UPPSALA
UNIVERSITET

UPTEC IT 20046
Examensarbete 30 hp
November 2020

Mapping medical expressions to MedDRA using Natural Language Processing

Vanja Wallner

Institutionen för informationsteknologi
Department of Information Technology

UPPSALA
UNIVERSITET

Teknisk- naturvetenskaplig fakultet
UTH-enheten

Besöksadress:
Ångströmlaboratoriet
Lägerhyddsvägen 1
Hus 4, Plan 0

Postadress:
Box 536
751 21 Uppsala

Telefon:
018 – 471 30 03

Telefax:
018 – 471 30 00

Hemsida:
http://www.teknat.uu.se/student

Abstract

Mapping medical expressions to MedDRA using Natural Language Processing

Vanja Wallner

Pharmacovigilance, also referred to as drug safety, is an important science for identifying risks related to medicine intake. Side effects of medicine can be caused by for example interactions, high dosage and misuse. In order to find patterns in what causes the unwanted effects, information needs to be gathered and mapped to predefined terms. This mapping is today done manually by experts which can be a very difficult and time consuming task. In this thesis the aim is to automate the process of mapping side effects by using machine learning techniques.

The model was developed using information from preexisting mappings of verbatim expressions of side effects. The final model that was constructed made use of the pre-trained language model BERT, which has received state-of-the-art results within the NLP field. When evaluating on the test set the final model performed an accuracy of 80.21%. It was found that some verbatims were very difficult for our model to classify mainly because of ambiguity or lack of information contained in the verbatim. As it is very important for the mappings to be done correctly, a threshold was introduced which left for manual mapping the verbatims that were most difficult to classify. This process could however still be improved as suggested terms were generated from the model, which could be used as support for the specialist responsible for the manual mapping.

Handledare: Lucie Gattepaille
Ämnesgranskare: Robin Strand
Examinator: Lars-Åke Nordén
UPTEC IT 20046
Tryckt av: Reprocentralen ITC

Sammanfattning

Läkemedelsövervakning är viktigt för att identifiera risker relaterade till medicinintag. Biverkningar av medicin kan till exempel orsakas av interaktioner, hög dosering och missbruk. För att hitta mönster i vad som orsakar de oönskade effekterna måste information samlas in och mappas till fördefinierade termer. Denna kartläggning görs idag manuellt av experter, vilket kan vara en mycket svår och tidskrävande uppgift. I denna avhandling är syftet att undersöka om vi kan automatisera processen av att kartlägga biverkningar med hjälp av maskininlärning.

Modellen utvecklades med hjälp av information från redan existerande mappningar av rapporterade biverkningar. Den slutliga modellen som konstruerades använde sig av BERT, som har visat mycket goda resultat för olika uppgifter inom det språkteknologiska området. Vid utvärderingen av test datat utförde den slutliga modellen en *accuracy* på 80,21 %. Det visade sig att vissa uttryck för biverkningar var mycket svåra för vår modell att klassificera på grund av saker som tvetydighet eller brist på information. Eftersom det är mycket viktigt att dessa kartläggningar görs korrekt infördes en *threshold* som uteslöt de uttryck som var svårast att klassificera. Dessa uttryck lämnades istället för manuell kartläggning. Processen av manuell kartläggning kunde däremot underlättas då föreslagna termer genererades från modellen och skulle därmed kunna användas som stöd för den ansvarige specialisten.

Acknowledgements

First of all, I would like to thank UMC for the opportunity of working with this thesis project. It has been a fun, challenging and educational experience! I would like to direct my gratitude to the staff working at UMC, especially my team, for always being helpful and making me feel welcome at the office. A special thanks to my supervisor Lucie Gattepaille for her invaluable guidance. She has given me great advice and feedback on my work throughout this project. I would also like to thank the terminology specialists at UMC that took their time to evaluate some of my data, which made it possible for me to carry out the error analysis. Last but not least thank you to my reviewer Robin Strand for reading and providing feedback on this report.

Contents

Glossary

ADR	Adverse Drug Reaction
BERT	Bidirectional Encoder Representations from Transformers
GS	Gold standard
HLGT	High level group term (MedDRA hierarchy)
HLT	High level term (MedDRA hierarchy)
ICSR	Individual case safety report
LLT	Lowest level term (MedDRA hierarchy)
MedDRA	the Medical Dictionary for Regulatory Activities
PIDM	Programme for International Drug Monitoring
PT	Preferred term (MedDRA hierarchy)
SOC	System organ class (MedDRA hierarchy)
UMC	Uppsala Monitoring Centre

1 Introduction

Pharmacovigilance is defined by the World Health Organisation (WHO) as ”the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other possible drug-related problems.” [1] Pharmacovigilance is important for identifying risks related to medicine intake caused by interactions, high dosage, misuse etc. where the detection of these risks can be crucial to ensure patient safety.

The Uppsala Monitoring Centre (UMC) is working alongside WHO as part of the WHO Programme for International Drug Monitoring (PIDM) for the common goal of “a systematic collection of information on serious adverse drug reactions during the development and particularly after medicines have been made available for public use”.[2] One tool developed and maintained by UMC, for the purpose of drug safety, is VigiBase.[3] It is WHOs global database, containing millions of individual case safety reports (ICSRs) from countries all over the world that are part of PIDM. The ICSRs are reports containing suspected adverse drug reactions (ADRs) which are reported and collected from both patients and healthcare professionals, by the National Authorities of each member country of the PIDM. An ADR is a term used to describe the unintended side effects of drugs that pharmacovigilance experts are trying to detect.

When expressing an ADR in free text, the very same reaction can be explained in many different ways: ”I have a headache”, ”my head hurts” and ”I have a pain in my head” are all verbatim expressions of the same condition: Headache. It is of great importance to classify these expressions as equal in order to find possible correlations between drugs and side effects. In order to do this, normalization can be performed, mapping the different verbatim expressions to the same condition label. As of today this mapping is done manually by coding specialists who choose a fitting label based on the verbatim. The labels used for mapping are the terms found in The Medical Dictionary for Regulatory Activities (MedDRA) which is a terminology that contains several 10,000 of medical terms. This manual mapping can be a very time consuming task and requires the work of specialists. When leaving this task for human evaluation there is also the aspect of subjectivity which can result in similar verbatim being mapped to different labels by different coding specialists.

1.1 Motivation

This project aims to develop an algorithm that can automate the process of mapping verbatim expressions of ADRs to MedDRA terms. The results of this project can be beneficial for the supervisor in multiple ways. For example in improving their mapping of verbatims in the side effect reports. At the time they only rely on direct matches to MedDRA and therefore might be missing valuable information. Both in pharmacovigilance and clinical trials experts are manually performing data entry. The proposed algorithm could therefore be a resource to improve these processes.

1.2 Problem formulation

During this project we will focus on answering the following question:

- How can we use verbatim descriptions of adverse drug reactions to create an automatic mapping to MedDRA terms?

As a guidance, to help answer the above mentioned main question, a few more specific questions were formulated:

- *How do we handle lack of training data?*
  Even though we have access to a few million rows of training data we also have thousands of classes to map to. This means that there might not be enough data to successfully train a classifier with good results. There might also be class imbalances meaning that some classes are less represented than others in the training data.
- *How do we deal with our training data being inadequate?*
  In some cases different verbatims with the exact same words can be mapped to different labels. With a high number of classes there might also be cases where there are multiple labels that fit the same verbatim, so there might be multiple correct answers even if the training data will only contain one.
- *Can machine learning techniques be used to improve the results?*
  The verbatims used as input are generally short collections of words. They can be actual sentences but also just descriptive words and/or measured values that can be tricky to classify with classic NLP methods. So the question is if improvements can be made by making use of algorithms from the ever advancing machine learning field?

1.3 Delimitations

Some limitations were made for this project. The project aims to create an algorithm that can map the verbatim expressions from the side effect reports to the predefined MedDRA terms. The algorithm will thereby assumably only be applied to text that we know contain descriptions of medical conditions. The verbatims are in most cases derived from a more descriptive source, like medical records or more thorough reports. Limiting ourselves to solely using the verbatim for creating this algorithm means we might be losing information that was available to the coding specialist performing previous mappings, but this choice was made for the project to stay within a reasonable scope.

