



DEGREE PROJECT IN TECHNOLOGY,
SECOND CYCLE, 30 CREDITS
STOCKHOLM, SWEDEN 2021

Explainable Antibiotics Prescriptions in NLP with Transformer Models

Omar Emilio Contreras Zaragoza

Authors

Omar Emilio Contreras Zaragoza <oecz@kth.se>
Master in Machine Learning
KTH Royal Institute of Technology

Place for Project

Stockholm, Sweden

Examiner

Amir H. Payberah
KTH Royal Institute of Technology

Supervisors

Magnus Boman
KTH Royal Institute of Technology

Karl Fredrik Erliksson
Peltarion

Marie Korneliusson
Peltarion

Abstract

The overprescription of antibiotics has resulted in bacteria resistance, which is considered a global threat to global health. Deciding if antibiotics should be prescribed or not from individual visits of patients' medical records in Swedish can be considered a text classification task, one of the applications of Natural Language Processing (NLP). However, medical experts and patients can not trust a model if explanations for its decision are not provided. In this work, multilingual and monolingual Transformer models are evaluated for the medical classification task. Furthermore, local explanations are obtained with SHapley Additive exPlanations and Integrated Gradients to compare the models' predictions and evaluate the explainability methods. Finally, the local explanations are also aggregated to obtain global explanations and understand the features that contributed the most to the prediction of each class.

Keywords

Transformer models, NLP, Explainable AI, Medical domain, Antibiotics prescription, SHAP, Integrated Gradients

Abstract

Felaktig

utskrivning av antibiotika har resulterat i ökad antibiotikaresistens, vilket anses vara ett globalt hot mot global hälsa. Att avgöra om antibiotika ska ordinerats eller inte från patientjournaler på svenska kan betraktas som ett textklassificeringsproblem, en av tillämpningarna av Natural Language Processing (NLP). Men medicinska experter och patienter kan inte lita på en modell om förklaringar till modellens beslut inte ges. I detta arbete utvärderades flerspråkiga och enspråkiga Transformers-modeller för medicinska textklassificeringsproblemet. Dessutom erhöles lokala förklaringar med SHapley Additive exPlanations och Integrated gradients för att jämföra modellernas förutsägelser och utvärdera metodernas förklarbarhet. Slutligen aggregerades de lokala förklaringarna för att få globala förklaringar och förstå de ord som bidrog mest till modellens förutsägelse för varje klass.

Nyckelord

Transformator-modeller, NLP, förklarbar AI, medicinsk domän, antibiotikarecept, , SHAP, Integrated Gradients

Acknowledgements

I would like express my gratitude to Marie Korneliusson and Karl Fredrik Erliksson, my supervisors at Peltarion, for their advice, the weekly discussions and feedback during this work. I would like to thank all the people in Peltarion for letting me work in such an interesting project and all the support given. Moreover, thank you to the people involved in the Swedish Medical Language Data Lab, and especially Folktandvården Västra Götalandsregionen for providing the dataset and research problem. Furthermore, thank you Amir H. Payberah for the guidance during the project and Magnus Boman for your help during the work. Finally, I want to thank my family for the strong support during my studies, this accomplishment is also yours.

Acronyms

ALBERT	A Lite BERT
ANUG	Acute Necrotizing Ulcerative Gingivitis
AI	Artificial Intelligence
BERT	Bidirectional Encoder Representation from Transformers
CNN	Convolutional Neural Networks
ELECTRA	Efficiently Learning an Encoder that Classifies Token Replacements Accurately
FC	Fully Connected
FFN	Feed Forward Network
GELU	Gaussian Error Linear Unit
IG	Integrated Gradients
ML	Machine Learning
MLM	Masked Language Model
mBERT	Multilingual BERT
mT5	Multilingual Text-to-Text Transfer Transformer
NER	Name Entity Recognition
NLP	Natural Language Processing
NMT	Neural Machine Translation
NSP	Next Sentence Prediction
NaN	Not a Number
RNN	Recurrent Neural Networks
RTD	Replaced Token Detection
RoBERTa	Robustly Optimized BERT Pretraining Approach
SHAP	SHapley Additive exPlanations
SOP	Sentence-Order Prediction
T5	Text-to-Text Transfer Transformer
XAI	Explainable AI
XLM-R	XLM-RoBERTa
WmT5	Why mT5?
WT5	Why T5?

