Simo Ihalainen

Automatic training data labeling for Finnish clinical narrative NLP tasks

Master's thesis of mathematical information technology

January 5, 2022

University of Jyväskylä

Faculty of Information Technology

Author: Simo Ihalainen

Contact information: sami.ayramo@jyu.fi

Supervisors: Sami Äyrämö, Toni Ruohonen, Miika Moilanen

Title: Automatic training data labeling for Finnish clinical narrative NLP tasks

Työn nimi: Suomenkielisten potilaskertomusten automaattinen opetusdatan luonti NLP-malleja varten

Project: Master's thesis

Study line: Teknis-matemaattinen mallintaminen ja päätösanalytiikka

Page count: 43 + 0

Abstract: Large amounts of patient data is stored in electronic health records in unstructured data form as clinical narratives. The efficient use of clinical narratives in day-to-day care and clinical research requires advanced natural language processing methods to extract data from the texts. The common problem for many deep learning algorithms is the requirement for vast amounts of labeled training data, which is time consuming and expensive to acquire in the clinical narrative context. The purpose of this thesis was to assess a weak supervision based approach in automatic training data labeling, and the subsequent machine learning model performance in classifying two medical risk factors in Finnish language clinical narratives: high cholesterol and alcohol consumption. Heuristic rules were developed to automatically label sentences collected from clinical narratives to create a training dataset. Different machine learning models were trained with automatically labeled training dataset and with 200 manually labeled sentences. BERT model achieved the highest overall classification accuracy of 94 % in cholesterol task and 91 % in alcohol task. BERT model was able to capture hidden patterns in the data and leverage the natural language understanding to produce better classification results and classify cases which were not captured by the rules used to create the training data. All machine learning models trained with the automatically labeled data produced better classification results compared to the models trained with a small manually labeled dataset. Weak supervision approach might be a valuable tool to reduce the

costs of applying machine learning algorithms in low-resource settings, where manual labeling process is time consuming, expensive, or requires the expertise of subject specialist.

Keywords: Natural language processing, clinical narratives, text analytics, medical risk factors, weak supervision, automatic training data labeling

Suomenkielinen tiivistelmä: Terveydenhuollossa suuri määrä dataa on tallennettuna elektronisiin potilastietojärjestelmiin potilaskertomusten muodossa. Potilaskertomustekstien tehokas hyödyntäminen päivittäisessä hoitotyössä ja kliinisessä tutkimuksessa vaatii edistyneiden luonnollisen kielen käsittelyalgoritmien käyttöä oleellisen data poimimiseksi potilasteksteistä. Monet tähän tarkoitukseen soveltuvat koneoppimisen menetelmät vaativat suuria määriä luokiteltua opetusdataa käytettäväksi mallin koulutukseen, mikä on potilaskertomusten tapauksessa aikaa vievää ja kallista toteuttaa. Tämän opinnäytetyön tarkoituksena oli tutkia automaattista opetusdatan luokittelua ja automaattisesti luodulla opetusdatalla koulutettujen mallien suorituskykyä kahden lääketieteellisen riskitekijän (korkea kolesteroli, haitallinen alkoholinkäyttö) luokitteluun potilaskertomuksista. Kehitettyjen sääntöjen avulla luotiin automaattisesti luokiteltu opetusdatasetti, jota käytettiin eri koneoppimismallien kouluttamiseen. Samat mallit koulutettiin myös manuaalisesti luokitellulla 200 lauseen opetusdatasetillä. BERT-malli saavutti parhaan luokittelutarkkuuden sekä kolesterolin (94 %) että alkoholin (91 %) tapauksessa. BERT-malli pystyi hyödyntämään luonnollisen kielen ymmärrystä ja saavuttamaan paremman luokittelutarkkuuden kuin mihin opetusdatan luomiseen käytetyt säännöt pystyivät. Kaikki automaattisesti luodulla opetusdatalla koulutetut mallit pääsivät parempaan luokittelutarkkuuteen kuin mihin vastaavat pienellä manuaalisesti luokitellulla opetusdatalla koulutetut mallit pystyivät. Automaattinen opetusdatan luokittelu saattaisi olla arvokas työkalu koneoppimisprojektien kustannusten pienentämiseen tilanteissa, joissa opetusdatan manuaalinen luokittelu on aikaa vievää, kallista ja vaatii sovellusalan asiantuntijan työpanosta.

Avainsanat: potilaskertomus, tekstianalytiikka, lääketieteelliset riskitekijät, automaattinen opetusdatan luokittelu

Glossary

AUROC	Area under receiver operating curve
BERT	Bidirectional encoder representations from transformers
EHR	Electronic health record
ELMo	Embeddings from Language Model
FN	False negative
FP	False positive
NB	Naïve Bayes
NLP	Natural language processing
ROC	Receiver operating curve
SVM	Support vector machine
TF-IDF	Term frequency – inverse document frequency
TN	True negative
TP	True positive

List of Figures

List of Tables

Contents

1 Introduction

Electronic health records (EHRs) provide a clinically relevant data source which serves in two main functions. EHRs aid decision making processes in patient day-to-day point of care at the hospitals, providing health care professionals fast and reliable access to patient data. On the other hand, secondary use of EHRs data in clinical research can be utilized to improve quality of care and patient safety, optimize health care costs, and provide population level health statistics.

Clinical narratives have traditionally been the main form of communication within the health care system. Clinical narratives provide a convenient and concise summary of patient history for health care professionals, enabling them to make quick and informed decisions in patient care. However, patient data in clinical narratives is stored as unstructured text data, complicating the secondary use of patient data in research and development settings. During the recent years, immense efforts have been made to transform these unstructured data sources into structured ones.

In clinical narrative text analytics and data extraction, one frequently used technology has been an artificial intelligence subfield called natural language processing (NLP). NLP has shown great promise in various tasks including patient diagnosis extraction and tobacco and alcohol consumption identification. One of the downsides in the application of NLP models in clinical text analytics is the requirement for huge amounts of labeled training data. Training data for different clinical narrative NLP tasks is not readily available and using health care professionals for manual data labeling would be extremely expensive. Fortunately, the expensive training data labeling problem is present in many machine learning applications, and different solutions have been designed to tackle this issue.

The techniques addressing low resource machine learning scenarios, where labeled training data is scares, include weak supervision, data augmentation and active learning. In weak supervision, heuristic rules are utilized to automatically label vast amounts of training data. This noisy and imprecise training data (weak labels) is then used to train the final model. Data augmentation and active learning on the other hand rely on a small set of labeled training data. In data augmentation small transformations are made to the training data without

changing the labels, increasing the amount of unique training data samples. In active learning an iterative process is performed, where the learning algorithm can query the user to label new most relevant data points during the model training and gradually increase the model precision through each iteration.

The previously reported use cases in clinical narrative text analytics have covered mostly English language texts. No studies so far have reported the applicability of the low resource machine learning techniques (weak supervision, data augmentation, active learning) to Finnish language clinical narrative text analytics problems. A low resource solution to Finnish clinical narrative text analytics would enable more efficient use of existing medical records and provide researchers with new possibilities for clinical studies. Therefore, the purpose of this thesis was to assess the applicability of weak supervision, a low resource machine learning technique, for clinical narrative text analytics tasks.

2 Review of the literature

2.1 Electronic health records

Electonic health records (EHRs) have been widely adopted in hospitals as a means to store, search and utilize patient data. EHRs contain longitudinal information related to the health and healthcare of an individual. Both structured data (for example billing information, medications, diagnosis, and laboratory test results) as well as unstructured data (admission documents, nursing notes, clinical narratives) are stored in EHRs. (Koleck et al., 2019). The advantages of EHRs have been related to three different sections: clinical, organizational, and societal outcomes (Menachemi & Collum, 2011).

2.1.1 EHRs and clinical outcomes

Improvements in the quality of care and reductions in medical errors have been described as main clinical outcomes and benefits in adopting EHRs to hospital use (Menachemi & Collum, 2011). EHRs often include clinical decision support systems, automated reminders and computerized physician order entry systems, which have been shown to increase adherence to vaccination guidelines (Dexter et al., 2001; Ledwich et al., 2009), reduce redundant unnecessarily repeated laboratory tests (Bates, Kuperman, et al., 1999; Chen et al., 2003; Niès et al., 2010; Tierney et al., 1990; Wilson et al., 1982), reduce serious medication errors (Bates et al., 1998; Bates, Teich, et al., 1999; Devine et al., 2010), and increase the frequency of appropriate medication orders (Chertow et al., 2001).

In addition to the patient level clinical outcomes described above, studies have also shown that hospitals with greater investments in EHRs have lower mortality rates, fewer complications, and lower costs (Amarasingham et al., 2009; Menachemi et al., 2008). However, some studies have shown only small benefits or mixed results in adopting EHRs to hospital use (DesRoches et al., 2010; McCullough et al., 2017). One possible reason for the conflicting results might be the differences in EHR systems provided by different vendors (T. Wang & Gibbs, 2019).

2.1.2 EHRs and organizational outcomes

The benefits of organizational outcomes in adopting EHRs have included increased revenue, diminished costs, and improved job satisfaction among medical personnel (Schmitt & Wofford, 2002; S. J. Wang et al., 2003). The increased revenue in adopting EHRs has been attributed to accurate and timely patient charges and reductions in lost or disallowable bills (Agrawal, 2002; Erstad, 2003; Schmitt & Wofford, 2002). EHRs can also provide automatic reminders of routine health checks, leading to increased patient visits and revenues (Menachemi & Collum, 2011).

The diminished costs related to EHRs include decreases in staff resources dedicated to patient management, substitution of paper records and removal of supplies needed to maintain paper files, decreased transcription costs, and lower malpractice claims for physicians using an EHR compared to physicians without EHR (Ewing & Cusick, 2004; Miskulin et al., 2009; Tierney et al., 1993; Virapongse et al., 2008). For example, Tierney et al. (1993) reported $887 lower per admission charges in intervention teams using EHRs for inpatient order writing compared to control teams. This reduction was estimated annually to amount to savings over $3 million in the given hospital. Another study showed nearly a 50% decrease in time spent by the dialysis unit staff in patient anemia management, when an EHR based decision support system was adopted (Miskulin et al., 2009). Regarding physician malpractice claims, Virapongse et al. (2008) reported that 6.1% of physicians with an EHR had malpractice claims compared to 10.8% of physicians without an EHR.