1.4 Thesis overview

The report is split up into multiple sections. After this introduction where motivation, problem formulation and delimitation have been presented, follows section 2 that presents the background for the project. In section 3 the theories behind the project will be presented. Section 4 covers the methods used and section 5 presents the results achieved. Section 6 covers a discussion followed by final conclusions in section 7. Lastly section 8 presents ideas about future work.

2 Background

2.1 UMC

This project is done in cooperation with the Uppsala Monitoring Centre (UMC) which is a non-profit organisation that works with the goal of improving medicine safety all over the world. Among other things they maintain Vigibase, the World Health Organization's (WHO) database of individual case safety reports. Vigibase contains over 20 millions of reports from over 130 countries worldwide with cases of suspected adverse events from medicines. On UMC's website they describe the purpose of Vigibase as to ”ensure that early signs of previously unknown medicines-related safety problems are identified as rapidly as possible.” [3]

2.2 MedDRA

The Medical Dictionary for Regulatory Activities (MedDRA) is a terminology that contains several 10,000 of medical terms presented in a hierarchical order containing five layers displayed in figure 1.[4] MedDRA is continuously updated with new medical concepts being added or existing concepts being modified. In this thesis MedDRA version 22.1, released in September 2019, is used. [5] In this version the highest level layer of MedDRA 'System Organ Class' (SOC) contained 27 terms while the lowest level layer 'Lowest Level Term' (LLT) contained over 80,000 terms. The highest level layer (SOC) contains the most general terms and for each layer the terms get more specific. The 'Preferred term' (PT) is the term used to label the side effects in VigiBase and it consists of almost 24,000 terms. The most specific term (LLT) can contain for example synonyms or different spellings of the PT, as well as the PT itself.

Every PT is primarily assigned to one SOC but can also be secondarily assigned to several other SOCs. An example is the PT “Asthma” that is found under its primary SOC “Respiratory, thoracic and mediastinal disorders” (SOC) but also “Immune system disorders” (SOC) as secondary. Each LLT is however uniquely related to one PT.

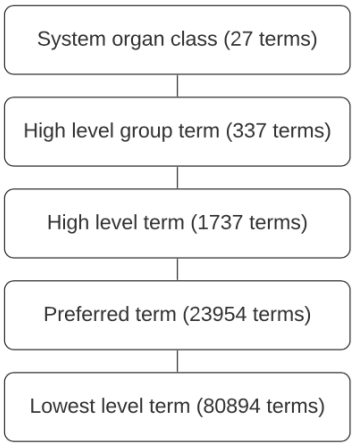

Figure 1: The five layers of the MedDRA hierarchy. The number of terms present in each layer (as of 2020-01-15) are shown in parenthesis.

Figure 2 shows an example of how 'Lactose Intolerance' is categorized in MedDRA. First we have 'Metabolism and nutrition disorders' (SOC) which is a very general term, afterwards comes the more specific 'Food intolerance syndromes' which is a 'High level group term' (HLGT). From the HLGT of food intolerance we specify even more unto 'Sugar intolerance (excl glucose intolerance)' which is a 'High level term' (HLT). Under this category we find the PT 'Lactose intolerance' that we were looking for. As can be seen the chosen PT has four corresponding LLTs. Among the LLTs we find the PT 'Lactose intolerance', 'Lactose intolerant' as well as synonyms including different spellings 'Lactose diarrhea/diarrhoea'.

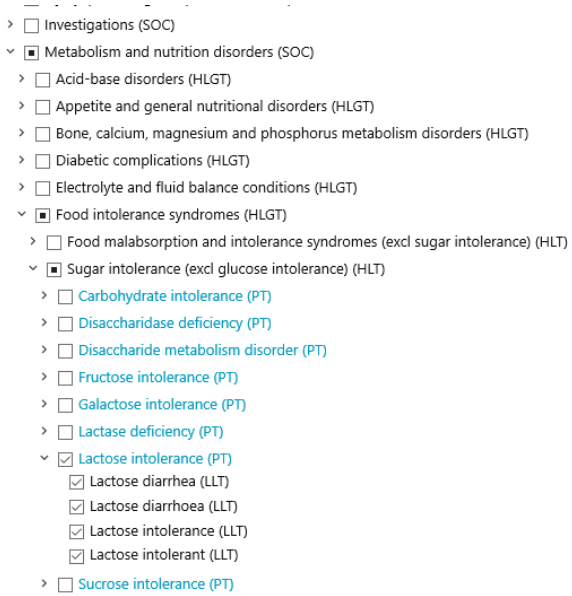

Figure 2: An example showing the MedDRA hierarchy for the PT 'Lactose intolerance'.

3 Theory

In this chapter the theory and techniques used in this project will be introduced and explained.

3.1 Natural language processing

Natural language processing (NLP) concerns the interaction between the human language and computers. In theory, translating human language to computers could be an easy task: words are just collections of characters and sentences are collections of words. However it can be more difficult in practice since the human language can be ambiguous as well as ever changing and evolving. Yoav Goldberg states in his book *Neural Network Methods for Natural Language Processing* that "People are great at producing language and understanding language, and are capable of expressing, perceiving, and interpreting very elaborate and nuanced meanings. At the same time, while we humans are great *users* of language, we are also very poor at formally understanding and describing the rules that *govern* language." [6]

3.2 Classification

Classification is the technique of categorizing data to a set of discrete output values, referred to as classes or labels. The classification algorithm is created from patterns found in preexisting mappings. By finding the features that create these patterns and determine how the data is mapped, any new observations can be mapped according to this information. The goal is to find enough features to be able to correctly map any new (unknown) data to the correct class. In an example of classifying vehicles to the labels ”bus”, ”car” or ”motorcycle” the features could for example be the number of wheels, the length of the vehicle and the presence of a steering wheel. An unknown vehicle with more than 4 wheels should probably be classified as ”bus” while an unknown vehicle without a steering wheel should be classified as ”motorcycle”. These patterns could be found by looking at multiple buses, cars and motorcycles and finding how they distinguish from one another.

Classification is seen as a supervised learning technique (see section 3.4) in machine learning since it uses previously made classifications to make future predictions. The preexisting mappings are referred to as the gold standard (GS) and they are seen as the benchmark. Within the NLP-field the data used is in text format, usually as words or sentences. When classifying data in this form the task is more specifically referred to as text classification. Some well-known examples are: sentiment analysis (text classified as having positive or negative sentiment) and language detection (predicting what language the text is written in).

Multi-class classification
The simplest form of classification is called binary classification and is done with only two classes. Examples of binary classification are classifying data to True/False based on some criterion, for example classifying e-mails to being spam/not-spam or reviews of a product to being positive/negative.

When there are more than two classes it is considered a multi-class classification problem. As the number of classes increases the classification problem gets increasingly difficult to solve. To explain this increasing difficulty we can compare a binary classification problem (2 classes) with a multi-class problem of 100 classes. To exemplify the problem a ”dummy classifier” could be used, that simply classifies everything to the same class. Statistically (not considering imbalanced classes) this would mean that in the binary classification we get 50% accuracy while in the case of 100 classes we get 1% accuracy. More generally we would get $accuracy = \frac{1}{nrOfClasses}$, clearly showing the relation between a decreased accuracy with an increased number of classes. In practice the algorithms are usually better than this ”dummy classifier”, but as the number of classes increases any algorithm will have more outputs to consider, decreasing the

possibility for a correct classification.

Imbalanced classes
Having imbalanced classes means that the number of observations from different classes, used to train the classifier, is disproportionate. This can lead to bias within the model as it is trained to classify more often towards the most represented classes which can give results that seem more promising than they really are. Let us say there is a binary classification problem where the goal is to detect spam e-mails and the observations are 95% non-spam and 5% spam. The ”dummy classifier” that always classifies to non-spam would then give an accuracy of around 95%, which seems great, even if nothing has really been implemented.

