micro	0.31	**0.81**	0.79	0.75
	macro	0.30	0.73	**0.74**	0.63
	weighted	0.31	**0.81**	0.79	0.75
F1	micro	0.33	**0.81**	0.79	0.75
	macro	0.26	0.75	**0.76**	0.66
	weighted	0.29	**0.80**	0.79	0.74

Table 26: In-class performance of predicting only the **Ketenzorg** examples of the test set (176 EMR examples)

	Ketenzorg-trained BERTje *1.407 training examples*			all-trained BERTje *4.978 training examples*			non-Ketenzorg trained BERTje *3.571 training examples*		
	precision	recall	F1	precision	recall	F1	precision	recall	F1
SMOKING	0.67	0.57	0.62	**0.82**	**0.64**	**0.72**	0.71	0.36	0.48
NON-SMOKING	**0.78**	0.75	**0.76**	0.74	**0.76**	0.75	0.71	0.70	0.70
EX-SMOKING	**0.84**	**0.88**	**0.86**	0.82	0.83	0.82	0.78	0.84	0.81

E.3 Analyzing Confusion Matrices

We also analyzed the confusion matrices for the Ketenzorg trained model predicting the 176 EMRs in the test set that are ketenzorg. We see that the model also when only trained and tested on ketenzorg is primarily good at identifying the NEVER and EX SMOKING status, with Figure 14 showing most confusion happens when the model predicts EX-SMOKER for NEVER-SMOKER (15 items), or predicts NEVER=SMOKER for EX-SMOKEr (also 15 items). In Figure 15 it is visible that 61% of all smokers are correctly labelled, while up to 75% of NEVER-SMOKERS and 81% of EX-SMOKERS. Apparently, these are easier to detect for the model. Figure 16 does show there is least confusion of the SMOKER class with another class: 83% of all predicted SMOKER items are really SMOKER, against only 72% of NEVER smokers and 77% of EX-SMOKERS.

Figure 14: Confusion Matrix on the **ketenzorg** trained model, showing the distribution of predicted labels (in the vertical columns) and the true labels (horizontal columns) of the test set. The test set has a total of 176 examples with 14 SMOKER examples, 63 NEVER SMOKER, and 99 EX-SMOKER. We see most examples are correctly predicted, though there is considerable confusion between the NEVER and EX class.

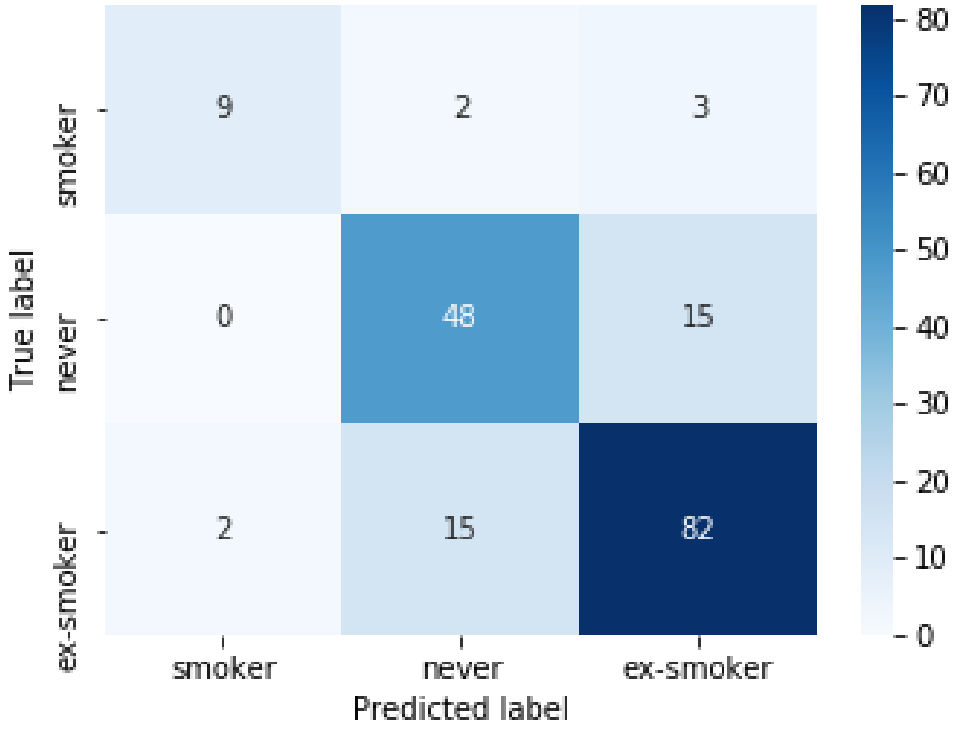


Figure 15: Confusion Matrix on the **ketenzorg** trained model, showing the **relative** (in percentage) distribution over True labels (in the vertical columns) and the true labels (horizontal columns) for the ketenzorg examples in the test set (176 examples). We see 61% of all smokers are correctly predicted, with 75% of NEVER and 81% of EX