Contents

1	Introduction	1
1.1	Research Questions	2
1.2	Contribution	2
1.3	Thesis structure	3
2	Background	4
2.1	Transformer models	4
2.1.1	Transformer model architecture	5
2.1.2	BERT based architectures	7
2.1.3	Monolingual models	10
2.1.4	Multilingual models	10
2.1.5	Transformer models for the medical classification task	11
2.2	Explainability	11
2.2.1	Integrated Gradients	13
2.2.2	SHapley Additive exPlanations	13
2.3	Domain specific NLP	14
2.4	XAI for NLP	15
2.5	Antibiotics prescription	15
3	Methodology	16
3.1	Dataset	16
3.2	Evaluation of Transformer models	20
3.3	Explanations	22
3.3.1	Local explanations	22
3.3.2	Global explanations	26
4	Results	27
4.1	Classification task	27
4.2	Explanations	28
4.2.1	Local explanations	29
4.2.2	Global explanations	34
5	Conclusions	43
	References	45

Chapter 1

Introduction

The treatment for bacterial infections is usually antibiotics. Nevertheless, bacteria can become resistant to antibiotics, which makes infections harder to treat. This problem is considered a threat to global health since common infections could lead to death. To prevent antibiotics resistance, healthcare professionals have to follow the WHO [32] guidelines to avoid prescribing antibiotics when they are not needed [32].

Dental care accounts for around 7-11% of antibiotics prescribed internationally, whereas, in Sweden, the percentage is approximately 7% of outpatient care (care that does not involve overnight stay in the hospital). Moreover, there exist regional differences in the number of antibiotics prescribed in Sweden, highlighting the importance of thoroughly evaluating the correctness of antibiotics prescription for specific dental visits [24]. As a consequence, in articles like [2], prescription of antibiotics is evaluated, suggesting some criteria for correctly prescribing antibiotics.

Predicting if a patient should be prescribed antibiotics from medical records can be considered a text classification task, one of the popular applications of Natural Language Processing (NLP). In recent years, different Transformer models have demonstrated state-of-the-art results in multiple tasks [3, 6, 21] outperforming previous NLP models. However, these models are pre-trained to be used in English language, to use NLP models for other languages, multilingual models (models pre-trained in multiple languages) have been proposed [7, 23, 45]. Another approach, offering more particular solutions by training non-English models (monolingual models) are presented in [26, 27, 42].

Obtaining a prediction from an NLP model of whether to prescribe antibiotics or not could be helpful to the medical field to reduce unnecessary use of antibiotics. For this work, the dental records used consist of all the information needed to decide if antibiotics should be prescribed or not and individual visits are considered to make an individual decision. Although, medical tasks are hard to approve in real-world scenarios with actual patients if the model is unable to explain the reasoning behind its prediction. In [9] they stress the importance of explanations to

enable trust and acceptance of automated systems and [12] mentions interpretation of predictions as an issue to implement Artificial Intelligence (AI) technologies in medicine applications. Different explainability methods [25, 39], have been proposed to understand the decisions of deep learning models and build trust for domain experts. The explainability methods that this work employs are post-hoc methods, i.e., the explanations are obtained after the model is trained and the decision is taken. Moreover, additional operations to the model are required. Additionally, these methods are also feature attribution methods, they assign a score to each input feature.

This thesis examines the process of using NLP models for a medical text classification task and at the same time, provides explanations of the models' decisions. The goal is to understand their rationales and compare the results with the correct criteria for prescribing antibiotics. If the model is capable of learning the correct criteria, it could be used by medical experts to avoid prescribing antibiotics when they are not needed.