Multi-label classification
Commonly within classification each observation is mapped to a single class. With multi-label classification however the observations are mapped to a set of classes, one or multiple ones.

Hierarchical classification
Usually all classes are equally differentiated from one another. If however the classes are part of a hierarchy they will be more or less related. This relation can be used with a hierarchical classifier that can start mapping data to a low-level and increase the level of detail.

3.3 Text representation

For any classification problem the input needs to be numerical since that is the only representation that a computer can comprehend. When working with NLP-problems we are using text as input and before doing any calculations we need some method to translate the text into numbers. There are multiple proposed solutions for text representation some of which are presented in this section.

Bag of words
One simple approach of representing text is Bag Of Words (BOW) which takes the words and its number of occurrences in a document into account. If two documents consists of the same words, they are seemingly similar and could therefore belong to the same class. By creating vectors that reflect on the term frequency, similarities between documents could be found by vector comparisons.

The data representation will be a vector were each position corresponds to a word that is present in some of the documents. Each document will then have their own vector were each number represents the occurrence of the word in the document. This way of rep-

resenting each word with a vector of N positions with a ”1” in the position representing that word and ”0” for the other N-1 positions is called a one-hot encoding.

Document 1 = ”Headache”
Document 2 = ”Drug exposure during pregnancy”
Document 3 = ”Drug exposure”

Using the three documents above as an example the corpus used would consist of the words: ”drug” ”during” ”exposure” ”headache” ”pregnancy”. Since the corpus consists of five words, the vector representation will be five dimensions. Comparing the vector representation below it is clear to see that document 2 has more in common with document 3 than document 1, as expected.

Document 1 = ”Headache” = [0 0 0 1 0]
Document 2 = ”Drug exposure during pregnancy” = [1 1 1 0 1]
Document 3 = ”Drug exposure” = [1 0 1 0 0]

Tf-idf
The Bag of words representation is based on term frequency but it doesn't take into account the fact that words are more or less commonly used. Some words like ”the”, ”of”, ”a”, ”that” appear more often in the English language and these words might have a high frequency in multiple documents, even if these documents should not be seen as similar. ”Term frequency - inverse document frequency”, often shortened tf-idf, deals with this weakness by weighting the frequency of each term with the number of documents where they are present.

Word embeddings
There are multiple problems with the earlier mentioned representations (BOW and tf-idf). One being the high dimensionality of the vectors, which will be growing with the number of terms in the corpus. Another problem is that similar words are not connected in any way; these representations lack awareness of word meaning. With BOW and tf-idf there will probably be similarities found between the vector representation of ”I feel pain in my head” and ”I feel pain in my arm” but not between ”she felt pain in her head” and ”he had a headache” since the last sentences have no common words.

In his book *Speech and Language processing* Jurafsky mentions terms such as *word similarity* and *word relatedness* [7]. These concepts can be used to understand the insufficiency of using term frequency for representation, which is simply based on the words and not the meaning behind them. Even if two words are not synonyms they can still be more or less similar or related to one another. Cat is not a synonym of dog, but cats and dogs are still similar words used in similar contexts. In the same manner coffee

and cup are neither synonyms nor similar words, but they are still related and associated to one another.

Word embeddings are a collection of techniques used for creating word vectors and they often include the use of neural networks. This results in vector representations that are much more dense then the one-hot encodings mentioned in previous sections. Another advantage of using word embeddings is that the vectors capture semantic meaning of words from the contexts of where it appears. When training a word embedding model with sentences, not only the target word is considered but also its surrounding words. In 2013 Google introduced a word embedding model which later became known as word2vec [8] that became very popular for creating word vectors. It makes use of two architectures called CBOW and Skip-gram and produces word vectors from unsupervised training on a large text corpus.

3.4 Machine learning

Machine learning is a field within computer science. The objective is for the computer to ”learn” how to solve a problem (that it is not explicitly programmed to solve) based on data. There are different kinds of machine learning algorithms which can be divided into separate categories. The three most common ones are supervised learning, unsupervised learning and reinforcement learning.

Supervised learning can be used if we have access to a labeled data set of observations. The model can learn from this set and find patterns that will help make future predictions. When there is no labeled dataset to begin with, *unsupervised learning* algorithms can be used. These models tries to group data together based on underlying patterns. Lastly *reinforcement learning* is based on interaction with the environment. The system learns by rewards, were better choices are rewarded higher and thereby effecting future choices.

3.4.1 Deep learning

Deep learning is a sub-field within machine learning that is based on artificial neural networks (ANNs). ANNs are a set of algorithms with a structure inspired by the signal transmission of neurons in the brain. An ANN is built in multiple layers: starting with the input layer, ending with the output layer and then a number of hidden layers in between. ANNs operates on numerical data and the input must be of fixed size. When working with data that is not numerical by default, for example text, it needs to be translated into a numerical data representation. When using deep learning algo-

rithms the features are extracted from the data without human intervention, as opposed to traditional machine learning algorithms. This however comes at a cost of needing a relatively high amount of training data for the algorithm to be successful.

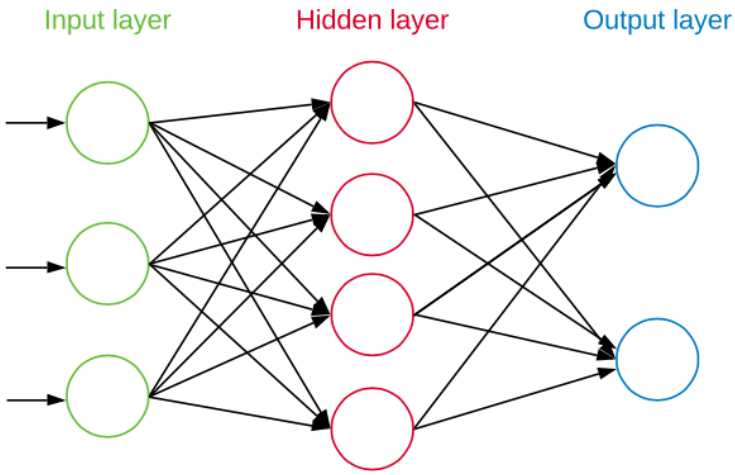

Figure 3: The figure shows the structure of a simple ANN with a single hidden layer.

Recurrent neural networks (RNNs) is a field within neural networks with algorithms that, opposed to the simpler neural networks, take sequential structures into account (through time or space depending on the application). This is accomplished by internal feedback loops in the network that creates what can be referred to as ”memory”.

In a paper from 2017 [10] researchers from Google presented the concept of Transformers, built in an encoder-decoder architecture. The conventional encoder-decoder model has a sequence of connected RNNs where each RNN inputs a token and outputs a vector that is based on the token as well as all the previous tokens. One disadvantage of this model is that the input has to be fed to the system sequentially, as each step is dependent on previous calculations. The Transformer introduced an alternative to the RNN architecture, which uses something called attention instead of recurrence. As opposed to the RNN architecture with the sequential dependency of the input, the Transformer reads the entire word sequence at once and can learn its context both from the previous as well as the following words. The model is thereby considered to be bidirectional.

3.4.2 BERT

BERT, which stands for Bidirectional Encoder Representations from Transformers, is a pre-trained language model that was released by Google in 2018. As the name

might reveal, the BERT architecture consists of multiple layers of transformer encoders. The model has received state-of-the-art results on multiple NLP tasks [9]. BERT is trained on a huge data set consisting of the English Wikipedia (2,500M words) and the BooksCorpus (800M words). The training is unsupervised and two different tasks are performed, namely *Mask language model* (MLM) and *Next sentence prediction* (NSP), which creates the word representation. For MLM the model looks at an entire sentence or piece of text trying to predict the word that has been masked out. This task is seen as a key technical innovation as it uses the bidirectional training of Transformers for a language modelling task. For the NSP task, the model received pairs of sentences and had to predict if the second sentence followed the first in the original document.

The pre-trained BERT model has a general knowledge of the English language but the model needs to be fine-tuned to perform a specific task. For a classification task this means training the model using training data to detect how the input relates to the classes.