2.1.3 EHRs and societal outcomes

The societal benefits of EHRs have been attributed to improved research opportunities, aggregated population level health statistics, monitoring and detection of disease outbreaks and biological threats, and job and career satisfaction among the hospital personnel (Aspden et al., 2004; Elder et al., 2010; Kukafka et al., 2007; Menachemi et al., 2009). The data availability and standardization offer many research opportunities, such as identifying evidence based best practices in patient care, providing data for national and international health surveys, as well as recruitment of suitable patients matching the acceptance criteria for different clinical trials (Kukafka et al., 2007). It is also worth noting the improvement in physicians'

job and career satisfaction through the adoption of EHRs, since physician career satisfaction has been associated with improved quality of care (Elder et al., 2010; Linzer et al., 2000; Menachemi et al., 2009; Pathman et al., 1996).

2.1.4 Disadvantages related to EHRs

In addition to the benefits of EHRs listed in many publications, a number of disadvantages and problems have been reported with the adoption of EHR use. These issues include financial issues, loss of productivity related to EHR implementation, unfavorable changes in previously established workflows, and patient privacy and data security (Agrawal, 2002; Schmitt & Wofford, 2002; S. J. Wang et al., 2003; Westin, 2005; Zurita & Nøhr, 2004).

In addition to the aforementioned issues, the data format commonly used in day-to-day patient care has not been designed to be used in research settings. A large amount of clinical information is stored in the form of clinical narratives. It has been estimated that 80 percent of currently available health data is in the form of unstructured data sources (Martin-Sanchez & Verspoor, 2014), such as patient clinical history, description of present illness, narrative description of physical exam findings, radiology reports, and operative reports. Unstructured data sources present a major challenge for the use of EHR data in clinical research settings. Regarding this issue, many different natural language processing (NLP) approaches have been developed to extract information from clinical narratives in order to be used in clinical research (Juhn & Liu, 2020).

2.2 Natural language processing

Natural language processing (NLP) is a subfield of artificial intelligence and machine learning, encompassing a wide variety of methods and algorithms to process and analyze natural language data. The purpose of NLP is to understand the content and context of natural language data sources, and then apply this understanding to for example information extraction or categorization of the data sources. A typical NLP task involves mapping textual data into vector representation (feature extraction) followed by a suitable model which takes the natural language feature representations as input and produces the desired output, for example classifies text into predefined categories. (Goldberg, 2017)

2.2.1 Feature extraction

Feature extraction in NLP refers to the process of transforming textual data into real valued vectors, i.e. feature representations of the data. The feature extraction process varies depending on the use case and the model selected to achieve the NLP task in question, but common feature extraction methods include tokenization, stopword removal, stemming, lemmatization, and word or sentence vectorization.

Tokenization

The tokenization process refers to splitting the text into pieces called tokens, and possibly removing some characters, such as punctuation, from the text data used in subsequent processing steps. The purpose of the tokenization process is to produce semantically useful units encompassing a sequence of characters grouped together. In the simplest form this tokenization process is achieved by splitting the text at whitespace and removing all punctuations from the text. (Manning et al., 2008). Many NLP programming libraries provide readily available tokenizers to be used with different machine learning models.

Stopword removal

The stopword removal is designed to remove extremely common words from the text, which have little or no value to the NLP task in question. Usually, a list of stopwords is provided, and all instances of these stopwords are removed from the tokens used for subsequent feature extraction tasks. The stopword list might include words like {*me, to, from, the, and, not*}, which occur frequently in all texts, but provide no relevant information to for example classifying texts into different categories. This process helps to reduce the size of the vocabulary used in the machine learning model training phase and thus reduce the computational requirements in the later stages of the NLP task. (Manning et al., 2008). However, caution should be taken in selecting the appropriate stopword list, since some NLP tasks might be heavily reliant on common stopwords to distinguish the semantic meaning in a text. A common stopword list used in Natural language toolkit (https://www.nltk.org/) would reduce the sentence “I am not happy” into “happy”, which would make NLP tasks such as sentiment analysis impossible.

Stemming and lemmatization

The stemming and lemmatization processes aim to transform words with semantically similar meaning into one common base form without changing the meaning of the sentence. For grammatical reasons, natural language text contains various forms of the same word, such as look, looked, and looking. All these words have a similar meaning easily understood by a human reader. However, from the machine learning model perspective, all the different forms of a word with similar semantic meaning only increases the difficulty for the model to interpret natural language. This is why stemming or lemmatization process is applied to derive the base forms for words. The choice between stemming and lemmatization as a processing method depends on the use case and the available resources. Lemmatization requires morphological analysis to accurately identify the correct base form (lemma) for each word, which might not be available for a given language. (Manning et al., 2008).

Stemming and lemmatization differ in the process of how words are transformed. Stemming is a simpler approach, where the word prefixes and suffixes are removed, leaving only the word base form, stem, to be used in the following data transformations. Lemmatization on the other hand tries to transform words into their dictionary form. (Goldberg, 2017; Manning et al., 2008). In the case of the words *look, looked, and looking*, both stemming and lemmatization processes would yield the same results, *look*, for all of these words. However, the word *flew* might not be changed at all by the stemming process, whereas lemmatization would yield *fly* as a lemma for *flew*.

Word, sentence, and document vectorization

After all the previously mentioned natural language processing steps, the data is still in text form and has to be transformed into numerical vector representation (feature vector) in order to be used in a machine learning model. This transformation can be achieved with many different methods, such as bag of words, n-grams, term frequency-inverse document frequency (TF-IDF), and word or sentence embedding (Liang et al., 2017).

Bag of words is a commonly used method in text classification to produce feature vectors representing the documents to be classified. In bag of words a feature vector is calculated by summing the word occurrences for all the different words in a text or collection of texts and combining these counts into a vector. In this approach the exact ordering of the words is ignored, and only the number of occurrences of each word is preserved. (HaCohen-Kerner

et al., 2020). This means that the semantic meaning is lost in the process, since bag of words would represent the texts “Usain Bolt is faster than John” and “John is faster than Usain Bolt” identically. However, for the text classification task two documents with similar bag of words representations are usually also similar in content. (Manning et al., 2008). An extension of the bag of words approach is n-gram, where in addition to single word counts, also word pair counts (bigrams), triplet counts (trigrams), and up to n consecutive word counts are calculated and included in the feature vector (Majumder et al., 2002).

One critical flaw in bag of words approach is that it considers all terms as equally important. However, for text classification tasks many words do not have any discriminating power in determining the class of a document. These would be words which appear in almost all documents. To tackle this issue, TF-IDF method weights each term frequency in a given document by the inverse of the document frequency (IDF) for the corresponding term. IDF is calculated as a logarithm of the total number of documents divided by the number of documents where the term is present. This weighing of the term frequencies produces highest scores for those terms which occur many times within the given document but only in small number of other documents, thus implying a strong discriminatory power of the given term. (Manning et al., 2008).

All the feature extraction methods presented above are well suited for text classification tasks but fail to represent syntactic and semantic meaning of the text. Word embeddings are a form of word representations in N-dimensional vector space, where words whit similar semantic meanings and context the words appear in are close to each other. Many different approaches have been developed to achieve this word embedding goal, such as continuous bag of words, skip-gram, and deep contextualized models. (B. Wang et al., 2019). A widely used word embedding algorithm is word2vec, which utilizes continuous bag of words or skip-gram to produce the word embeddings (Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013). During the recent years, huge advances have been made in deep learning, and models such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) and Embeddings from Language Model (ELMo) (Peters et al., 2018) have been developed to produce context sensitive word embeddings.

2.2.2 Common machine learning models in NLP classification tasks

After the textual data has been transformed into numerical feature vectors, the choice of machine learning model depends on the task at hand. The different machine learning algorithms applicable to NLP classification tasks include Naïve bayes (NB), support vector machine (SVM), decision tree and random forest, k-nearest neighbors, k-Means, and various deep neural networks.

Naïve bayes classifier is a probabilistic classifier, which learns the conditional posterior probability of different classes with a given feature vector. Bayes rule is used to calculate the conditional posterior probabilities. The model assumes that the features are conditionally independent. In practice this assumption is frequently violated since features are usually dependent. However, the resulting model is easy to fit and works well in many different classification tasks. Another advantage of NB classifier is the low computational requirement during inference, i.e. classifying new data samples. (Murphy, 2006).

SVM classifier is based on the construction of a hyper plane which separates datapoints belonging to different classes. Many different hyperplanes might separate the classes, so the optimal hyperplane which produces the maximal margin to positive and negative class examples is used in the SVM model. This means that the position of the optimal hyperplane is determined by the few examples closest to the hyperplane. (Bottou & Lin, 2007). When the original datapoints are not linearly separable, a kernel function is used to map the datapoints (feature vectors) into higher dimensional space, where the different classes are linearly separable (Pradhan, 2012). In the case of noisy datasets with non-linearly separable classes, a soft margin SVM which allows for some datapoints to be classified incorrectly is recommended instead of complicated high dimensional kernel transformations (Bottou & Lin, 2007).

BERT model has been one of the most influential deep learning models which has been used in many different NLP tasks with superior results compared to the previously published models. At the time of its release, BERT reached state-of-the-art results in 11 natural language processing tasks, such as question answering. The same model architecture can be used for various different NLP tasks with only minor variations in the output layer. (Devlin et al., 2018). The initial BERT model was for English language, but later pre-trained BERT

models have been presented for various other languages including Finnish (Virtanen et al., 2019).

The BERT model architecture is presented in Figure 1. The model architecture is a bidirectional transformer encoder, which produces token embeddings as output. There are two steps in BERT model training: unsupervised pre-training and task specific fine-tuning. The pre-training phase consisted of masked token prediction, where 15% of the tokens were randomly masked, and the task for the model was to predict these masked words based on the surrounding unmasked words. In the second phase of the pretraining, sentence pairs were formed from the training data, where 50% of the sentence pairs were consecutive sentences, and 50% of the sentence pairs were not consecutive sentences. The model task was to classify the sentence pairs as consecutive or not consecutive sentence pairs. Books corpus and English Wikipedia were used as unlabeled data for both masked token and next sentence prediction pre-training tasks (Devlin et al., 2018). The data input embedding process is shown in Figure 2.

The second step in BERT model training is fine-tuning. During fine-tuning, a smaller dataset with task specific labels is used to fine-tune the pre-trained model parameters to suit the specific NLP task at hand. The base model architecture during fine-tuning is identical to the pre-training phase. Only an additional output layer (classification head) is added to the base model to achieve the desired BERT model task at hand, and this output layer differs between different types of NLP tasks. (Devlin et al., 2018).