1.1 Research Questions

The research questions are divided into two main subjects: Text classification in the medical domain and explainable predictions for text classification. Each section has its main research question and at least one sub-question.

Text classification in the medical domain

How can an NLP model predict if a patient should be prescribed antibiotics or not from individual visits of patients' dental records using a low-resource language, such as Swedish?

1. How do multilingual and monolingual models' performance compare in predicting if antibiotics should be prescribed?

Explainable predictions for text classification

How can a post-hoc method provide high-quality explanations of an NLP model?

2. What attribution method is better for a medical classification task?
3. How can local explanations be used to obtain global explanations?

1.2 Contribution

The contribution of this work is divided into two sections. Text classification in the medical domain:

1. Evaluation of the performance of monolingual and multilingual models in a medical classification task.

2. In this work, it is demonstrated that Transformer models can be used for a binary classification task using a Swedish medical dataset.
3. Evaluation and analysis of monolingual and multilingual models resulting in a better performance with monolingual models.

Explainable predictions for text classification:

1. It is demonstrated that attribution methods are capable of explaining predictions from NLP models trained with a Swedish medical dataset.
2. Both explainability methods are compared showing that it is easier to interpret SHAP since the method has a sparser explanation, although, IG or SHAP without regularization are more mathematically correct .
3. It is demonstrated that explanations can be used as guidance for data cleaning.

1.3 Thesis structure

The structure of the thesis is the following: In Chapter 2, the background needed for this work is introduced, describing the NLP models and the explainability methods that were employed. Moreover, the related work that other people have done in domain-specific NLP and Explainable AI (XAI) applications is presented, additionally, what the Machine Learning (ML) field has done regarding antibiotics prescription will be mentioned. In chapter 3, the dataset is described; the processes of data cleaning and training the NLP models are specified, furthermore, how the explanations were obtained are detailed. Chapter 4 presents the results obtained from the medical classification task and the local and global explanations and the analysis of those results. Finally, Chapter 5 presents the conclusions from this thesis and some future work is proposed.

Chapter 2

Background

This chapter provides the theory necessary to understand this thesis. First, the NLP models are introduced and then, the explainability methods are presented. The NLP models are selected for being the models with better state-of-the-art results in Swedish or multiple languages. Similarly, Integrated Gradients (IG) and SHapley Additive exPlanations (SHAP) are the two explainability methods selected since they have become very popular for the explanations they obtain. Furthermore, some approaches that the field of NLP and XAI have applied to a particular domain are also presented. Finally, the relevant work where ML has been applied for antibiotics prescription is discussed.

2.1 Transformer models

Up until 2017, the state-of-the-art models for NLP were based on Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN), both methods use an encoder-decoder architecture. A problem with RNNs and CNNs is that they are sequential models, thus, it is impossible to use parallelization and the computation time is significant [41]. The best models used for Neural Machine Translation (NMT) task, before the Transformers, connected the encoder and decoder by an attention mechanism, helping the model focus on different words of a sentence despite the distance between them [41].

In "Attention is all you need" [41], the first Transformer model proposed. The model takes the attention mechanism used before and instead of employing it to connect the encoder and decoder parts, it solely relies on it without using conventional networks like CNNs or RNNs, which allowed parallelization and improved the performance on different NLP tasks. This model is trained to perform translation tasks, achieving better results than previous NLP models.

2.1.1 Transformer model architecture

The Transformer model consists of an encoder-decoder architecture that uses a multi-head attention mechanism, the full architecture is shown in Figure 2.1.1 and an explanation of each part comes afterward.