As input BERT takes a sequence of tokens. BERT has a corpus of tokens that can be numbers, words or segments of words, which is used to represent the input. One benefit of the token representation is that any word can be represented. If a word is not found within the corpus it can be broken down into multiple tokens and possibly keep some of the original meaning. As an example BERT has no single token to represent the word ”chills” but it can instead be represented with the tokens ”chill” and ”##s”, where the ”##” represents that the token is part of the same word as the previous token. If the word ”headache” was not found in the token corpus it would be represented by the tokens ”head” and ”##ache” which in this case still contains most of the meaning. The worst case scenario would be for a word to be broken down into each character that it consists of: not keeping a lot of meaning but still being able to represent the word. The number of tokens is fixed length for a given model and corresponds to the size of the input layer. If the number of tokens for a given model is 64 it means that any input that can be represented by less tokens will be padded using a [PAD] token and if an input needs more than 64 tokens it will be cut off.

When using the fine tuned BERT model for classification it produces an output layer of logits, with a layer size corresponding to to the number of classes. Logits are integer values ranging from -∞ to ∞ and represents the unnormalised predictions of the model. The logits can be turned into a distribution of probabilites using the softmax function. Each class will then be represented by a value showing the models probability for each possible class being the correct class of a given input. When using the softmax function on the logits layer to create a vector of probabilities, taking the sum of that vector will

always add up to 1. The softmax function is defined as:

$$S(y) = \frac{e^{y_i}}{\sum_{j=1}^{n} e^{y_j}}$$

Where *S* is the resulting softmax vector, *e* is the standard exponential function and *y* is the vector of logits ranging from position *i = 1,2,...n*.

3.5 Evaluation

3.5.1 Confusion matrix

In binary classification there are two classes, in this example referred to as the positive labels (P) and the negative labels (N). When we want to evaluate how well our model can classify class P we'll use a confusion matrix where we compare the predictions made to the actual classes. Each observation will be part of one of the sets: TP, TN, FP or FN.

True positives (TP) - All the observations of class P correctly predicted as P.
True negatives (TN) - All the observations of class N correctly predicted as N.
False positives (FP) - All the observations of class N incorrectly predicted as P.
False negatives (FN) - All the observations of class P incorrectly predicted as N.

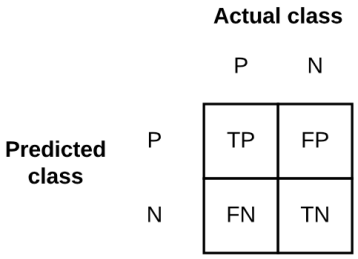

Figure 4: Confusion matrix for binary classification

3.5.2 Accuracy

Accuracy is the ratio of correct predictions over the total number of observations and can be a good measurement to understand the overall performance of a system. However there are other metrics like *Precision* and *Recall* that look at FN and FP separately which can expose imbalances in these rates that are not shown in the accuracy.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

3.5.3 Precision

Precision is the ratio of correct positive predictions over all positive predictions. It shows how many of all the observations that were classified as P that are actually part of P.

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

3.5.4 Recall

Recall is the ratio of all correct positive predictions over all positive observations. It shows how many of all the observations that are part of P that we correctly classified to P.

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

3.5.5 F-score

The F-score is a measure that considers both the precision and the recall. It is a score between 0 and 1, with 1 being the perfect score. The general formula looks like this:

$$F_\beta = (1 + \beta^2) * \frac{precision * recall}{\beta^2 * precision + recall}$$

The most common f-score is called f_1-score meaning that $\beta = 1$ in the formula. This makes recall and precision have an equal impact on the f-score. There are different possible f-scores that can be used, like f_2-score, $f_{0.5}$-score and $f_{0.2}$-score, with different values for β. Using a value $\beta > 1$ will weigh recall higher than precision while a

value $\beta < 1$ will weigh precision higher than recall. The choice of β therefore depends on the importance of recall and precision for the classification task. If for example the task is to detect tumors in x-rays it is very important that no tumor goes undetected and less important if a tumor is falsely detected (recall over precision). A completely different task could be providing product recommendations to customers based on previous purchases. For this task it is of higher importance that the products recommended are actually good suggestions and less important that all possible ”good suggestions” are shown (precision over recall).

3.5.6 Error rate

The error rate is the ratio of incorrect predictions over the total number of observations.

$$errorrate = \frac{FP + FN}{TP + TN + FP + FN}$$

3.6 Data division

In machine learning tasks it is a common practice to split the obtained data set into separate disjoint subsets for training, validation and testing. The training set is used to fit the model. It is from this set that the model learns patterns. The validation set is used to repeatedly evaluate the model. Since the validation set is disjoint from the training set it will provide new data for the model. From the results of the evaluation of the validation set, parameter tuning can be performed. In conclusion the model indirectly learns from the validation set. Lastly the test set is used to evaluate the final model. No further changes should be made to the model after evaluating against the test set as these results are seen as the actual performance of the model.

4 Methods and data

4.1 Data

Data was retrieved from a frozen version of VigiBase, containing all reports up to 5th of January 2020. From this source *9,869,169* rows of annotated data could be retrieved, each containing one verbatim and one label. The verbatims are the inputs to our system, the observations that we want to classify. They are freely entered text containing a single up to multiple tens of words, expressing ADRs in multiple different languages. The labels are the preferred terms (PTs) from the MedDRA hierarchy (see section 2.2) that the verbatims have been mapped to. This mapping is done by an expert while making the ICSR. MedDRA is, like VigiBase, ever changing. In this project MedDRA version 22.1 was used [5]. In this version there are 23,954 PTs, however in our labeled data set only 16,408 ($\sim$ 68,5%) are represented.

4.1.1 Explorations

An initial approach was exploring the data to get a feeling of what kind of difficulties there were and what methodologies could be fitting to solve them. An early finding was that even as the verbatims were pulled from free text fields, they were not all unstructured. A large number of verbatims were already in the form of LLTs, the lowest level term of the MedDRA hierarchy.

The verbatims are expressed in free text and can contain all sorts of characters. In the data there exist alphabetical characters, special characters and numbers. They can contain abbreviations and spelling mistakes to list a few. Among the English verbatims a common use of non-alphabetic characters are numbers used to report on medical measurements. Examples of verbatims including such measurements are ”mxd raised 1.7x10ˆ9/l”, ”lipase (over 4000u/l)” and ”high white blood cell count 80”.

The verbatims in the given data set are of varying length, consisting of between 1 to 53 words. However, around 70% of the verbatims only contain up to 3 words. For the purpose of word counting a word was defined to be a number of characters separated from other characters with spacing. ”他服後倒了” will therefor count as 1 word and ”increase in the white blood count” is 6 words.

Each row of data is only mapped to one class, although the verbatim can contain multiple reported side effects. This means that even though there will only be a single correct class for each row of data (that is our gold standard), there might be multiple fitting classes for the verbatim. This will complicate our classification process and is

something to keep in mind when evaluating the algorithm.

As presented in section 2.2 the PT-level in MedDRA contains 23,954 terms. Having this enormous selection of labels means dealing with an extreme multi-class problem. When examining the annotated data the imbalances of the classes is a fact. Looking at the training data (which will be gathered in section 4.2.2) only 13,978 classes are represented, meaning that we already lost ∼42% of possible classes. In table 1 the imbalances of the data becomes clear. Even if the training data set contains 13,978 classes, the table shows that only the 100 most common classes are used to label 54,9% of the data set. This implies that the 45,1% left is split between the other 13,878 classes in different measures.

Number of PTs	Data covered
100	54,9 %
500	80,7 %
1,000	89,1 %
2,000	94,9 %
3,000	97,1 %
4,000	98,2 %
5,000	98,8 %
7,000	99,5 %
10,000	99,8 %

Table 1: This table shows how much of the training set that is covered for a number of PTs most commonly used for labeling this set

4.2 Preparations

4.2.1 Language filtering

Since the data consists of verbatims from ICSRs retrieved from countries all over the world, multiple languages will be present in the data. This thesis will be limited to working with English verbatims and therefore it is important to find a method that can successfully separate the English from the non-English verbatims. A number of methods where tested to find the most effective one for this task. In this section the different methods are presented. The methods were all evaluated on the same set of 2000 data rows that were randomly sampled from the whole data set. The 2000 verbatim were manually labeled as ”English” or ”non-English”, resulting in 478 non-English and 1522 English verbatims.