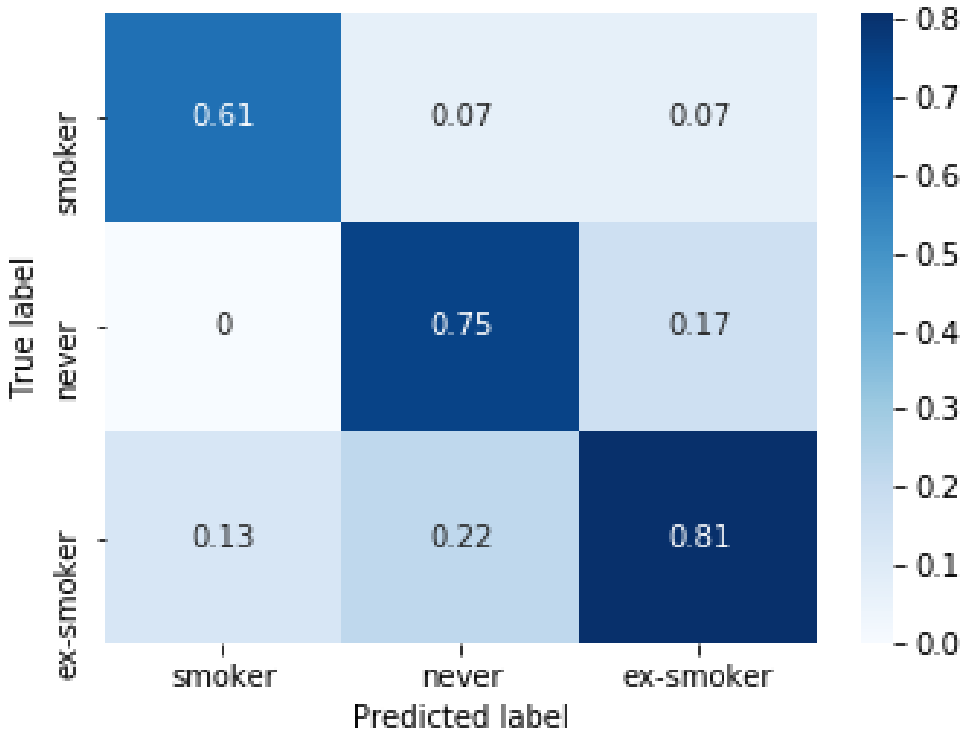
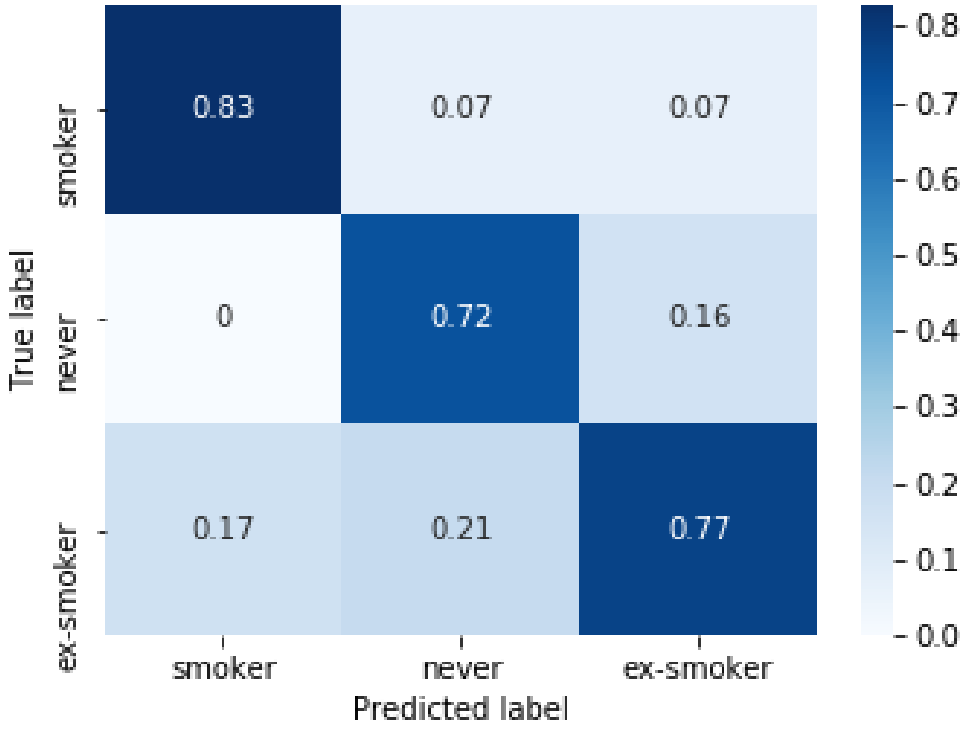