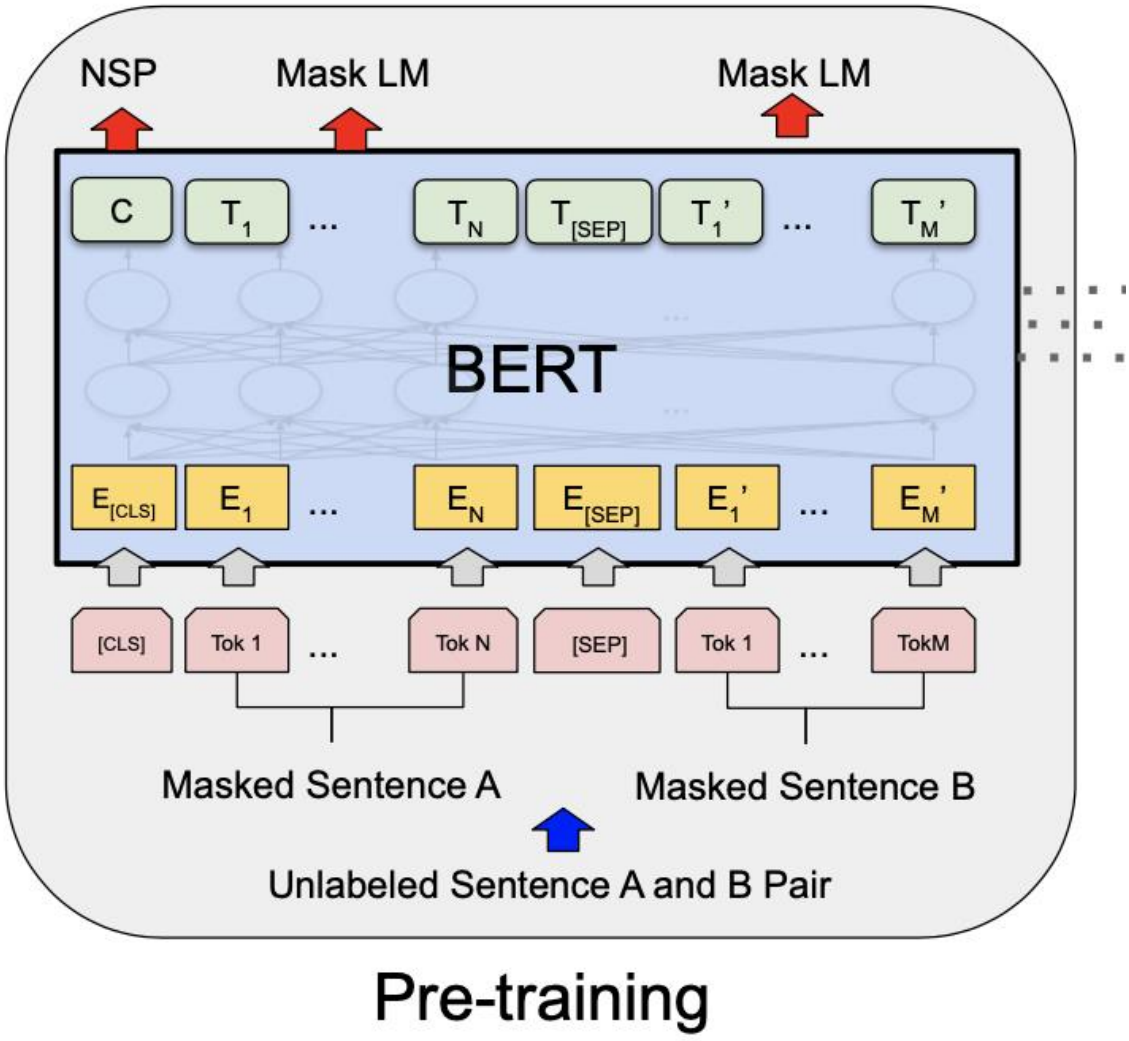

Figure 1. BERT model architecture (modified from Devlin et al., 2018).

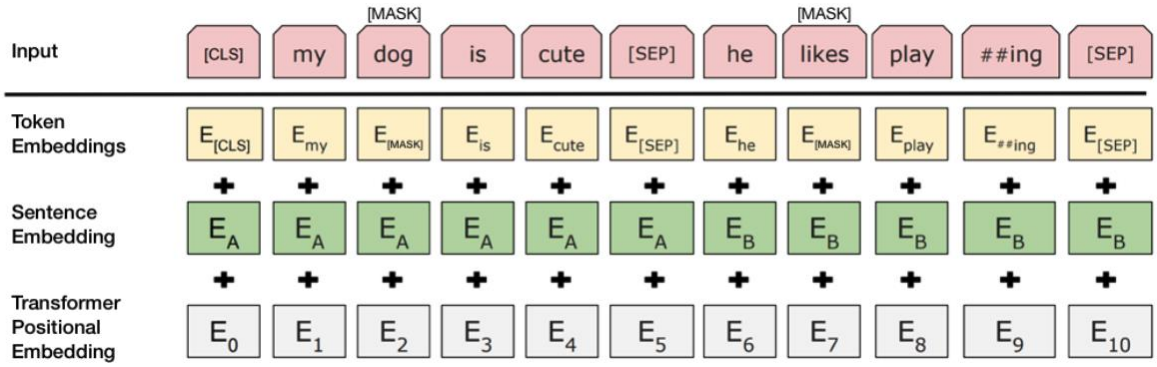

Figure 2. BERT model embeddings for input tokens (Devlin et al., 2018).

2.2.3 NLP in clinical narrative information extraction tasks

Several examples of NLP applications in clinical narrative information extraction tasks are present in the literature. NLP has been used to extract cancer stage information (Cheng et al., 2010), disease characteristics (Zhu et al., 2012), pathological conditions (Séverac et al., 2015), and medical risk factors from clinical narratives. Wang et al. (2019) presented weak supervision based model results for SVM, random forest, multilayer perceptron neural network, and convolutional neural network in classifying smoking status and proximal femur fracture in clinical narratives. Their unique approach was to use rule-based training data labeling to automatically produce weak labels for vast amounts of training data, and then use this weakly labeled training data to train different machine learning models. The binary classification cases presented in the study showed that convolutional neural network achieved higher classification accuracies compared to rule-based classification method or other machine learning models studied. Convolutional neural network binary classifier precision, recall, and F1-score achieved were 0.93, 0.92, and 0.92 for smoking status classification, and 0.97, 0.97, and 0.97 for proximal femur fracture classification. (Y. Wang et al., 2019). A previous study without deep learning NLP methods reported lower precision, recall and F1-scores of 0.86, 0.87, and 0.87, respectively, for the smoking status classification (Khalifa & Meystre, 2015).

A more conventional smoking status classification NLP approach was used by Karlsson et al. (2021), who created a training dataset by manually labeling 5000 Finnish language medical narrative smoking related sentences into never, former and current smoker categories. A BERT model pretrained in Finnish language was fine-tuned with the manually labeled training dataset. This BERT model reached 88.2% total classification accuracy for never, former, and current smoker, with 96%/96%, 96%/73%, and 90%/97% specificity and sensitivity values for the classes, respectively. (Karlsson et al., 2021).

In many classification cases, deep learning models have shown superior performance compared to older traditional methods. However, in the case of identifying sudden cardiac death risk factors (syncope, family history of sudden cardiac death, or family history of hypertrophic cardiomyopathy) from clinical narratives, traditional rule-based NLP methods reached sensitivity and specificity values between 0.9 and 0.98 (Moon et al., 2019). The modern deep learning methods are often complex and computationally expensive, and one

should carefully select the appropriate method to meet the demands of the task in question. In the case of identifying sudden cardiac death risk factors, the accuracy acquired with the traditional rule-based method might be high enough to meet the demands of the task in question, and the use of more complex and computationally expensive methods would be useless.

One commonly faced problem in applying deep learning to clinical narrative NLP tasks is the annotation bottleneck (Spasic & Nenadic, 2020). Supervised learning requires large amounts of labeled training data, and in clinical domain the annotation process is harder than on many other application fields. Firstly, there are no publicly available datasets because of patient privacy concerns. Secondly, the annotation process requires medical expertise to correctly classify the clinical narratives, making the annotation process expensive and time consuming. The privacy concerns and requirements for medical expertise also prevent the use of crowdsourcing as a means to produce large amounts of training data. Lastly, the training data is in many cases restricted to the use case at hand and cannot be used for other NLP tasks. (Spasic & Nenadic, 2020; Y. Wang et al., 2019). The above-mentioned problem of the lack of labelled training data has been identified on many different application fields, and different solutions have been designed to tackle this issue.

2.3 Weak supervision and training data labeling in low resource settings

Low resource settings in NLP context refer to the task of carrying out NLP model training in a specific language or domain, where there are no parallel corpora, extensive monolingual corpora, annotated data or existing NLP tools available. In situations where the small amount of available labeled training data is the cause for low resource settings, methods such as active learning, data augmentation and weak supervision have been proposed as solutions to the problem. In active learning the amount of labeled training data required for model training is reduced by training the model iteratively and allowing the active learning training algorithm to query new labels from human labeler for the most informative data points (Settles, 2009). Data augmentation is the process of constructing synthetic data from the available dataset by introducing small changes in the data (Shorten et al., 2021). In case of labeled text data, the data augmentation process should keep the semantic meaning of the sentence intact, while introducing variation to the training dataset with such methods as synonym replacement, back translation, and spelling mistake insertions.

In addition to data augmentation and active learning, weak supervision has emerged as one promising technique to automatically create weakly labeled training datasets. In weak supervision, heuristic rules are generated to automatically label large volumes of training data, producing noisy and imperfect labels. Weak supervision and deep learning approaches have produced good results with medical narrative smoking status classification, hip fracture classification, and suicidal ideation identification (Cusick et al., 2021; Y. Wang et al., 2019).

One well documented library for applying weak supervision and heuristic rules on the training data is Snorkel (https://www.snorkel.org/). Snorkel has been used in many scientific publications (Bach et al., 2017, 2019; Callahan et al., 2019; Hancock et al., 2018; A. Ratner et al., 2016, 2017, 2017, 2018; Varma et al., 2019; Wu et al., 2017). The workflow in Snorkel encompasses the generation of labeling functions. Labeling functions apply heuristic rules, for example regex functions to the training data sample, and vote for a label to be assigned to the data sample. Labeling functions might also abstain from voting, when the data sample in question does not match the specified labeling function rules. The votes from all the labeling function are aggregated through a majority voter or probabilistic label model to produce the final weak labels for the training dataset.

Weak supervision can be used in the scenarios where vast amounts of unlabeled training data is available. In these scenarios, similar or better results compared to weak supervision can be achieved by manually labeling more training data. In a topic classification task, 85000 manually labeled training data samples were required to reach the same accuracy as weak supervision process with 684000 unlabeled data samples was able to produce (Figure 3). In a user study comparing the weak supervision process to hand labeled training data, participants with various backgrounds in machine learning and coding were able to produce more accurate models in 2.5 hours developing and applying Snorkel labeling functions and weak supervision, compared to models trained on 7 hours (2500 instances) of manually labeled data (A. Ratner et al., 2017).