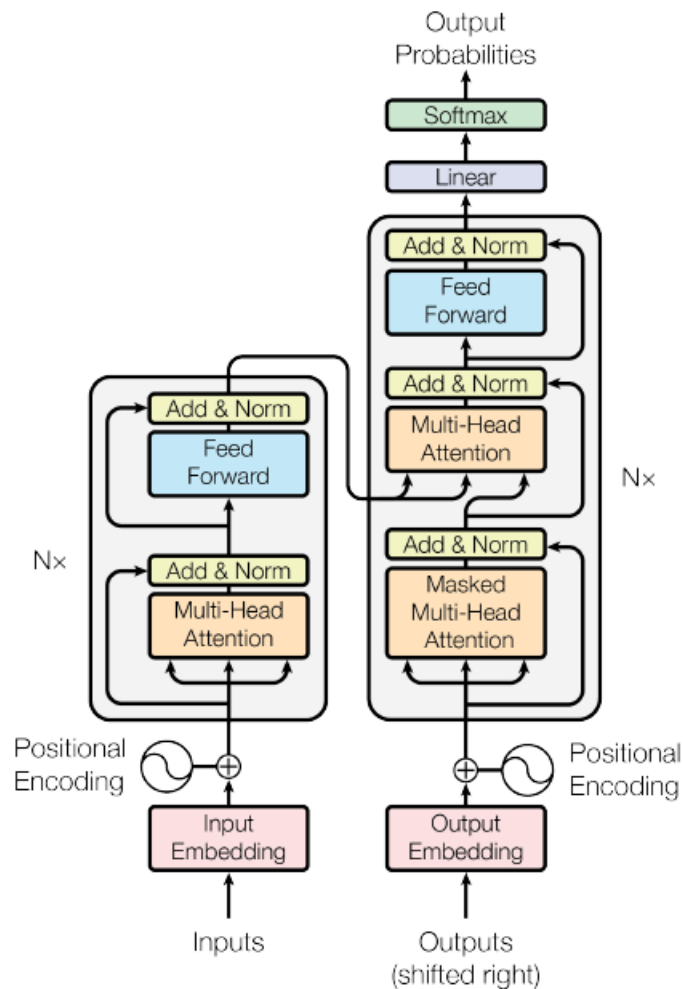


Figure 2.1.1: Architecture of Transformer model [41].

Encoder

The encoder proposed in [41] has six stacked identical encoding layers; each layer consists of two parts, the first one is a multi-head attention and the second one, a Fully Connected (FC) Feed Forward Network (FFN). These two parts use a residual connection followed by a normalization layer.

Decoder

The decoder also has six stacked identical layers; each layer consists of three parts, two of them are exactly the same as the encoders' layers while the middle part of the decoder

(see Figure 2.1.1), is different, it is another multi-head attention over the encoder's output.

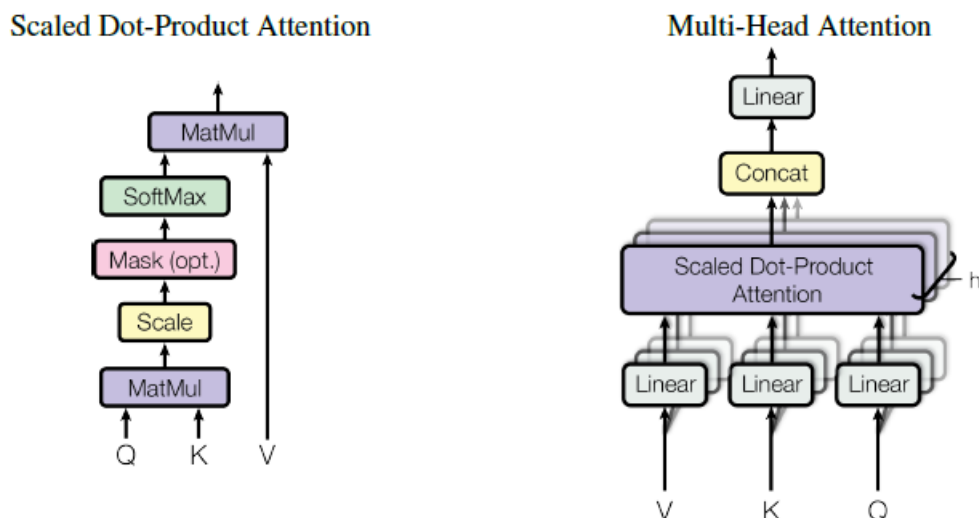


Figure 2.1.2: Attention and multi-head attention functions [41].