Figure 16: Confusion Matrix on the **ketenzorg** trained model, percentage of predicted labels. In the most left column we see the percentages for smoking prediction: 83% of predicted smokers are actually smokers.



References

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., & McDermott, M. (2019). Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Beeksma, M., Verberne, S., Van Den Bosch, A., Das, E., Hendrickx, I., & Groenewoud, S. (2019). Predicting life expectancy with a long short-term memory recurrent neural network using electronic medical records. *BMC medical informatics and decision making*, *19*(1), 36.

Bowker, G. C., & Star, S. L. (2000). *Sorting things out: Classification and its consequences*. MIT press.

Centraal Bureau Statistiek. (2018). *Helft van laagopgeleide 25- tot 45-jarige mannen rookt.* `https://www.cbs.nl/nl-nl/nieuws/2018/22/helft-van-laagopgeleide-25-tot-45-jarige-mannen-rookt`. ((Accessed on 04/08/2020))

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1* (pp. 4171–4186). Association for Computational Linguistics. Retrieved from `https://www.aclweb.org/anthology/N19-1423` doi: 10.18653/v1/N19-1423

Filip, I., Tidman, M., Saheba, N., Bennett, H., Wick, B., Rouse, N., ... Radfar, A. (2017). Public health burden of sleep disorders: underreported problem. *Journal of Public Health*, *25*(3), 243–248.

Ford, E., Carroll, J. A., Smith, H. E., Scott, D., & Cassell, J. A. (2016). Extracting information from the text of electronic medical records to improve case detection: a systematic review. *Journal of the American Medical Informatics Association*, *23*(5), 1007–1015.

Ghelani, S. (2019, June). *Snorkel — a weak supervision system - towards data science.* `https://towardsdatascience.com/snorkel-a-weak-supervision-system-a8943c9b639f`. ((Accessed on 04/10/2020))

Hecht-Nielsen, R. (1992). Theory of the backpropagation neural network. In *Neural networks for perception* (pp. 65–93). Elsevier.

Horng, S., Sontag, D. A., Halpern, Y., Jernite, Y., Shapiro, N. I., & Nathanson, L. A. (2017). Creating an automated trigger for sepsis clinical decision

support at emergency department triage using machine learning. *PloS one*, *12*(4).

Hossain, M. A., Amin, A., Paul, A., Qaisar, H., Akula, M., Amirpour, A., ... others (2018). Recognizing obesity in adult hospitalized patients: a retrospective cohort study assessing rates of documentation and prevalence of obesity. *Journal of clinical medicine*, *7*(8), 203.

Islami, F., Torre, L. A., & Jemal, A. (2015). Global trends of lung cancer mortality and smoking prevalence. *Translational lung cancer research*, *4*(4), 327.

Jie, M., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., van Calster, B., et al. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of clinical epidemiology*.

Jurasky, D., & Martin, J. H. (2000). Speech and language processing: An introduction to natural language processing. *Computational Linguistics and Speech Recognition. Prentice Hall, New Jersey*.

King, C. C., Piper, M. E., Gepner, A. D., Fiore, M. C., Baker, T. B., & Stein, J. H. (2017). Longitudinal impact of smoking and smoking cessation on inflammatory markers of cardiovascular disease risk. *Arteriosclerosis, thrombosis, and vascular biology*, *37*(2), 374–379.

Kreimeyer, K., Foster, M., Pandey, A., Arya, N., Halford, G., Jones, S. F., ... Botsis, T. (2017). Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. *Journal of biomedical informatics*, *73*, 14–29.

Kunde, S., Mishra, M., Pandit, A., Singhal, R., Nambiar, M. K., Shroff, G., & Gupta, S. (2020). Recommending in changing times. In *Fourteenth acm conference on recommender systems* (p. 714–719). New York, NY, USA: Association for Computing Machinery. Retrieved from `https://doi.org/10.1145/3383313.3418492` doi: 10.1145/3383313.3418492

Mani, S., Chen, Y., Elasy, T., Clayton, W., & Denny, J. (2012). Type 2 diabetes risk forecasting from emr data using machine learning. In *Amia annual symposium proceedings* (Vol. 2012, p. 606).