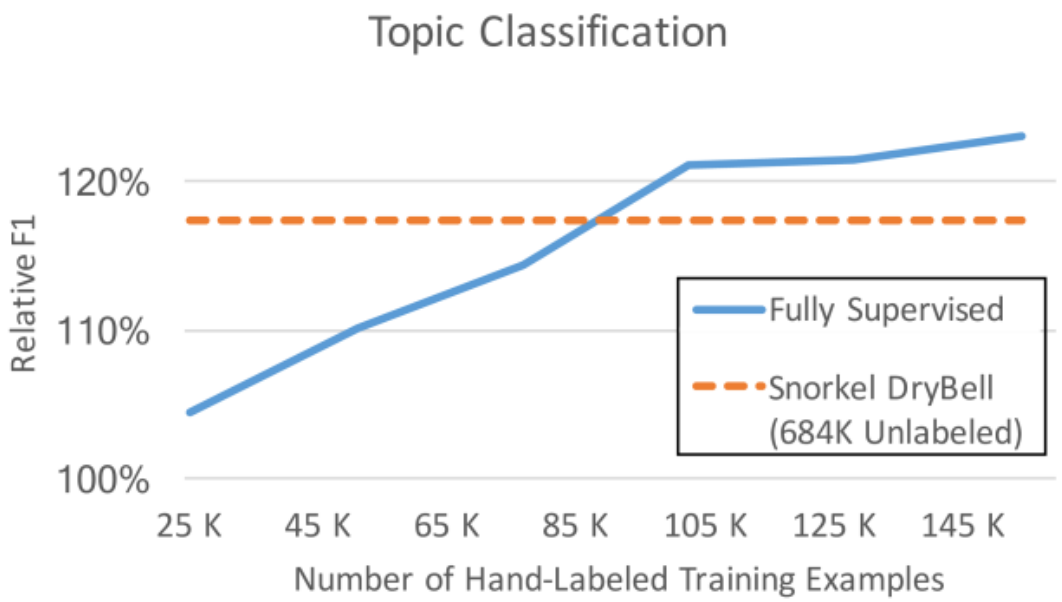

Figure 3. Number of manually labeled training examples required to reach the accuracy achieved by weak supervision approach (Bach et al., 2019).

3 Research questions

Electronic health records contain vast amounts of patient medical information in the form of unstructured clinical narratives. Efficient use of this data source both at the point of care in hospitals as well as secondary use in medical research require advanced automatic text analytics methods to extract relevant information from the clinical narratives. Acquisition of accurate labeled training data in medical narrative text analytics tasks is time consuming and expensive and has been identified as one major bottleneck in the application of NLP algorithms for clinical narrative text analytics tasks. Weak supervision based approaches have been applied successfully to English language clinical narrative NLP tasks, greatly reducing the requirements for labeled training data. However, no study so far has investigated the use of weak supervision in the context of Finnish language clinical narrative NLP tasks. Therefore, the purpose of this thesis was to assess weak supervision based approaches in Finnish clinical narrative text analytics tasks. More specifically, this thesis is focused on automatic rule-based training data labeling methods and subsequent NLP model performance in labeling two common medical risk factors in clinical narratives: high cholesterol and alcohol consumption. The research questions were formulated as follows:

1) What is the machine learning model performance with manually labeled training data compared to automatically labeled training data in classifying medical risk factors in Finnish language clinical narrative sentences?
2) What is the classification accuracy achieved with automatic training data labeling?

4 Methods

4.1 Train, test and validation datasets

4.1.1 Cholesterol

The dataset for high cholesterol labeling task was collected by combining 2000 sentences in clinical narratives containing 1) “kolesterol” or “cholesterol”, 2) "hyperkolesterol", "hyperlipid" or "dyslipid", 3) “rasva-arvo”, 4) “lipid”, 5) “ldl”, “hdl” or “trigly, 6) list of cholesterol medication names (collected from https://sydan.fi/fakta/kolesterolilaakkeet/), and 7) none of the above mentioned words. Combining 7 categories of 2000 sentences yielded a total of 14000 sentences as the total dataset used.

The collected dataset was randomly sampled into training and validation datasets with 95%-5% split, yielding a cholesterol training dataset of 13300 sentences and a validation dataset of 700 sentences. The validation dataset was not used or viewed until all the models had been trained, and validation dataset was only used to assess the generalizability of the models. The training dataset was further randomly sampled into train and test datasets (70%-30% split with NB and SVM, 90%-10% split with BERT).

4.1.2 Alcohol consumption

The dataset for bad alcohol consumption labeling task was collected by combining 3000 sentences in clinical narratives containing 1) “alkohol”, 2) " viina”, “olut”, “olue”, “kalja”, “viini”, “lonkero”, “siideri”, “konjak”, or “viski", 3) “juoppohullu”, “delirium tremens”, or “alkodelirium”, 4) “humala”, “humaltun”, “päihdyksissä, “päihtynyt”, or “krapula”, 5) none of the above-mentioned words. Combining 5 categories of 3000 sentences yielded a total of 15000 sentences as the total dataset used.

The collected dataset was randomly sampled into training and validation datasets with 95%-5% split, yielding an alcohol training dataset of 14250 sentences and a validation dataset of 750 sentences. The training dataset was further randomly sampled into train and test datasets (70%-30% split with NB and SVM, 90%-10% split with BERT).

4.2 Automatic training data labeling

4.2.1 Cholesterol

Snorkel (https://github.com/snorkel-team/snorkel), an open-source weak supervision framework was used to create labeling functions to automatically label training data with regex based heuristic rules. Labeling function regex rules were developed based on example sentences drawn from the training data set in order to capture different ways of expressing high cholesterol, normal cholesterol, and no mention of cholesterol levels. A total of seven labeling functions were created, where each labeling function voted for bad cholesterol or good cholesterol label to be assigned for the sentence or abstained from voting if the regex rule conditions were not met for the particular labeling function and sentence in question. Snorkel majority label voter was then used to aggregate all labeling function votes and assign final labels for the training data set. In line with the purpose of this thesis, the automatically labeled training data set was converted into a binary dataset by combining the "normal cholesterol" and "no mention of cholesterol levels" into one class, namely "not bad cholesterol".

4.2.2 Alcohol consumption

Similar process was carried out with alcohol consumption as described in the case of automatic cholesterol labeling. Snorkel labeling function regex rules were developed based on example sentences drawn from the training data set in order to capture different ways of expressing alcohol overuse, normal alcohol consumption, and no mention of alcohol consumption. A total of five labeling functions were created, and a Snorkel majority label voter was used to aggregate the labeling function votes as in the case of automatic cholesterol labeling. Lastly, the automatically labeled training data set was converted into a binary dataset by combining the "normal alcohol consumption" and "no mention of alcohol consumption" into one class, namely "not bad alcohol consumption".

4.3 Manual training and validation data labeling

200 sentences were randomly sampled from the cholesterol and alcohol training datasets and manually labelled. Additional 200 sentences were randomly sampled from the cholesterol

and alcohol validation datasets and manually labelled. The labeler had no medical degree or medical background, which has to accounted for as a limitation of the study.

4.4 Data preprocessing

For the SVM and NB model training and evaluation pipelines, training, test and validation datasets were preprocessed with the same text processing pipeline. Non-character symbols except whitespaces were removed and all characters lower cased. Sentences were tokenized using whitespace as delimiter. Stopwords were removed using Natural language toolkit (NLTK, nltk.org) Finnish stopword list. Both stemmed and un-stemmed versions of data preprocessing pipelines were tested in model training and evaluation. Stemming was carried out with NLTK SnowballStemmer.

Different feature extractions were tested, where Ngrams in range of one to four were followed by either term frequeny (Tf) or term frequency-inverse term frequency (Tf-idf) calculations. The hyperparameter tuning for Ngram followed by Tf or Tf-idf vectorization was carried out using Sklearn.

The BERT model implementation was carried out with HuggingFace tranformers library and a pretrained Finnish language model (Virtanen et al., 2019). The data preprocessing for the BERT model was carried out with HuggingFace tokenizer corresponding to the selected Finnish language BERT model.

4.5 Model training

The model trainings consisted of NB, SVM, and BERT binary classifier model trainings with automatically labeled training datasets and manually labeled training datasets. For the NB and SVM models, hyperparameter tuning was carried out using 5-fold cross validation. For NB, alpha values from 0.009 to 0.03 were used for hyperparameter tuning. For SVM, alpha, loss, learning rate, and eta parameters were used in hyperparameter tuning. After the hyperparameter tuning, the models were trained with the whole training dataset (manually labeled or automatically labeled) using the best obtained hyperparameter values for each

model. The NB and SVM model hyperparameter tuning and training were implemented using Sklearn.

A pretrained Finnish BERT (Virtanen et al., 2019) was fine-tuned using HuggingFace pipeline for text classification with both automatically labeled training data and manually labeled training data. The manually labeled training data was randomly sampled into training and test datasets using a 70%-30% split. Because of GPU cost and time restrictions, no hyperparameter tuning was possible for the BERT model. The only regularization used was early stopping (test set classification accuracy with automatically labeled training data, test set loss with manually labeled training data) with early stopping patience set at 5 epochs (number of epochs tolerated without improvement in the early stopping metric). Training and evaluation batch sizes were set at 4, dictated by the available GPU cluster memory restrictions. Maximum number of training epochs was set at 100.

4.6 Model evaluation

Classification prediction accuracy, defined as the fraction of all correctly classified instances, was calculated for all models with training, test, and validation datasets (when applicable). Snorkel labeling functions did not have training and test datasets, so the classification prediction accuracy was calculated only for the validation dataset. NB and SVM models trained with manually labeled dataset did not have a test dataset, so the classification prediction accuracy was calculated only for the training and validation datasets.

Model precision was calculated for both positive and negative classes with true positive (TP), false positive (FP), false negative (FN), and true negative (TN) predictions as defined in equations 1 and 2. Model recall and F1 scores were calculated for both positive and negative classes as defined in equations 3, 4, and 5.

$$1)\ Precision = \frac{TP}{TP + FP}$$

$$2)\ Precision = \frac{TN}{TN + FN}$$

$$3)\ Recall = \frac{TP}{TP + FN}$$

$$4)\ Recall = \frac{TN}{TN + FP}$$

$$5)\ F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

The receiver operating characteristic (ROC) values and area under receiver operating characteristic curve (AUROC) were calculated using sklearn.metrics roc_curve and roc_auc_score functions.