Sort by country
An early and easy approach was to examine if we could simply choose to include data from countries where a majority was written in English. The countries were chosen based on a manual overview of the data. The countries chosen were: United Arab Emirates, Australia, Canada, Egypt, United Kingdom of Great Britain and Northern Ireland, Greece, India and Korea. However this method led to a giant data loss since many countries had to be excluded because of the presence of non-english verbatims, even though there was a lot of useful english data also present in these data sets.

Regular expression
When examining the differences in the English vs Non-English data it was clear to see a common difference in the characters present. There was data containing only non-alphabetical characters (Chinese/Japanese) and data containing vowels like "à, é and ï" (French/Italian). By using regular expression these rows could be found and discarded.

Python libraries
The python library langid was used to remove some non-English rows. The function langid.classify(verbatim) was invoked for every verbatim. Every call made with a verbatim that resulted in "en" (classified as English by langid) was kept and all other verbatim were removed.

Dictionary
A dictionary was created using a combination of all words in the lexical database WordNet [11] and all the words present in any Lowest Level Term in MedDRA (any numbers or special characters where not added). When evaluating this method each verbatim was split up into seperate words and each word compared against the dictionary. The verbatim then received a score of $score = \frac{englishWords}{words}$ where *words* refer to the number of words in the verbatim and *englishWords* the number of words in the verbatim found in the dictionary. Finally, since the goal of this language sorting was to make sure that all data we operate the project on is in actually in English, we based the threshold of the score upon the precision. The precision was set to be .99 allowing for a .01 error rate of non-English verbatim.

From evaluation of the results (see section 5.1) the final choice for retrieving only the english data was a combination of the *Dictionary* and *Regular expression* methods mentioned above. Using this method left 6,986,110 rows of data ($\sim$70% of the original set).

4.2.2 Division of data

After filtering away non-english verbatims, the remaining 6,986,110 rows were divided in a three-way split of 70% training data, 10% validation data and 20% test data. Di-

vision was made with the data sorted over time, based on ReportID:s, meaning that we are using old data to predict newer data. If the data or how it is labeled has changed over time, this might be noticeable during the evaluation. If the sets would instead have been divided by randomly sampling data from across the whole set, the possible data changes over time would go unnoticed.

4.2.3 Preprocessing

The only preprocessing that was done explicitly was lower-casing all the data. This since many cases of verbatims in all capital letters were observed in the data. It was earlier detected that multiple verbatims were structured as LLTs and the lowercasing will make sure that verbatims such as *"HEADACHE"*, *"Headache"* and *"headache"* are handled as equals before performing any string matching.

4.3 Modules

The modules are different approaches taken to solve the problem of this thesis. They are based on simple string matching algorithms as well as the more advanced technology of the BERT model. The reason for creating these modules was to compare the results of different algorithms as well as exploring if a combination of different algorithms would be more beneficial for solving the problem.

4.3.1 String matching - LLT

From explorations of the training data it was discovered that there were many verbatims that were already written as MedDRA terms. To further explore this finding, a string matching algorithm for classification was constructed. The algorithm was designed to compare each verbatim to a dictionary consisting of all LLTs. All the LLTs were lowercased to match the lowercased verbatims. If a match was found, that verbatim would be classified to the LLTs corresponding PT.

4.3.2 String matching - training

The second algorithm used the same approach of string match comparisons, but comparing the verbatim we want to classify to the verbatims in the training data. We wanted to make use of previous data by classifying accordingly. If any match was found, the

verbatim would be labeled as it was labeled in the training data. An initial problem with this approach was that the very same verbatim expression can be labeled differently in the training data, resulting in multiple labeling options. To solve this problem all the verbatims in the training data was compiled into a dictionary of distinct verbatim expressions. Each distinct verbatim worked as a dictionary key connected to a PT label value. When creating the dictionary, each distinct verbatim would get the PT that it was most commonly labeled as in the training data.

4.3.3 BERT

We used a pre-trained BERT base model and fine tuned it for our classification task using the 4,890,274 rows present in the training data set and trained for 4 epochs. The input layer was set to 32 and the output layer to 5,000, meaning we input 32 tokens and have 5,000 possible classes as output. As can be seen in table 1 considering the whole data set the 5,000 most commonly used labels covers 98.8% of the data labeling.

When using the fine tuned BERT model for classification it produces an output layer of logits, with a layer size corresponding to to the number of classes, in our case 5,000. We take the softmax of the logits and classify the verbatim to the class with the corresponding logit of highest softmax value.

Thresholds
Because the verbatims are constructed in free text fields some might be very difficult to classify. They could for example include measurements, abbreviations, multiple symptoms or other ambiguity. In order to avoid misclassification, one option would be to not classify the most difficult verbatims. The values of the logit layer reflects on the confidence of the BERT model making good predictions. By taking the values of the logits into account in the classification process we could decide how confident we need the model to be.

Since the value of the logits represent the confidence of the corresponding class being correct, we decided to use this value as a threshold. This is referred to as the val-threshold. As discovered in the explorations (section 4.1.1) there are cases where multiple side effects are reported in the same verbatim. This could lead to multiple logits getting high values. To increase the certainty of the prediction we make, we chose to also include another threshold based on the difference between the highest and second highest logit. The smaller the difference between the two highest logits, the less certainty that the highest value results in a correct prediction, as we have high confidence in multiple classes. The second threshold is referred to as the diff-threshold.

In practice each verbatim is classified with the PT corresponding to the highest logit

value, if the logit value exceeded the val-threshold and the difference between the highest and second highest logit values exceeded the diff-threshold. For any verbatim for which the logit values do not satisfy the thresholds, no classification is made. But even if the confidence of BERT's prediction is not seen as good enough to classify a verbatim, there might still be good suggestions among the top predictions. With any verbatim that is left unclassified the top 5 highest ranked PTs (based on the highest logits from the BERT output layer) will therefore be provided. If these verbatims are left for manual mapping it means that the 5 suggestions could be a resource in the process.

Three different modules of BERT were created with the thresholds based on the maximum f_1-score, the maximum $f_{0.5}$-score and the maximum $f_{0.2}$-score for both the highest value logit (val-threshold) and the difference between the highest and second highest logit (diff-threshold).

4.4 Evaluation

The modules were evaluated in different combinations to find a pipeline of modules that gave us the best results on the validation set. The pipelines were evaluated by accuracy (number of correct predictions) and error rate (number of incorrect predictions). If this pipeline were to be used in clinical trials it would be very important to not be making incorrect mappings. Because of this we want to keep the error rate as low as possible.

The evaluation of the pipelines was based on comparisons between the predictions and the gold standard (GS) as well as an error analysis performed on the different modules of the final pipeline. The basis for the error analysis was produced by a panel of terminology specialists at UMC. They were asked to review 200 randomly selected verbatims from each module, where the predicted PT was different from the GS. The terminology specialists looked at each verbatim and chose a PT that they would code that verbatim to. Each verbatim was then given a label that shows how the terminology specialists' PT relates to the (by our system) predicted PT and the gold standard PT.

In table 2 the possible labels and their corresponding description are shown. **TS** refers to the PT chosen by the terminology specialists, **GS** refers to the PT that is our gold standard and **P** refers to the (by our system) predicted PT. The label "-1" was given when the specialists felt there was not enough information in the verbatim to give it a PT label. "0" was given when the specialists chose a PT that was not predicted by our system, neither the gold standard. "1" is the case when the specialist chose the same PT as our system predicted, and "2" when they chose the same as the gold standard. The last label "3" was chosen when the verbatim contained information linking to multiple PTs, were the specialists would split up the verbatim and code the parts separately.

As mentioned in section 4.3.3 the output of the manual mapping module is five suggested PTs. In this evaluation ”0” was given when none of the five suggestions (or the GS) was the same as the specialists’ choice and ”1” when one of the five suggestions matched their chosen PT.