Marston, L., Carpenter, J. R., Walters, K. R., Morris, R. W., Nazareth, I., White, I. R., & Petersen, I. (2014). Smoker, ex-smoker or non-smoker? the validity of routinely recorded smoking status in uk primary care: a cross-sectional study. *BMJ open*, *4*(4), e004958.

McCormick, C., & Ryan, N. (2019). *Bert fine-tuning tutorial with pytorch.* (`https://mccormickml.com/2019/07/22/BERT-fine-tuning/`)

Medisch Contact. (n.d.). *Elektronisch patiëntendossier.* `https://`

www.medischcontact.nl/nieuws/dossiers/dossier/elektronisch-patientendossier.htm. ((Accessed on 04/08/2020))

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Ng, A. (2020). *Machine learning | coursera*. https://www.coursera.org/learn/machine-learning. ((Accessed on 04/08/2020))

Ngai, E. W., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision support systems*, *50*(3), 559–569.

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future—big data, machine learning, and clinical medicine. *The New England journal of medicine*, *375*(13), 1216.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

Palmer, E. L., Hassanpour, S., Higgins, J., Doherty, J. A., & Onega, T. (2019). Building a tobacco user registry by extracting multiple smoking behaviors from clinical notes. *BMC medical informatics and decision making*, *19*(1), 141.

Partnership Stop met Roken. (2019). *Zorgstandaard Tabaksverslaving*. http://trimbos-assets.e-vision.nl/docs/e61c8ed9-50d3-4e79-aa2b-5f1f8e0821ad.pdf. ((Accessed on 04/08/2020))

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations..

Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., & Re, C. (2017). Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the vldb endowment. international conference on very large data bases* (Vol. 11, p. 269).

Ratner, A., De Sa, C. M., Wu, S., Selsam, D., & Re, C. (2016). Data programming: Creating large training sets, quickly. In *Advances in neural information processing systems* (pp. 3567–3575).

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215.

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune bert for text

classification? In *China national conference on chinese computational linguistics* (pp. 194–206).

Suster, S., Tulkens, S., & Daelemans, W. (2017). A short review of ethical challenges in clinical natural language processing. *First Workshop on Ethics in Natural Language Processing (EACL)*.

Text Mining Research Leiden. (n.d.). *Textdata.nl :: Language resources for all.* `http://textdata.nl/`. ((Accessed on 04/09/2020))

The Pandas Development Team. (2020, February). *Pandas.* Zenodo. doi: 10

Tulkens, S., Šuster, S., & Daelemans, W. (2019). Unsupervised concept extraction from clinical text through semantic composition. *Journal of biomedical informatics*, *91*, 103120.

Uzuner, Ö., Goldstein, I., Luo, Y., & Kohane, I. (2008). Identifying patient smoking status from medical discharge records. *Journal of the American Medical Informatics Association*, *15*(1), 14–24.

van Smeeden, M. (2020, Feb). *Table Terminology.* `https://twitter.com/MaartenvSmeden/status/1225055692828160002`. Twitter. ((Accessed on 07/09/2020))

Vries, W. d., Cranenburgh, A. v., Bisazza, A., Caselli, T., Noord, G. v., & Nissim, M. (2019, December). BERTje: A Dutch BERT Model. *arXiv:1912.09582 [cs]*. Retrieved from `http://arxiv.org/abs/1912.09582`

Wang, Y., Sohn, S., Liu, S., Shen, F., Wang, L., Atkinson, E. J., ... Liu, H. (2019). A clinical text classification paradigm using weak supervision and deep representation. *BMC medical informatics and decision making*, *19*(1), 1.

Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., & Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PloS one*, *12*(4).

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... Rush, A. M. (2019). Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, *abs/1910.03771*.

Young-Wolff, K. C., Klebaner, D., Weisner, C., Von Korff, M., & Campbell, C. I. (2017). Smoking status and opioid related problems and concerns among men and women on chronic opioid therapy. *The Clinical journal of pain*, *33*(8), 730.

Zhao, J., Gu, S., & McDermaid, A. (2019). Predicting outcomes of chronic kidney disease from emr data based on random forest regression. *Mathematical biosciences*, *310*, 24–30.

Zorginstituut Nederland. (n.d.). *Ketenzorg (zvw) | verzekerde zorg | zorginstituut nederland.* `https://www.zorginstituutnederland.nl/Verzekerde+`

zorg/ketenzorg-zvw. ((Accessed on 04/09/2020))