5 Results

5.1 Cholesterol

Summary of binary classifier model performances in classifying medical narrative sentences into bad and not bad cholesterol classes is presented in Table 1. Bert models achieved highest validation data classification accuracies both among models trained with automatically labeled training data as well as among models trained with manually labeled training data. It is worth noting that Snorkel labeling functions achieved a 0.91 classification accuracy on the validation dataset, which tied as a second-best classification accuracy after Bert model trained with automatically labeled training data (classification accuracy 0.94).

Among the models trained with automatically labeled training data, Bert model achieved the highest validation data classification accuracy (0.94), followed by SVM (0.91) and Naïve Bayes (0.85). The receiver operating characteristic (ROC) curves and area under receiver operating characteristic curve scores (AUROC) for these models are presented in Figure 4.

The ROC curves and AUROC scores for models trained with manually labeled training data are presented in Figure 5. All models trained with manually labeled training data had lower validation data classification accuracies compared to the models trained with automatically labeled training data. The ROC curve comparison between the best model trained with automatically labeled training data and the best model trained with manually labeled training data is presented in Figure 6.

Table 1. Binary classifier model accuracy, precision, recall and F1-score for classifying cholesterol sentences (train, test, and validation datasets). Results are presented for models trained with automatically labelled training data (Auto) or manually labelled training data (Manual).

Model	Training data	Accuracy train-test-val	Class	Precision test-val	Recall test-val	F1-score test-val
Labeling functions		N/A-N/A-0.91	Bad chol	N/A-0.95	N/A-0.86	N/A-0.90
			Not bad chol	N/A-0.93	N/A-0.94	N/A-0.94
Naïve bayes	Auto	0.90-0.91-0.85	Bad chol	0.90-0.85	0.90-0.81	0.90-0.83
			Not bad chol	0.91-0.84	0.91-0.88	0.91-0.86
Naïve bayes	Manual	0.84-N/A-0.81	Bad chol	N/A-0.78	N/A-0.82	N/A-0.80
			Not bad chol	N/A-0.83	N/A-0.80	N/A-0.82
SVM	Auto	0.93-0.93-0.91	Bad chol	0.94-0.94	0.93-0.85	0.93-0.89
			Not bad chol	0.93-0.88	0.94-0.95	0.94-0.91
SVM	Manual	0.91-N/A-0.85	Bad chol	N/A-0.88	N/A-0.78	N/A-0.83
			Not bad chol	N/A-0.83	N/A-0.91	N/A-0.87
BERT	Auto	1.00-0.99-0.94	Bad chol	0.99-0.97	0.98-0.90	0.98-0.93
			Not bad chol	0.98-0.92	0.99-0.97	0.99-0.95
BERT	Manual	0.99-0.73-0.87	Bad chol	0.73-0.90	0.62-0.80	0.67-0.85
			Not bad chol	0.74-0.84	0.82-0.93	0.78-0.88

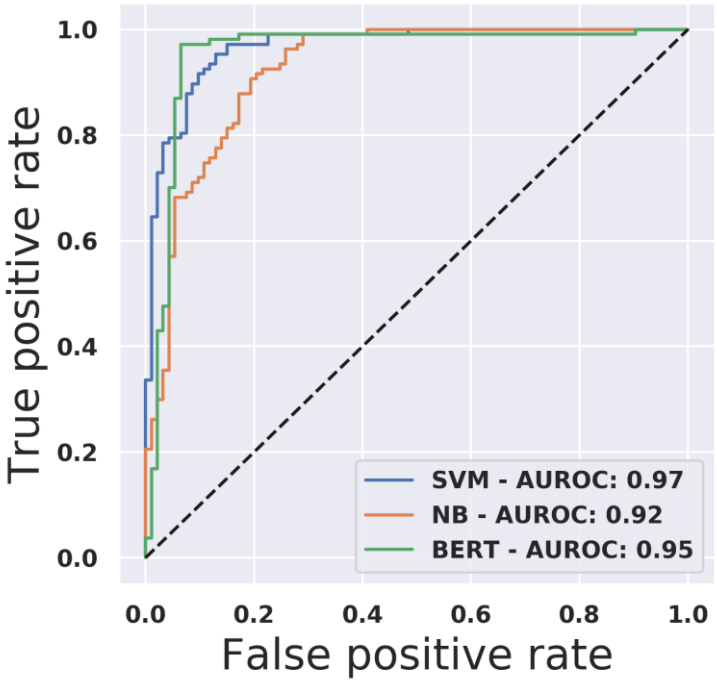

Figure 4. ROC curves for models trained with automatically labeled training data to classify bad cholesterol sentences.

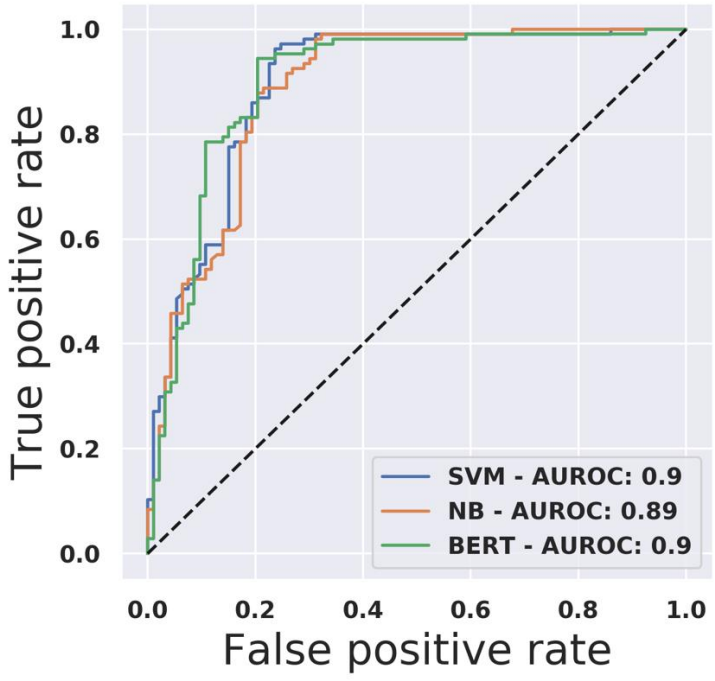

Figure 5. ROC curves for models trained with manually labeled training data to classify bad cholesterol sentences.

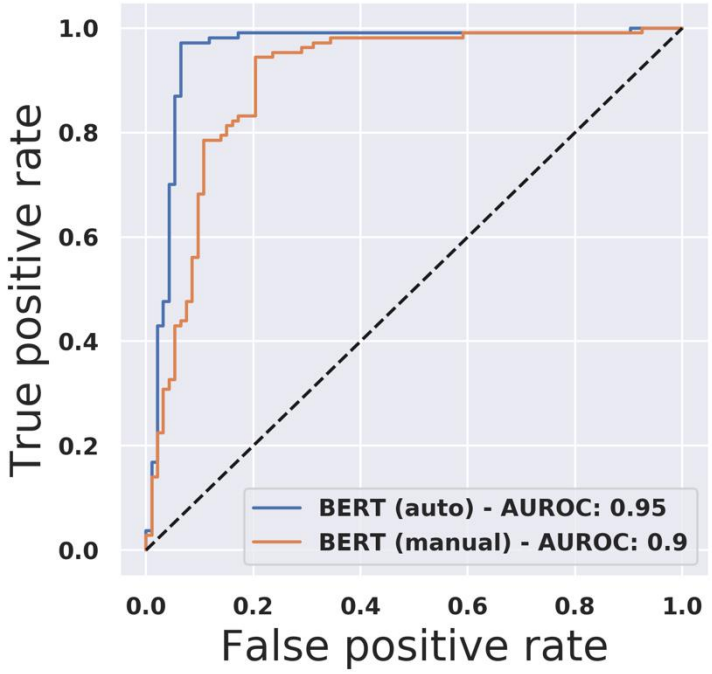

Figure 6. ROC curves for the best validation accuracy model trained with manually labeled training data (BERT (manual)), and best validation accuracy model trained with automatically labeled training data (BERT (auto)) to classify bad cholesterol sentences.

5.2 Alcohol consumption

Summary of binary classifier model performances in classifying medical narrative sentences into bad and not bad alcohol consumption classes is presented in Table 2. As in the case of cholesterol classification models, Bert models achieved highest validation data classification accuracies both among models trained with automatically labeled training data as well as among models trained with manually labelled training data. Again, Snorkel labeling functions achieved a 0.90 classification accuracy on the validation dataset, which tied as a second-best classification accuracy after Bert model trained with automatically labeled training data (classification accuracy 0.91).

Table 2. Binary classifier model accuracy, precision, recall and F1-score for classifying alcohol sentences (train, test, and validation datasets). Results are presented for models trained with automatically labelled training data (Auto) or manually labelled training data (Manual).

Model	Training data	Accuracy train-test-val	Class	Precision test-val	Recall test-val	F1-score test-val
Labeling functions		N/A-N/A-0.90	Bad alco	N/A-0.86	N/A-0.90	N/A-0.88
			Not bad alco	N/A-0.93	N/A-0.90	N/A-0.92
Naïve bayes	Auto	0.84-0.84-0.81	Bad alco	0.79-0.76	0.83-0.76	0.81-0.76
			Not bad alco	0.88-0.84	0.85-0.84	0.87-0.84
Naïve bayes	Manual	0.74-N/A-0.68	Bad alco	N/A-0.59	N/A-0.64	N/A-0.61
			Not bad alco	N/A-0.75	N/A-0.71	N/A-0.73
SVM	Auto	0.87-0.88-0.88	Bad alco	0.85-0.82	0.85-0.89	0.85-0.85
			Not bad alco	0.90-0.92	0.90-0.87	0.90-0.89
SVM	Manual	0.78-N/A-0.75	Bad alco	N/A-0.73	N/A-0.59	N/A-0.65
			Not bad alco	N/A-0.76	N/A-0.86	N/A-0.80
BERT	Auto	1.00-0.98-0.91	Bad alco	0.99-0.86	0.97-0.93	0.98-0.89
			Not bad alco	0.98-0.95	0.99-0.90	0.99-0.92
BERT	Manual	1.00-0.80-0.81	Bad alco	0.70-0.75	0.70-0.78	0.70-0.76
			Not bad alco	0.85-0.85	0.85-0.82	0.85-0.84

Among the models trained with automatically labeled training data, Bert model achieved the highest validation data classification accuracy (0.91), followed by SVM (0.88) and Naïve Bayes (0.81). ROC curves and AUROC scores for these models are presented in Figure 7.