Label	Description
-1	The verbatim has no fitting label
0	(TS != P) AND (TS != GS)
1	TS = P
2	TS = GS
3	The verbatim should be coded to multiple labels

Table 2: The labels used for evaluating the incorrect samples from the different modules

5 Results

5.1 Language sorting

The results of methods presented in section 4.2.1 for sorting out English data is shown in the table 3. The different methods were evaluated on a set of 2,000 randomly sampled rows which were manually labeled as English or non-English.

From table 3 we find that the method *Countries* had a high precision of almost .99, but a low recall of .71. It shows that when data was solely selected from a few countries, the selected data was mainly in English. However a lot of English data (from other countries) were filtered out.

The *RE* approach had a prefect recall of 1 meaning that all the English samples were classified as English. The low precision of around .78 however shows that this method did not filter out non-English data strictly enough as there were still much left in the data set.

Python LangID had relatively good results in both precision (.92) and recall (.89) but was out-performed by the *Dictionary* method that got a precision of .99 and a recall of .96.

When combining the *Dictionary* method with the *RE*, having a perfect recall, the precision was slightly improved without any negative effect on the recall. This led to the best precision and fscore of all the methods which led to the final choice of the *Dictionary + RE* as the method for language sorting.

Method	Precision	Recall	F_1-score
Countries	0.9899	*0.7063*	*0.8244*
RE	*0.7841*	**1**	0.8790
Python LangID	0.9192	0.8890	0.9038
Dictionary	0.9898	0.9560	0.9726
Dictionary + RE	**0.9905**	0.9560	**0.9729**

Table 3: The different methods for distinguishing english verbatims, evaluated with a sample set of 2,000 manually labeled rows of data.

5.2 Thresholds

Table 4 shows the three different thresholds used for the BERT module. Each one of these three thresholds corresponds to one ”diff”-threshold and one ”val”-threshold. The values used for ”diff” (difference in highest and second highest logit) and ”val” (the value of the highest logit) were chosen because they were maximizing three different f-scores (F_1-score, $F_{0.5}$-score and $F_{0.2}$-score).

Threshold	Diff	Val
F_1	0.4	9.2
$F_{0.5}$	2.0	10.8
$F_{0.2}$	4.1	12.6

Table 4: The different thresholds used to improve the predictions made by BERT

When evaluating the validation set on the BERT module, the distribution of the correctly classified verbatims are displayed to the left in figure 5. The y-axis shows the value of the highest value logit (referred to as ”val”). The x-axis shows the difference between the highest and second highest value logits (referred to as ”diff”). To the right in figure 5 the distribution of the incorrectly classified data is displayed. It may look as if the graph showing the incorrect predictions has more data because of the intensity of the heat map. However this is a result of the graphs being generated separately, thereby the intensity is not comparable.

Comparing the distributions in the graphs, the incorrectly classified verbatims are much more centered towards the lower values of both ”diff” and ”val”, while the correctly classified verbatims are more centered around higher values. The three different boxes present in both graphs show how the three different thresholds introduced in table 4 affect the number of occurrences of correct and incorrect predictions in the validation set. Everything inside the box will be left unclassified for that specific threshold, while everything outside the box is classified by BERT. As can be seen the higher the threshold boundary, the fewer incorrect predictions are made. However, this also means that more of the correct predictions will be left unclassified. The overlapping distribution of the two graphs shows that no threshold will completely eliminate the errors. Choosing a threshold will really be a trade-off of getting the best possible accuracy without overstepping the accepted error rate.

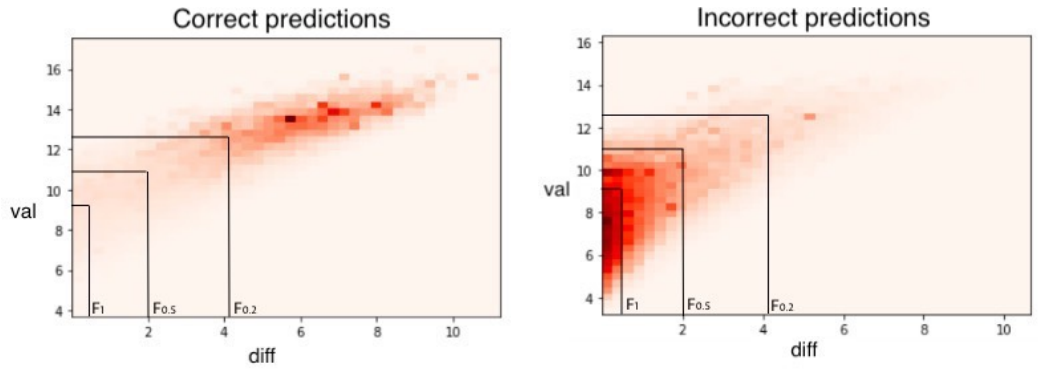

Figure 5: Two heat map graphs displaying the data distribution from the by BERT correctly and incorrectly predicted verbatims. The graphs also show the three different thresholds from table 4. The graphs were generated separately and the intensity can therefore not be compared.

5.3 Pipeline

A number of possible pipelines were evaluated based on different combinations of modules presented in section 4.3 with different thresholds presented in table 4. The pipelines are shown in table 5 and the evaluation is done on the validation set.

In *pipeline 1* the BERT model is classifying the whole validation set. The accuracy is 0.8376 which can be interpreted as a good result considering the difficult examples that exists in the data set (see explorations of the data in section 4.1.1). However the drawback of this pipeline is the high error rate of 0.1624, which would never be accepted for the potential use case of the system. To let BERT classify every verbatim seemingly was a too simple approach. When restricting the classification with different thresholds in *pipeline 2-4* both the accuracy and the error rate drops as the threshold boundary is increased.

The results of *pipeline 5* shows that more than half of the verbatims in the validation set are actually in the form of LLTs. Even as the error rate for this pipeline setup is relatively low (around 1%) the presence of these errors is still surprising. They occur when a verbatim, that is also a LLT, is mapped to another PT than the one that the LLT is corresponding to.

When apart from LLTs also basing the predictions on training data (meaning classifying a verbatim in the validation set as it was classified in the training set) the error rate increases from 1,1 % to 3,58% (comparing *pipeline 5 and 6*). This shows that the exact same verbatim can be classified differently, which strengthen the hypothesis that the mapping is sometimes based on additional information. Because of the high error rate

brought by basing future predictions on exact matches in the training data (as can be seen in *pipeline 6, 8 and 9*) this module was not kept in the final pipeline.

Pipeline 7 is a combination of the modules separately run in *pipeline 4 and 5*. When comparing these pipelines it is found that the combination really improves the accuracy without increasing the error rate notably. The improved accuracy of adding the LLT-matching before predicting with BERT can possibly be a result of BERT being limited to a number of classes (in this case 5,000). The LLT-matching can predict any of the almost 24,000 classes as long as the verbatims are in the form of LLTs, which a notable section seemingly is.

Pipeline 10 can also be compared to *pipeline 7*. The pipelines consist of the same modules but using different thresholds for the BERT prediction. With the relaxed threshold in *pipeline 10* the accuracy approximately increases from 71% to 79%. However *pipeline 7* was considered a better option as the error rate more than doubled with the relaxed thresholds.

id	Module 1	Module 2	Module 3	Accuracy	Error rate	Unclassified
1	BERT	-	-	**0.8376**	*0.1624*	**0**
2	BERT F_1	-	-	0.8054	0.0714	0.1232
3	BERT $F_0.5$	-	-	0.7520	0.0334	0.2146
4	BERT $F_0.2$	-	-	0.6258	0.0111	0.3631
5	LLT	-	-	*0.5549*	**0.0110**	*0.4341*
6	LLT	Training	-	0.7046	0.0358	0.2596
7	LLT	BERT $F_0.2$	-	0.7106	0.0169	0.2725
8	LLT	BERT $F_0.2$	Training	0.7649	0.0387	0.1964
9	LLT	Training	BERT $F_0.2$	0.7612	0.0389	0.1999
10	LLT	BERT $F_0.5$	-	0.7923	0.0365	0.1712

Table 5: Evaluation of possible pipelines on the validation set

5.4 Final model

For the final model, pipeline number 7 from table 5 was chosen because of its low error rate and relatively high accuracy. In figure 6 the details of how the selected pipeline works are shown. The first two modules *LLT* and *BERT* are the automatic part of the pipeline that perform classification. The resulting correct and incorrect classifications are displayed in the figure. The data that could not get automatically classified is left to the last module *Manual mapping* where the 5 top suggestions from BERT are given. The figure shows if the correct label is found within the top 5 suggestions or not.