Attention

The attention mechanism is a method that Transformer models use to understand better the input by allowing the model to look at multiple words at the same time resulting in an encoding of words that has a better context. During the training phase, the model learns three matrices W^Q , W^K and W^V that try to find similar words in the input. By multiplying these three matrices with the input embedding, we obtain Queries (Q), Keys (K) and Values (V) respectively which are abstractions of the input text. The scaled dot-product attention, involves these Queries, Keys and Values (see Figure 2.1.2) and its computation is shown in Equation 2.1.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2.1)$$

Multi-head Attention

Multi-head attention is the computation of multiple attention functions in parallel, the results of each attention function are concatenated and multiplied by the W matrix (a matrix also learned during training), as depicted in Figure 2.1.2 and Equation 2.2. In the architecture shown in Figure 2.1.1, when there's a multi-head attention, eight different attention mechanism are used in parallel.

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O \quad (2.2)$$

where

$$head_i = \text{Attention}(XW_i^Q, XW_i^K, XW_i^V)$$

Where W^Q , W^K and W^V are the learned matrices. X is the input embedding vector and XW_i^Q, XW_i^K, XW_i^V are Queries (Q), Keys (K) and Values (V) respectively.

2.1.2 BERT based architectures

In this subsection, the Transformer models Bidirectional Encoder Representation from Transformers (BERT) [6], Efficiently Learning an Encoder that Classifies Token Replacements Accurately (ELECTRA) [3] and A Lite BERT (ALBERT) [21] are explained. All of these models are based on the same architecture (BERT), which is detailed first.

BERT

BERT (Bidirectional Encoder Representation from Transformer) [6] is a model that is designed to be used for different tasks, unlike the model proposed in [41]. The BERT model is a multi-layer bidirectional Transformer that stacks several encoders from [41] together. It has achieved state-of-the-art results by just adding an output layer to the encoder by pre-training on deep bidirectional representations from unlabeled text and allowing it to learn contextualized word embeddings.

BERT consists of two phases:

1. Pre-training. The model is trained on unlabeled text where it learns language and context from it. A representation of pre-training is shown in the left side of Figure 2.1.3.
2. Fine-tuning. Fine-tuning starts with the pre-trained parameters and they are updated to best fit the downstream task of choice. The model is capable of performing different tasks by feeding the specific input and output of the downstream task into the model, as shown in the right side of Figure 2.1.3.

The pre-training is sub-divided into two parts:

1. Masked Language Model (MLM). MLM task is used to train a deep bidirectional representation, 15% of the tokens are masked (hidden from the model using a mask token) and the model tries to predict the masked word with the context of it (words around the masked token). Nevertheless, since mask tokens would only appear during pre-training, this generates a discrepancy between pre-training and fine-tuning. As a result, from all masked words, 80% of them are indeed replaced with the mask token, 10% of them are replaced with a random word and another 10% of them are not changed, the original word is not masked.
2. Next Sentence Prediction (NSP). NSP task predicts whether a sentence is followed by another sentence or not. 50% of the time, the sentence is indeed the

next one and 50% of the time the sentence is a random one from the corpus. This helps the model understand context and relationships across different sentences.