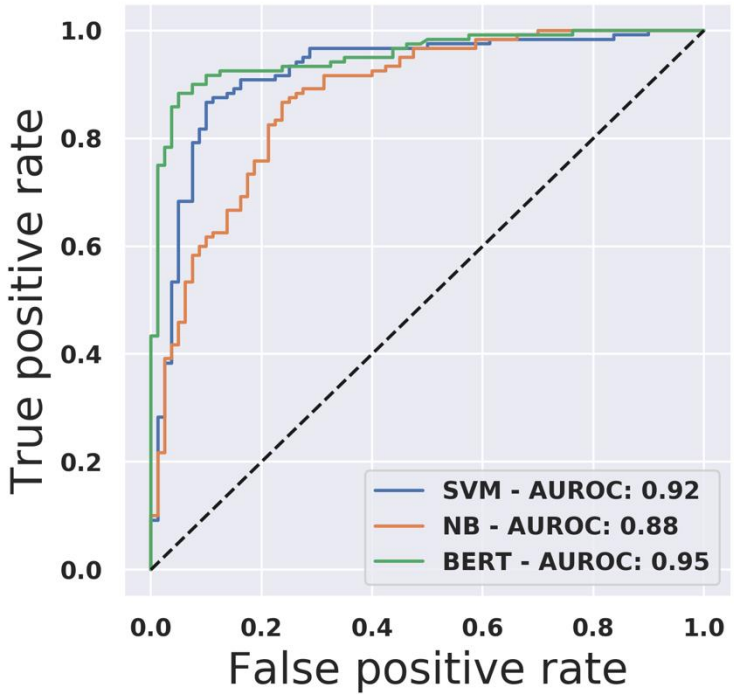

Figure 7. ROC curves for models trained with automatically labeled training data to classify bad alcohol consumption sentences.

The ROC curves and AUROC scores for models trained with manually labeled training data are presented in Figure 8. All models trained with manually labeled training data had lower validation data classification accuracies compared to the models trained with automatically labeled training data. The ROC curve comparison between the best model trained with automatically labeled training data and the best model trained with manually labelled training data is presented in Figure 9.

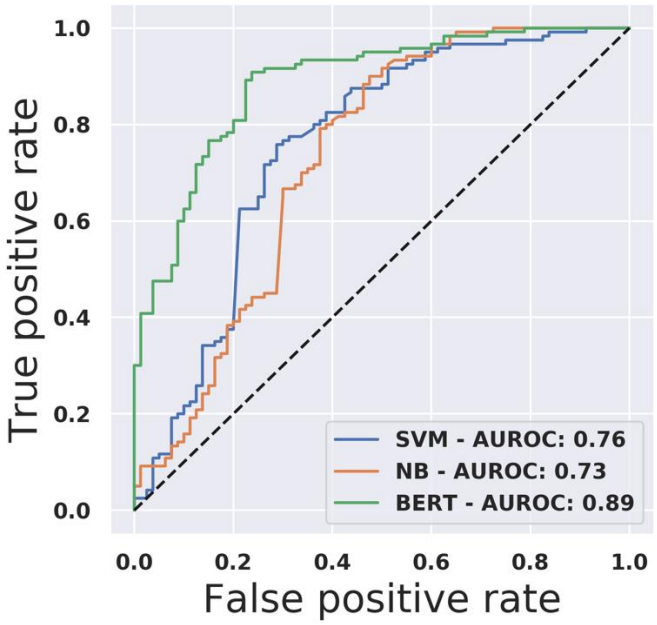

Figure 8. ROC curves for models trained with manually labeled training data to classify bad alcohol consumption sentences.

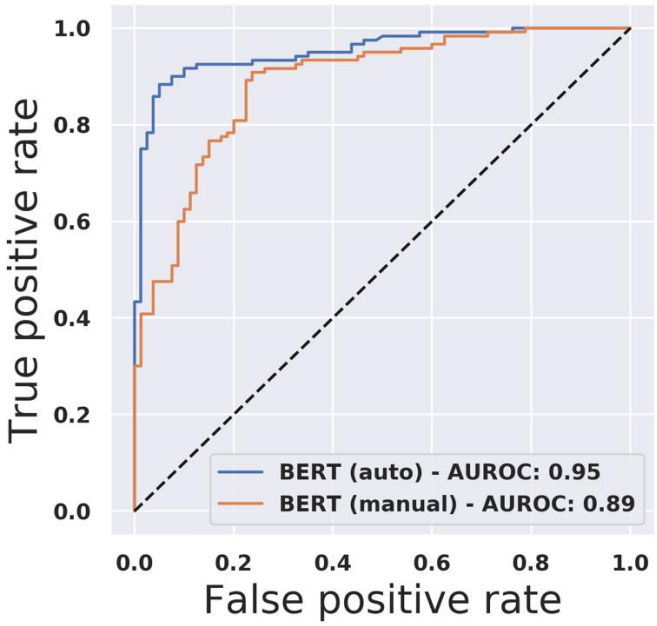

Figure 9. ROC curves for the best validation accuracy model trained with manually labeled training data (BERT (manual)), and best validation accuracy model trained with automatically labeled training data (BERT (auto)) to classify bad alcohol consumption sentences.

6 Conclusion

The purpose of this thesis was to evaluate and compare machine learning model performances in clinical narrative NLP classification tasks, when models were trained with either a weak supervision based approach using automatically labeled training data or manually labeled training data. Two NLP classification tasks of identifying medical risk factors in clinical narrative sentences were selected for this thesis: classifying sentences containing mentions of bad/high cholesterol level and excessive alcohol use. All tested machine learning models achieved higher classification accuracies for both cholesterol and alcohol classification tasks with automatically labeled training dataset compared to the training dataset of 200 manually labeled samples. The best classification accuracies achieved with automatically labeled training dataset were with BERT model, reaching 94 % overall classification accuracy for cholesterol and 91 % for alcohol.

The results of this thesis were in line with previously published results regarding weak supervision in clinical narrative NLP tasks. Wang et al. (2019) reported precision, recall, and F1-score values of 0.91, 0.91 and 0.91, respectively, for rule-based binary smoking status classification. For proximal femur fracture classification, the corresponding values were 0.97, 0.97, and 0.97. In the present study, the labeling functions (rule-based classification) achieved comparable precision, recall, and F1-score values of 0.95, 0.86, and 0.90 for cholesterol classification and 0.86, 0.90, and 0.88 for alcohol classification. In the present thesis the traditional simpler machine learning models SVM and NB reached the same or lower classification accuracies as the rule-based labeling functions, indicating no benefit in utilizing these machine learning models. This same finding was evident in the previous smoking status classification results, but not in the proximal femur fracture classification task (Y. Wang et al., 2019). Both the present thesis and the previous study by Wang et al. (2019) showed that machine learning models utilizing word embeddings could capture additional hidden patterns not presented in the rules used for automatic training data labeling, since BERT model in the current thesis and convolutional neural network used in the previous study were able to achieve higher classification accuracies compared to the rule-based classification.

The results of the present thesis showed that machine learning models trained with automatically labeled training data achieved 4-7 percentage points higher classification accuracies in cholesterol task and 10-13 percentage points higher classification accuracies in alcohol task compared to models trained with 200 manually labeled data samples. This result is not surprising, since a previous study showed that 85000 manually labeled data samples were required to reach similar classification accuracies in a topic classification task as acquired through weak supervision and automatic training data labeling (Bach et al., 2019). The requirement for the amount of manually labeled training data to achieve similar results compared to automatically labeled training data seems to be task specific, and the decision between investing in a manual labeling process or automatic rule-based process should be carefully analyzed.

In the present thesis, BERT model trained with automatically labeled training data reached highest overall classification accuracies for both cholesterol and alcohol classification tasks. The overall accuracy was 0.94 for cholesterol task, and the F1-score for bad cholesterol was 0.93. The overall accuracy for alcohol task was 0.91, and the F1-score for bad alcohol consumption was 0.89. In a previous study using readily available NLP packages and rule-based high cholesterol extraction, F1-score of 0.44 was reported, which is far lower compared to the F1-score for bad cholesterol classification in the present thesis (Khalifa & Meystre, 2015). Similar accuracy and F1-score values compared to the present thesis have been reported in previous studies regarding smoking status classification in clinical narratives with machine learning models. Palmer et al. (2019) used SVM to identify smoking status, reaching F1-score of 0.90. Another study used a BERT model pre-trained in Finnish and finetuned with 5000 manually labeled smoking-related sentences, reaching overall classification accuracy of 0.88 (Karlsson et al., 2021). And finally, a previous study using similar weak supervision approach as in the present thesis reached F1-score of 0.92 in smoking status classification (Y. Wang et al., 2019). Even though the classification tasks differ between all of these studies and direct comparison between the results is unwarranted, the results of the present thesis suggest that weak supervision and automatic training data labeling might be a valuable tool to reduce the costs of training data labeling in clinical narrative NLP tasks.

The weak supervision approach studied in the present thesis showed promising results in the two selected binary classification tasks. However, as noted also in a previous weak

supervision study (Y. Wang et al., 2019), it is still unclear how this automatic rule-based training data labeling approach would handle more complicated multiclass labeling tasks, and whether rules could even be constructed to meet the requirements for different types of NLP tasks. The application of weak supervision to more complex NLP tasks could be an interesting topic for future studies.

A number of limitations are present in the current thesis. Firstly, it should be noted that the models and the training and validation data did not classify numerical values and measurement results correctly. This was a conscious decision since language models such as BERT are ill suited to handle numerical values in classifications. The idea for a complete classification process was to combine BERT classifier with a rule-based classifier, which could be easily constructed to extract numerical measurement values and use threshold values to classify for example LDL cholesterol measurement values above a certain threshold to high cholesterol class.

Secondly, the validation data set was sampled from the same data source that was used in the creation of the training and test dataset. Even though the same data samples were not used in training and validation, the sentences were collected with the same algorithm to include sentences containing alcohol and cholesterol related content. Even though random sentences were also included in the datasets, some common expressions for high cholesterol or bad alcohol consumptions not captured by the dataset collection algorithm might have been missed. Furthermore, the labeling process was not carried out by a medical expert, which could result in misclassifications in the datasets. A manually labeled bigger validation dataset collected by medical experts would have been a better reference for the generalizability of the trained models. However, this was not attainable during the thesis process. Because of the above-mentioned limitations in the validation dataset, the results of this thesis might overestimate the accuracy of the presented models.