When evaluating the pipeline on the validation set the accuracy is measured to 71.06% (LLT-matching: 55.49% and BERT: 15.57%) compared to the 80,21% (LLT-matching: 66.50% and BERT: 13.71%) when evaluating on the test set. The increase in accuracy can evidently be explained by a bigger part of the test set consisting of LLT terms, compared to the validation set. As mentioned in section 4.2.2 the sets were split over time and the increase in LLTs among the later received verbatims could be explained by how newer reporting systems choose to input this information.

The results of the manual mapping module shows that the suggestions generated by BERT are rather accurate. For the validation set about 77% of the verbatims left for manual mapping has the correct label found in the top 5 suggestions, for the test set the corresponding results were about 82%. This shows that the suggestions could actually be a useful resource for someone who were to map these verbatims manually.

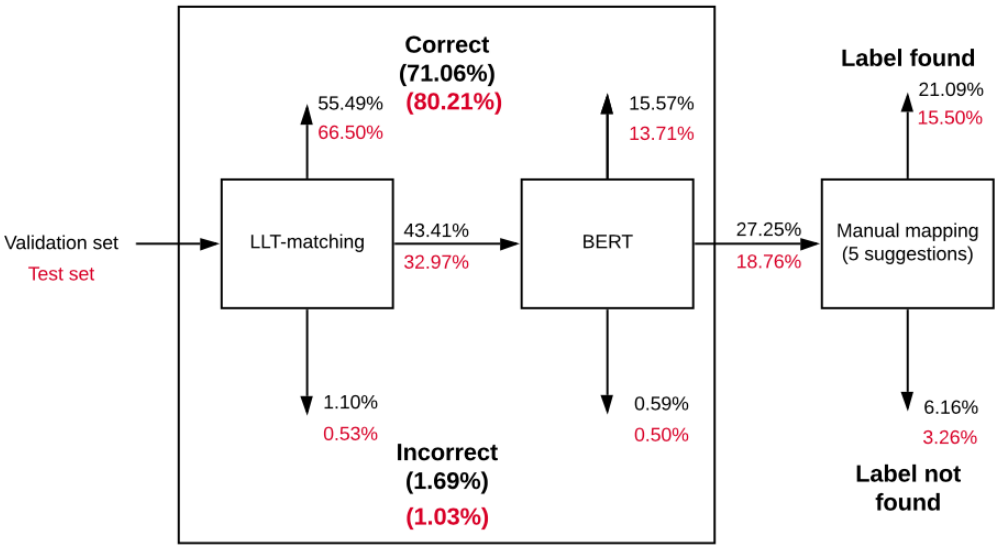

Figure 6: Evaluation of the final pipeline. The results on the validation set are shown in black while the results of the test set are shown in red. The percentages are given as fractions of each entire data set.

5.5 Classification examples

Table 6, 7 and 8 shows examples of data from the test set that is incorrectly predicted in the different modules of the pipeline. The examples were selected to show as many different scenarios of wrong predictions and classifications as possible.

Table 6 displays verbatims in the form of LLTs that was incorrectly classified. The verbatim *pregnant* is classified as *Pregnancy* but was labeled in the validation set as *Exposure during pregnancy*. Seemingly the label contains more information than the verbatim, making the classification, which is solely based on the verbatim, very difficult. Another example from the same table is the verbatim *feels bad* that is labeled *Malaise* but classified as *Feeling abnormal*.

In table 7 the classification *Lipase increased* made by BERT for a verbatim labeled as *Hyperlipasaemia* is found. The classification is incorrect as the correct label was not captured, however hyperlipasaemia is a diagnosis given for someone with increased lipase. A similar case, found in the same table, is a verbatim that is classified as *Pyrexia* (which is the diagnostic term for having a fever) while the label is set to *Body temperature increased*.

Presented in table 8 are examples of verbatims that were not classified by BERT and where the label was not found in the top 5 suggestions. The verbatim *swollen tongue, dyspnoea, dysphagia, drooling, cough* is labeled as *Drooling*. In the top 5 suggestions we find labels such as *Dyspnoea*, *Cough* and *Swollen tongue*, which are all symptoms expressed in the verbatim.

Verbatim	Label (PT)	Classification (PT)
pregnant	Exposure during pregnancy	Pregnancy
application site reaction	Skin reaction	Application site reaction
painful rash	Pain of skin	Rash
feels bad	Malaise	Feeling abnormal
bunion operation	Foot deformity	Bunion operation

Table 6: Examples of data incorrectly classified by LLT-matching

Verbatim	Label (PT)	Classification (PT)
lipase (over 4000u/l)	Hyperlipasaemia	Lipase increased
high temperature all over the body	Body temperature increased	Pyrexia
imbalance	Walking disability	Balance disorder
increased anger	Mood swings	Anger
does not have af	Off label use	Atrial fibrillation

Table 7: Examples of data incorrectly classified by BERT

Verbatim	Label (PT)	Top 5 suggestions (PT)
multiple tumors in mediastinum	Mediastinum neoplasm	['Colon cancer', 'Malignant neoplasm progression', 'Neoplasm malignant', 'Second primary malignancy', 'Neoplasm']
tablet breakage	Prescribed underdose	['Product physical issue', 'Product quality issue', 'Wrong technique in product usage process', 'Product container issue', 'Product complaint']
raised ketone	Acetonaemia	['Blood ketone body present', 'Urine ketone body present', 'Ketoacidosis', 'Dehydration', 'Bone disorder']
fracture (traumatic)	Femoral neck fracture	['Fracture', 'Upper limb fracture', 'Fall', 'Stress fracture', 'Multiple fractures']
swollen tongue, dyspnoea, dysphagia, drooling, cough.	Drooling	['Dyspnoea', 'Cough', 'Swollen tongue', 'Anaphylactic reaction', 'Dysphagia']

Table 8: Examples of data that was unclassified by BERT (because the logit values did not pass the thresholds) where the correct label was not found in the top 5 suggestions.

5.6 Error analysis

In this section the results of the error analysis (see section 4.4) is shown in table 9. For this evaluation 200 random samples of incorrectly predicted verbatims were chosen from each module (LLT-, BERT- and the manual mapping module) and classified by terminology specialists at UMC.

For the LLT module the specialists classified all the 200 verbatims with the same PT as our system. They also noted that in 105 of the 200 cases (just over 50%) the verbatim was the exact same as the PT chosen by them and our LLT module. These results are not surprising as the LLT module is based on direct string matches and should therefore hypothetically have a high accuracy. As mentioned previously, looking at table 6, the correct labels sometimes seems to be based on more information than solely the verbatim. In this error analysis however, the terminology specialists, similarly to our pipeline, only consider and have access to the verbatim.

For the BERT module in 149 cases (covering 74,5% of the samples) the specialists chose the same PT as BERT had predicted, compared to the 15 cases (7,5%) where they chose the gold standard PT. There were also a few cases were the verbatim could not be classified either because the specialists wanted to split it up as it should be mapped

to multiple PTs (label ”3”) or because it could not be mapped to any PT (label ”-1”). 7% of the set was labeled as ”-1” and two examples of that are the verbatims: *”to be high”* and *”does not have af”*. Another 4.5% needed to be split up, one example is the verbatim: *”hyperglycemia/pain in hands/swelling on feet”*.