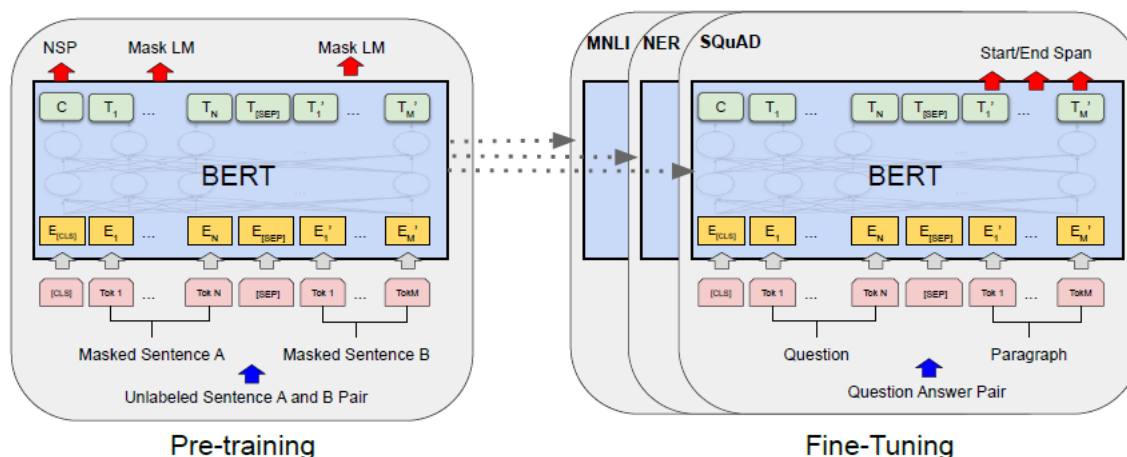


Figure 2.1.3: BERT’s pre-training and fine-tuning phases [6].

The input embeddings are a continuous representation of the input where words with similar meaning have a similar representation. The embeddings of the model are a result of:

1. WordPiece embeddings [37]. Tokenizing a text is splitting it into words or subwords. WordPiece is a subword tokenization (splitting the words into a bunch of characters), it starts with every character present in the training data and merges them to maximize the likelihood of the subword tokens in the training data.
2. Position embeddings. Since the attention mechanism does not maintain the order of words in a sentence, the position embeddings incorporate positional information into the model.
3. Segment embeddings. Indicates if the token belongs to the first or second input sentence.

ELECTRA

ELECTRA [3] has the same architecture as BERT and the same tokenizer (WordPiece [37]), the modification lies in the way each model learns. Specifically, the pre-training, where BERT seeks to learn bidirectional representations with the task of MLM, masking 15% of the tokens, making the model learn only from those mask tokens. While ELECTRA, proposes the Replaced Token Detection (RTD) task, which requires two models, a generator and a discriminator.

1. Generator. The generator performs the MLM task like BERT, some words are replaced with the mask token and the generator learns to predict the masked word as depicted in the left side of Figure 2.1.4.

2. **Discriminator.** The discriminator learns to predict if each word in the input is the original one or if it is replaced by the generator as shown in the right side of Figure 2.1.4.

According to [3], instead of just learning from a small sample, ELECTRA’s pre-training allows it to learn from all words making training faster than BERT’s and achieving better results than BERT in many downstream tasks.

Another difference between BERT and ELECTRA is that ELECTRA does not perform NSP task during the pre-training phase.

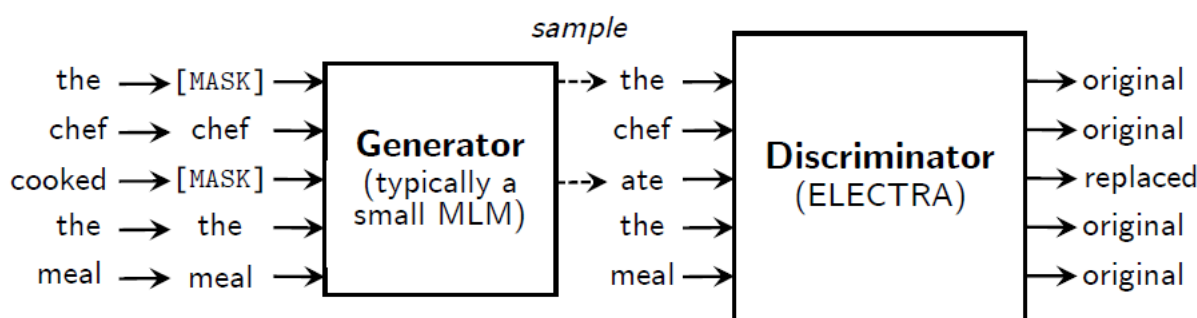


Figure 2.1.4: Replaced token detection. Electra’s pre-training task [3].