The third major limitation is the BERT model training process used in the present thesis. Due to time constraints and GPU-cluster availability, the BERT model training process was carried out without any hyperparameter tuning and changes in the model architecture. The early stopping used as regularization method did not prevent the BERT model from overfitting. The overfitting was evident in the training data classification accuracy of 1.00, whit a test data classification accuracy of 0.98 for alcohol and 0.99 for cholesterol. Even though the

validation data classification accuracies were high, they were far below the training and test data classification accuracies. The generalizability of the model might be compromised, and better regularization methods should be applied to develop the models further.

In addition to the limitations of the present thesis, a number of alternative approaches could be explored to increase the understanding and benefits of weak supervision processes, and to improve the results presented in the current thesis. Firstly, the BERT model used in the present thesis was pretrained with Finnish Wikipedia, news articles, and online discussion forum texts (Virtanen et al., 2019). A process of pre-training BERT with a medical corpus from scratch could allow more meaningful word embeddings to be used during the fine-tuning process, resulting in better classification accuracies in medical narrative context. Secondly, a larger manual training dataset could be used to assess the cost benefits of a weak supervision based approach compared to more traditional manual labeling process. Thirdly, different amounts of automatically labeled training data could be used to assess the requirements for the availability of training data, establishing guidelines on the size of the required training data for weak supervision approaches. And lastly, alternative low-resource machine learning methods such as active learning could be combined and compared with weak supervision approach.

As a conclusion, the results of the present thesis showed that weak supervision based approach was able to produce accurate models in classifying two medical risk factors, high cholesterol and alcohol consumption in Finnish language medical narratives. A machine learning model encompassing word embeddings was able to capture hidden patterns in the data and utilize the natural language understanding for better classification results and classifying cases which were not captured by the rules used to create the training data. Weak supervision approach was also able to produce more accurate classification models compared to the models trained with a small manually labeled dataset. Weak supervision approach might be a valuable tool to reduce the costs of applying machine learning algorithms in low-resource settings, where manual labeling process is time consuming, expensive, or requires the expertise of subject specialist.

Bibliography

Agrawal, A. (2002). Return on investment analysis for a computer-based patient record in the outpatient clinic setting. *Journal of the Association for Academic Minority Physicians : The Official Publication of the Association for Academic Minority Physicians*, *13*(3), 61–65. https://pubmed.ncbi.nlm.nih.gov/12362561/

Amarasingham, R., Plantinga, L., Diener-West, M., Gaskin, D. J., & Powe, N. R. (2009). Clinical information technologies and inpatient outcomes: a multiple hospital study. *Archives of Internal Medicine*, *169*(2), 108–114. https://doi.org/10.1001/ARCHINTERNMED.2008.520

Bach, S. H., He, B., Ratner, A., & Ré, C. (2017). Learning the Structure of Generative Models without Labeled Data. *34th International Conference on Machine Learning, ICML 2017*, *1*, 434–449. https://arxiv.org/abs/1703.00854v2

Bach, S. H., Rodriguez, D., Liu, Y., Luo, C., Shao, H., Xia, C., Sen, S., Ratner, A., Hancock, B., Alborzi, H., Kuchhal, R., Ré, C., & Malkin, R. (2019). Snorkel Drybell: A case study in deploying weak supervision at industrial scale. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 362–375. https://doi.org/10.1145/3299869.3314036

Bates, D. W., Kuperman, G. J., Rittenberg, E., Teich, J. M., Fiskio, J., Ma'luf, N., Onderdonk, A., Wybenga, D., Winkelman, J., Brennan, T. A., Komaroff, A. L., & Tanasijevic, M. (1999). A randomized trial of a computer-based intervention to reduce utilization of redundant laboratory tests. *The American Journal of Medicine*, *106*(2), 144–150. https://doi.org/10.1016/S0002-9343(98)00410-0

Bates, D. W., Leape, L. L., Cullen, D. J., Laird, N., Petersen, L. A., Teich, J. M., Burdick, E., Hickey, M., Kleefield, S., Shea, B., Vliet, M. vander, & Seger, D. L. (1998). Effect of computerized physician order entry and a team intervention on prevention of serious medication errors. *JAMA*, *280*(15), 1311–1316. https://doi.org/10.1001/JAMA.280.15.1311

Bates, D. W., Teich, J. M., Lee, J., Seger, D., Kuperman, G. J., Ma'Luf, N., Boyle, D., & Leape, L. (1999). The Impact of Computerized Physician Order Entry on Medication Error Prevention. *Journal of the American Medical Informatics Association : JAMIA*, *6*(4), 313. https://doi.org/10.1136/JAMIA.1999.00660313

Bottou, L., & Lin, C.-J. (2007). Support Vector Machine Solvers. *Large Scale Kernel Machines* , *3*(1), 301–320.

Callahan, A., Fries, J. A., Ré, C., Huddleston III, J. I., Giori, N. J., Delp, S., & Shah, N. H. (2019). Medical device surveillance with electronic health records. *Npj Digital Medicine*, *2*(94).

Chen, P., Tanasijevic, M. J., Schoenenberger, R. A., Fiskio, J., Kuperman, G. J., & Bates, D. W. (2003). A Computer-Based Intervention for Improving the Appropriateness of Antiepileptic Drug Level Monitoring. *American Journal of Clinical Pathology*, *119*(3), 432–438. https://doi.org/10.1309/A96XU9YKU298HB2R

Cheng, L. T. E., Zheng, J., Savova, G. K., & Erickson, B. J. (2010). Discerning tumor status from unstructured MRI reports--completeness of information in existing reports and utility of automated natural language processing. *Journal of Digital Imaging*, *23*(2), 119–132. https://doi.org/10.1007/S10278-009-9215-7

Chertow, G. M., Lee, J., Kuperman, G. J., Burdick, E., Horsky, J., Seger, D. L., Lee, R., Mekala, A., Song, J., Komaroff, A. L., & Bates, D. W. (2001). Guided medication dosing for inpatients with renal insufficiency. *JAMA*, *286*(22), 2839–2844. https://doi.org/10.1001/JAMA.286.22.2839

Cusick, M., Adekkanattu, P., Campion, T. R., Sholle, E. T., Myers, A., Banerjee, S., Alexopoulos, G., Wang, Y., & Pathak, J. (2021). Using weak supervision and deep learning to classify clinical notes for identification of current suicidal ideation. *Journal of Psychiatric Research*, *136*, 95–102. https://doi.org/10.1016/J.JPSYCHIRES.2021.01.052

DesRoches, C. M., Campbell, E. G., Vogeli, C., Zheng, J., Rao, S. R., Shields, A. E., Donelan, K., Rosenbaum, S., Bristol, S. J., & Jha, A. K. (2010). Electronic health records' limited successes suggest more targeted uses. *Health Affairs (Project Hope)*, *29*(4), 639–646. https://doi.org/10.1377/HLTHAFF.2009.1086

Devine, E. B., Hansen, R. N., Wilson-Norton, J. L., Lawless, N. M., Fisk, A. W., Blough, D. K., Martin, D. P., & Sullivan, S. D. (2010). The impact of computerized provider order entry on medication errors in a multispecialty group practice. *Journal of the American Medical Informatics Association : JAMIA*, *17*(1), 78–84. https://doi.org/10.1197/JAMIA.M3285

Devlin, J., Chang, M.-W., Lee, K., Google, K. T., & Language, A. I. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv Preprint ArXiv:1810.04805*. https://github.com/tensorflow/tensor2tensor

Dexter, P. R., Perkins, S., Overhage, J. M., Maharry, K., Kohler, R. B., & McDonald, C. J. (2001). A computerized reminder system to increase the use of preventive care for hospitalized patients. *The New England Journal of Medicine*, *345*(13), 965–970. https://doi.org/10.1056/NEJMSA010181

Elder, K. T., Wiltshire, J. C., Rooks, R. N., Belue, R., & Gary, L. C. (2010). Health Information Technology and Physician Career Satisfaction. *Perspectives in Health Information Management / AHIMA, American Health Information Management Association*, *7*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2921302/

Erstad, T. L. (2003). Analyzing computer based patient records: a review of literature. *Journal of Healthcare Information Management : JHIM*, *17*(4), 51–57. https://pubmed.ncbi.nlm.nih.gov/14558372/

Ewing, T., & Cusick, D. (2004). Knowing what to measure. *Healthcare Financial Management : Journal of the Healthcare Financial Management Association*, *58*(6), 60–63. https://pubmed.ncbi.nlm.nih.gov/17883234/

Goldberg, Y. (2017). Neural Network Methods for Natural Language Processing. *Synthesis Lectures on Human Language Technologies*, *10*(1), 1–311. https://doi.org/10.2200/S00762ED1V01Y201703HLT037

HaCohen-Kerner, Y., Miller, D., & Yigal, Y. (2020). The influence of preprocessing on text classification using a bag-of-words representation. *PLOS ONE*, *15*(5), e0232525. https://doi.org/10.1371/JOURNAL.PONE.0232525

Hancock, B., Bringmann, M., Varma, P., Liang, P., Wang, S., & Ré, C. (2018). Training Classifiers with Natural Language Explanations. *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, *1*, 1884–1895. https://doi.org/10.18653/v1/p18-1175

Juhn, Y., & Liu, H. (2020). Artificial intelligence approaches using natural language processing to advance EHR-based clinical research in Allergy, Asthma, and Immunology. *The Journal of Allergy and Clinical Immunology*, *145*(2), 463. https://doi.org/10.1016/J.JACI.2019.12.897

Karlsson, A., Ellonen, A., Irjala, H., Väliaho, V., Mattila, K., Nissi, L., Kytö, E., Kurki, S., Ristamäki, R., Vihinen, P., Laitinen, T., Ålgars, A., Jyrkkiö, S., Minn, H., & Heervä, E. (2021). Impact of deep learning-determined smoking status on mortality of cancer patients: never too late to quit. *ESMO Open*, *6*(3), 100175. https://doi.org/10.1016/J.ESMOOP.2021.100175

Khalifa, A., & Meystre, S. (2015). Adapting existing natural language processing resources for cardiovascular risk factors identification in clinical notes. *Journal of Biomedical Informatics*, *58*, S128–S132. https://doi.org/10.1016/J.JBI.2015.08.002

Koleck, T. A., Dreisbach, C., Bourne, P. E., & Bakken, S. (2019). Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. *Journal of the American Medical Informatics Association*, *26*(4), 364–379. https://doi.org/10.1093/JAMIA/OCY173

Kukafka, R., Ancker, J. S., Chan, C., Chelico, J., Khan, S., Mortoti, S., Natarajan, K., Presley, K., & Stephens, K. (2007). Redesigning electronic health record systems to support public health. *Journal of Biomedical Informatics*, *40*(4), 398–409. https://doi.org/10.1016/J.JBI.2007.07.001