When it comes to the last module, the manual mapping module, the chosen labels are more distributed over the different options then for the other modules. In 27.5% of cases the specialist chose the same PT as one of the five suggestions from our system while in 26.7% they chose the same as the gold standard. Another 24.5% of the verbatims needed to be split up, 8% of verbatims could not be labeled and in 13.5 % of cases none of the five suggestions nor the gold standard matched the PT chosen by the specialists.

Label	LLT module	BERT module	Manual mapping module
-1	0 (0)	0.070 (14)	0.080 (16)
0	0 (0)	0.065 (13)	0.135 (27)
1	1 (200)	0.745 (149)	0.275 (55)
2	0 (0)	0.075 (15)	0.265 (53)
3	0 (0)	0.045 (9)	0.245 (49)

Table 9: The results of the error analysis are shown as fractions for each module and the number of occurrences of the labels are shown in parentheses. The description of the labels can be found in table 2.

6 Discussion

The final pipeline design consists of three different modules namely the *LLT-matching* module, the *BERT* module and lastly the *Manual mapping* module. BERT is by far the most important and complex module as it can make automatic predictions for all kinds of English verbatims. The shortcomings of the other modules are that the *LLT-matching* can only make predictions for verbatims structured as LLTs and the *Manual mapping* is not automatic as it involves human judgement. However through evaluation of the pipelines in table 5 it shows that all the above mentioned modules serves a purpose in the final pipeline, leading up to that choice.

The overlapping distribution of the correctly and incorrectly predicted verbatims (figure 5), in the graphs displaying information from the output layer of BERT, shows that no threshold will completely eliminate the errors. Choosing a threshold is really a trade-off of getting the best possible accuracy without overstepping the accepted error rate. The threshold could be made more or less strict depending on the task of the model and its accepted error rate.

When comparing the correct labels with either the classification in tables 6 and 7 or with the corresponding suggestions in table 8 it is clear to see that one major problem for our model is the great number of classes and how similar they can be. We see that even if the classification or given suggestions are incorrect (as they differ from the actual label) the classes are in several cases similar in meaning. Another problem seems to be the limited information gained from solely using the verbatim as input, as some labels seems to be based on more information then what is given in the verbatim.

From the error analysis in section 5.6 it was discovered that, assuming that the verbatim is the only source of information, there are cases where the gold standard might not be appropriate. This could be a result of many things. It has been mentioned before that the gold standard might be based on additional information and as the mapping is done manually it would not be surprising for some human errors to occur as well. The specialists that performed the evaluation found cases of verbatims that, according to them, could not be mapped or that should be split up into multiple verbatims before being mapped. They also found that for many of the verbatims the model's classifications, which were evaluated as errors, were actually matching their own proposed labels. When looking at errors from the BERT module, the specialists had chosen the same label as BERT classified in 149 of the 200 samples (74,5%).

7 Conclusion

The aim of this thesis was to answer the question introduced in section 1.2, namely: *How can we use verbatim descriptions of adverse drug reaction to create an automatic mapping to MedDRA terms?*. This question can now be answered with the help of the more specific questions formulated.

- *How do we handle lack of training data?*
  Because of the great class imbalances in the training data, the BERT model was limited to consider the 5000 most common classes. The drawback of this choice was that the other 18,954 possible PT classes were excluded. However the string matching performed in the LLT-module still provided the possibility to match any class. As the 5000 classes that BERT was limited to covered 98,8 % of the training data (see table 1) we can conclude that the excluded classes are seldom used as labels.

- *How do we deal with our training data being inadequate?*
  The model was developed from the information in the preexisting mappings provided in the training data, namely our gold standard. In explorations of the data it was found the the same verbatim can be mapped to different classes and in some cases it seemed as if the mappings were based on more information then what our model was given (the verbatim). The errors that were obtained when evaluating the model by comparing the results with the gold standard, were further evaluated by terminology specialists at UMC. This evaluation showed a disagreement in suitable labels for the verbatims. It showed that the gold standard might not always be an adequate representation solely basing the input on the verbatims and that there are classifications made in our test set that could be considered correct even if they are part of the error rate.

- *Can machine learning techniques be used to improve the results?*
  As can be seen in the evaluation of the final model (figure 6), a large piece of the test set could be correctly classified with a simple string matching approach in the LLT-module. However the use of BERT shows that the more difficult verbatims could with a low trade-off of incorrect predictions increase the accuracy of the model.

The final model resulted in an accuracy of approximately 80% and an error rate of approximately 1%, which was viewed by the job initiator UMC as good results. The goal of automatically mapping verbatim expressions of ADRs to MedDRA terms was partially fulfilled. For this task it is very important to keep a low error rate and because

of this requirement, with the approach taken, it was concluded that some verbatims are too difficult to map automatically. The solution was instead to keep the most difficult verbatims unclassified and instead provide suggestions that can support the process of manually mapping them.

8 Future work

In some coding events the LLT is actually preferred to use as the MedDRA label (instead of the PT). A future work is therefore to improve the system to be able to classify to this more specific MedDRA term. As the MedDRA terms are arranged in a hierarchical structure this could suggestively be accomplished by extending the system to predict a LLT based on the ones that are related to the already predicted PT.

From the evaluation of the incorrect results in the pipeline (table 9), it can be concluded that the coding specialist in some cases base their choice of label upon more than just the verbatim. In order for the system to get the correct predictions we might therefore want to consider additional information related to the reported reaction instead of solely considering the verbatim. However, this of course depends on if basing the label on additional information is desirable or not, which should first be determined.

The BERT model that was used in this project is the original model trained on general English text. Since BERT's release other more specialized models have been proposed. One example is BioBERT which in addition to BERTs pre-training on English Wikipedia and BooksCorpus also includes text from biomedical domains in order to increase performance on tasks with text from this field [12]. Another example is ClinicalBERT that with its training on clinical text distinctly improves the performance from BERT, when evaluated on data within that domain. [13] A future improvement could be to examine if using domain specific BERT models would improve the performance compared to the general BERT.

References

[1] World Health Organization (2002), *The Importance of Pharmacovigilance*, Available at: https://apps.who.int/iris/bitstream/handle/10665/42493/a75646.pdf

[2] World Health Organization, *The WHO Programme for International Drug Monitoring*, Available at: http://www.who.int/medicines/areas/quality_safety/safety_efficacy/National_PV_Centres_Map/en

[3] VigiBase, *The unique global resource at the heart of the drive for safer use of medicines*, Available at: https://www.who-umc.org/vigibase/vigibase

[4] Mozzicato P (2009), *MedDRA An Overview of the Medical Dictionary for Regulatory Activities*, Available at: https://www.researchgate.net/publication/233524508_MedDRA_An_Overview_of_the_Medical_Dictionary_for_Regulatory_Activities

[5] MedDRA (2019), *Introductory Guide MedDRA Version 22.1*, Available at: https://admin.new.meddra.org/sites/default/files/guidance/file/000354_intguide_22.1.pdf

[6] Goldberg Y (2017), *Neural Network Methods for Natural Language Processing*

[7] Jurafsky D. and Martin J. H. (2019) *Speech and Language Processing*, Available at: https://web.stanford.edu/~jurafsky/slp3/

[8] Mikolov T, Chen K, Corrado G and Dean J (2013), *Efficient Estimation of Word Representations in Vector Space*, Available at: https://arxiv.org/pdf/1301.3781.pdf

[9] Devlin J, Chang M-W, Lee K and Toutanova K (2019), *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, Available at: https://arxiv.org/pdf/1810.04805.pdf

[10] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L and Polosukhin I (2017), *Attention is all you need*, Available at: https://arxiv.org/pdf/1706.03762.pdf

[11] Wordnet, Princeton University, Available at: https://wordnet.princeton.edu

[12] Lee J, Yoon W, Kim S, Kim D, Kim S, So C H and Kang J (2019) *BioBERT: a pre-trained biomedical language representation model for biomedical text mining*, Available at: https://arxiv.org/pdf/1901.08746.pdf

[13] Huang K, Altosaar J and Ranganath R (2019), *ClinicalBert: Modeling Clinical Notes and Predicting Hospital Readmission*, Available at: https://arxiv.org/pdf/1904.05342.pdf