ALBERT

The ALBERT [21] model uses the same architecture as BERT although the activation function has been replaced by a Gaussian Error Linear Unit (GELU) function [14] on the FFNs. Moreover, the tokenization it uses is SentencePiece [20], a tokenizer that does not assume that words are separated by spaces and includes spaces in its set of characters.

The improvements proposed by [21] are:

1. **Hidden layer size.** The size of the hidden layers usually is the same as the size of the vocabulary embeddings. For ALBERT though, they isolated the size of the hidden layers from the size of the vocabulary embeddings by projecting one-hot vectors into a lower-dimensional embedding space and then to the hidden space. As a result, this modification made it easier to increase the hidden layer size without significantly increasing the parameter size of the vocabulary embeddings.
2. **Shared parameters.** ALBERT shares parameters for the FFNs and the attention mechanisms (see Figure 2.1.1).
3. **Sentence-Order Prediction (SOP).** Instead of using NSP, during pre-training ALBERT performs SOP, where the model learns if every pair of sentences is consecutive or not, forcing the model to learn coherence properties.

With these changes, ALBERT has fewer parameters than BERT, achieving better results on different tasks.

2.1.3 Monolingual models

The first methods proposed in NLP for Transformer models with non-English datasets were multilingual models. Afterwards, monolingual models like CamemBERT [27] (French BERT model) and FinBERT [42] (Finnish BERT model) offered a more particular solution for low-resource languages outperforming multilingual models.

KB-BERT (Swedish BERT) [26], a model trained by the National Library of Sweden (KB), is trained using the code and instructions proposed in [6]. The Swedish model outperformed the existing multilingual model Multilingual BERT (mBERT) [7] at the time of publication, but it also outperformed the previous Swedish models by using more data from KB's resources.

The dataset included resources from 1940's to 2019, they were digitized newspapers, official government publications, legal e-deposits, social media and Wikipedia Swedish articles. After KB-BERT, the National Library of Sweden also released KB-ELECTRA and KB-ALBERT, the pre-training of those models is done according to the code and instructions of the original English model, similar to KB-BERT.

2.1.4 Multilingual models

Transformer models achieve state-of-the-art results in many tasks, nevertheless, before multilingual models, every NLP model is pre-trained on English datasets, limiting their use since 80% of the population does not speak English [45]. A general solution offer to that problem, as mentioned in Section 2.1.3, is to pre-train models with multiple languages making the model capable of performing downstream tasks in many languages.

mBERT

mBERT[7] has the same architecture as BERT, the difference is that mBERT is pre-trained with the top 100 languages from Wikipedia dataset [28]. Since the size of each language in the Wikipedia dataset varies, they used techniques to sample more text from low-resource languages and less from high-resource language, learning almost equally from all of them and preventing the model from overfitting or underfitting any language.

XLM-R

XLM-RoBERTa (XLM-R) [4] is a multilingual Transformer model based on Robustly Optimized BERT Pretraining Approach (RoBERTa) [23]. RoBERTa has the same

architecture as BERT but it has different hyperparameters, it does not perform the NSP task during pre-training, only the MLM task and it is pre-trained on larger mini-batches and learning rates, that is why part of the name of the model is optimized BERT. XLM-R is trained with a filtered version of CommonCrawl data with 100 languages and it uses a subword tokenization, SentencePiece [20] with a unigram language model [19].

mT5

Multilingual Text-to-Text Transfer Transformer (mT5) [45] is the multilingual version of Text-to-Text Transfer Transformer (T5) [33], a model whose main feature is converting all NLP problems to a sequence-to-sequence format, it is fed with text as input as every other Transformer model with the difference that it also generates text as output. mT5 uses the full architecture proposed in [41], not only the encoder part of the architecture as most of the other models and is pre-trained with the MLM task. mT5 is pre-trained using the mC4 dataset, which contains 101 languages and they are sampled with a similar approach to mBERT's [7].