Ledwich, L. J., Harrington, T. M., Ayoub, W. T., Sartorius, J. A., & Newman, E. D. (2009). Improved influenza and pneumococcal vaccination in rheumatology patients taking immunosuppressants using an electronic health record best practice alert. *Arthritis and Rheumatism*, *61*(11), 1505–1510. https://doi.org/10.1002/ART.24873

Liang, H., Sun, X., Sun, Y., & Gao, Y. (2017). Text feature extraction based on deep learning: a review. *Eurasip Journal on Wireless Communications and Networking*, *2017*(1), 1–12. https://doi.org/10.1186/S13638-017-0993-1/FIGURES/3

Linzer, M., Konrad, T. R., Douglas, J., McMurray, J. E., Pathman, D. E., Williams, E. S., Schwartz, M. D., Gerrity, M., Scheckler, W., Bigby, J. A., & Rhodes, E. (2000). Managed Care, Time Pressure, and Physician Job Satisfaction: Results from the Physician Worklife Study. *Journal of General Internal Medicine*, *15*(7), 441. https://doi.org/10.1046/J.1525-1497.2000.05239.X

Majumder, P., Mitra, M., & Chaudhuri, B. B. (2002). N-gram: a language independent approach to IR and NLP. *International Conference on Universal Knowledge and Language*.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. In *Introduction to Information Retrieval*. Cambridge University Press. https://doi.org/10.1017/CBO9780511809071

Martin-Sanchez, F., & Verspoor, K. (2014). Big Data in Medicine is Driving Big Changes. *Yearbook of Medical Informatics*, *9*(1), 14. https://doi.org/10.15265/IY-2014-0020

McCullough, J. S., Casey, M., Moscovice, I., & Prasad, S. (2017). The Effect Of Health Information Technology On Quality In U.S. Hospitals. *Health Affairs*, *29*(4), 647–654. https://doi.org/10.1377/HLTHAFF.2010.0155

Menachemi, N., Chukmaitov, A., Saunders, C., & Brooks, R. G. (2008). Hospital quality of care: Does information technology matter? The relationship between information technology adoption and quality of care. *Health Care Management Review*, *33*(1), 51–59. https://doi.org/10.1097/01.HMR.0000304497.89684.36

Menachemi, N., & Collum, T. H. (2011). Benefits and drawbacks of electronic health record systems. *Risk Management and Healthcare Policy*, *4*, 47. https://doi.org/10.2147/RMHP.S12985

Menachemi, N., Powers, T. L., & Brooks, R. G. (2009). The role of information technology usage in physician practice satisfaction. *Health Care Management Review*, *34*(4), 364–371. https://doi.org/10.1097/HMR.0B013E3181A90D53

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*. https://arxiv.org/abs/1301.3781v3

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems*. https://arxiv.org/abs/1310.4546v1

Miskulin, D. C., Weiner, D. E., Tighiouart, H., Ladik, V., Servilla, K., Zager, P. G., Martin, A., Johnson, H. K., & Meyer, K. B. (2009). Computerized decision support for EPO dosing in hemodialysis patients. *American Journal of Kidney Diseases : The Official Journal of the National Kidney Foundation*, *54*(6), 1081–1088. https://doi.org/10.1053/J.AJKD.2009.07.010

Moon, S., Liu, S., Scott, C. G., Samudrala, S., Abidian, M. M., Geske, J. B., Noseworthy, P. A., Shellum, J. L., Chaudhry, R., Ommen, S. R., Nishimura, R. A., Liu, H., & Arruda-Olson, A. M. (2019). Automated extraction of sudden cardiac death risk factors in hypertrophic cardiomyopathy patients by natural language processing. *International Journal of Medical Informatics*, *128*, 32–38. https://doi.org/10.1016/J.IJMEDINF.2019.05.008

Murphy, K. P. (2006). Naive Bayes classifiers. *University of British Columbia*, *18*(60), 1–8.

Niès, J., Colombet, I., Zapletal, E., Gillaizeau, F., & Durieux, P. (2010). Effects of automated alerts on unnecessarily repeated serology tests in a cardiovascular surgery department: a time series analysis. *BMC Health Services Research*, *10*. https://doi.org/10.1186/1472-6963-10-70

Aspden, P., Corrigan, J., Wolcott, J., & Erickson, S. (2004). Patient Safety: Achieving a New Standard for Care. *National Academies Press*. https://doi.org/10.17226/10863

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. https://doi.org/10.1186/S12911-019-0863-3/TABLES/3

Pathman, D. E., Williams, E. S., & Konrad, T. R. (1996). Rural Physician Satisfaction: Its Sources and Relationship to Retention. *The Journal of Rural Health*, *12*(5), 366–377. https://doi.org/10.1111/J.1748-0361.1996.TB00804.X

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*, 2227–2237. https://doi.org/10.18653/V1/N18-1202

Pradhan, A. (2012). Support vector machine-a survey. *International Journal of Emerging Technology and Advanced Engineering*, *2*(8). www.ijetae.com

Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., & Ré, C. (2017). Snorkel: Rapid Training Data Creation with Weak Supervision. *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, *11*(3), 269. https://doi.org/10.14778/3157794.3157797

Ratner, A., de Sa, C., Wu, S., Selsam, D., & Ré, C. (2016). Data Programming: Creating Large Training Sets, Quickly. *ArXiv:1605.07723*. http://www.spacemachine.net/views/2016/3/datasets-over-algorithms

Ratner, A., Hancock, B., Dunnmon, J., Sala, F., Pandey, S., & Ré, C. (2018). Training Complex Models with Multi-Task Weak Supervision. *ArXiv:1810.02840*.

Ratner, A., Bach, S. H., Ehrenberg, H. R., & Ré, C. (2017). Snorkel: Fast Training Set Generation for Information Extraction. *Proceedings of the 2017 ACM International Conference on Management of Data*. https://doi.org/10.1145/3035918.3056442

Schmitt, K. F., & Wofford, D. A. (2002). Financial analysis projects clear returns from electronic medical records. *Healthcare Financial Management : Journal of the Healthcare*

Financial Management Association, *56*(1), 52–57. https://pubmed.ncbi.nlm.nih.gov/11806319/

Settles, B. (2009). *Active Learning Literature Survey*. https://minds.wisconsin.edu/handle/1793/60660

Séverac, F., Sauleau, E. A., Meyer, N., Lefèvre, H., Nisand, G., & Jay, N. (2015). Non-redundant association rules between diseases and medications: An automated method for knowledge base construction. *BMC Medical Informatics and Decision Making*, *15*(1), 1–7. https://doi.org/10.1186/S12911-015-0151-9/TABLES/3

Shorten, C., Khoshgoftaar, T. M., & Furht, B. (2021). Text Data Augmentation for Deep Learning. *Journal of Big Data 2021 8:1*, *8*(1), 1–34. https://doi.org/10.1186/S40537-021-00492-0

Spasic, I., & Nenadic, G. (2020). Clinical Text Data in Machine Learning: Systematic Review. *JMIR Medical Informatics*, *8*(3). https://doi.org/10.2196/17984

Tierney, W. M., Miller, M. E., & McDonald, C. J. (1990). The effect on test ordering of informing physicians of the charges for outpatient diagnostic tests. *The New England Journal of Medicine*, *322*(21), 1499–1504. https://doi.org/10.1056/NEJM199005243222105

Tierney, W. M., Miller, M. E., Overhage, J. M., & McDonald, C. J. (1993). Physician inpatient order writing on microcomputer workstations. Effects on resource utilization. *JAMA*, *269*(3), 379–383. https://doi.org/10.1001/jama.1993.03500030077036

Varma, P., Sala, F., He, A., Ratner, A., & Ré, C. (2019). Learning Dependency Structures for Weak Supervision Models. *ArXiv:1903.05844*.

Virapongse, A., Bates, D. W., Shi, P., Jenter, C. A., Volk, L. A., Kleinman, K., Sato, L., & Simon, S. R. (2008). Electronic health records and malpractice claims in office practice. *Archives of Internal Medicine*, *168*(21), 2362–2367. https://doi.org/10.1001/ARCHINTE.168.21.2362

Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., & Pyysalo, S. (2019). Multilingual is not enough: BERT for Finnish. *ArXiv:1912.07076*.

Wang, B., Wang, A., Chen, F., Wang, Y., & Kuo, C. C. J. (2019). Evaluating word embedding models: methods and experimental results. *APSIPA Transactions on Signal and Information Processing*, *8*, 1–14. https://doi.org/10.1017/ATSIP.2019.12

Wang, S. J., Middleton, B., Prosser, L. A., Bardon, C. G., Spurr, C. D., Carchidi, P. J., Kittler, A. F., Goldszer, R. C., Fairchild, D. G., Sussman, A. J., Kuperman, G. J., & Bates, D. W. (2003). A cost-benefit analysis of electronic medical records in primary care. *The American Journal of Medicine*, *114*(5), 397–403. https://doi.org/10.1016/S0002-9343(03)00057-3

Wang, T., & Gibbs, D. (2019). A Framework for Performance Comparison among Major Electronic Health Record Systems. *Perspectives in Health Information Management*, *16*(Fall). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6931047/

Wang, Y., Sohn, S., Liu, S., Shen, F., Wang, L., Atkinson, E. J., Amin, S., & Liu, H. (2019). A clinical text classification paradigm using weak supervision and deep representation. *BMC Medical Informatics and Decision Making*, *19*(1), 1–13. https://doi.org/10.1186/S12911-018-0723-6/FIGURES/4

Westin, A. F. (2005). Public attitudes toward electronic health records. *AHIP Coverage*, *46*(4), 22–25. https://pubmed.ncbi.nlm.nih.gov/16149656/

Wilson, G. A., McDonald, C. J., & McCabe, G. P. (1982). The effect of immediate access to a computerized medical record on physician test ordering: a controlled clinical trial in the emergency room. *American Journal of Public Health*, *72*(7), 698. https://doi.org/10.2105/AJPH.72.7.698

Wu, S., Hsiao, L., Cheng, X., Hancock, B., Rekatsinas, T., Levis, P., & Ré, C. (2017). Fonduer: Knowledge Base Construction from Richly Formatted Data. *ArXiv:1703.05028* .

Zhu, H., Ni, Y., Peng, C., Qiu, Z., & Cao, F. (2012). Automatic Extracting of Patient-related Attributes: Disease, Age, Gender and Race. *Studies in Health Technology and Informatics*, *180*, 589–593. https://doi.org/10.3233/978-1-61499-101-4-589

Zurita, L., & Nøhr, C. (2004). Patient opinion--EHR assessment from the users perspective. *Studies in Health Technology and Informatics*, *107*(Pt 2), 1333–1336. https://pubmed.ncbi.nlm.nih.gov/15361031/