Sentence Transformers XLM-R

Sentence transformers were presented in [35], it is a solution that uses siamese and triplet network structures. These type of networks use the same architecture and are fed with two different input vectors to compute comparable output vectors. These networks find sentence embeddings that are easy to compare making tasks like sentence similarity faster to perform and with better results [35].

The sentence transformers version of multilingual models like XLM-R [34], because of being pre-trained similarly to the monolingual model, has the embeddings of multiple languages aligned, thanks to the siamese and triplet networks. Which makes it easier to train a model using a high-resource language and then extending its results to a lower-resource language with a similar performance.

2.1.5 Transformer models for the medical classification task

A comparison of the size of the Transformer models is shown in Table 2.1.1, where we can see listed each model with its number of parameters (obtained from the model used in code, not from the paper of each model since there is some discrepancy between them despite using the original code implementation).

2.2 Explainability

Deep learning has been responsible for great technological advances in a variety of tasks like NLP, obtaining state-of-the-art results and in some cases even improving human's performance [13, 44]. However, the benefits have come at the expense of

Table 2.1.1: NLP models and the number of parameters they use.

	Model	Parameters (M)
Swedish models	KB-BERT	124
	KB-ELECTRA	124
	KB-ALBERT	14
Multilingual models	mBERT	177
	XLM-R	278
	mT5	300

models being less interpretable [29], which lead to being untrusted [9, 12]. Therefore, a field has been growing to offer a solution to this problem, XAI. Explainability tries to help the end-user understand why the model came to a decision, increasing the trust in the model and allowing to improve the model with the explanations provided [5].

Explanations can be divided into different categories:

- Local and global explanations. Local explanation concern a specific prediction from the model, e.g., explaining why a model labeled a movie review as a positive or negative comment. On the other hand, global explanations tries to find an understanding of the model’s behavior, e.g., finding gender bias in a credit risk prediction model [5].
- Self-explaining and post-hoc methods. Self-explaining methods provide explanations inherently at the same time as the prediction, e.g., decision trees. On the contrary, post-hoc methods require additional operations since the explanations are created after the model is trained, e.g., SHAP, IG [5].

Explainability methods use different techniques to obtain such explanations, some methods use more than one technique, two of the most popular techniques are:

- Feature importance. Also called attribution importance, investigates the importance scores of different features. Attention mechanism and first-derivative saliency are two common operations for feature importance methods [5].
- Surrogate model. Predictions are explained by learning a second model, an easier model to interpret, such as decision trees or linear regression. Thus, this method is model agnostic.

In this thesis, two feature attribution methods (based on feature importance) were used, IG and SHAP. Although SHAP also uses a surrogate model technique. These methods are post-hoc since the NLP model itself can not output those type of explanations and additional operations are needed. Moreover, the type of explanation obtained by both methods is local.

2.2.1 Integrated Gradients

IG is a feature attribution method that aims to compute the importance of each input feature that contributes to the model’s predictions. This post-hoc attribution method does not modify the network to obtain explanations [39]. The model requires a baseline, which is supposed to be an input that results in a neutral prediction. The method considers a straight line path from the baseline x' to the specific input x that will be explained and computes the gradients at small steps α along the path, as depicted in Figure 2.2.1 and integrates those gradients along the line, shown in Equation 2.3 [39].

$$\text{IntegratedGrads}_i(x) ::= (x_i - x'_i) \times \int_{\alpha=0}^1 \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} d\alpha \quad (2.3)$$

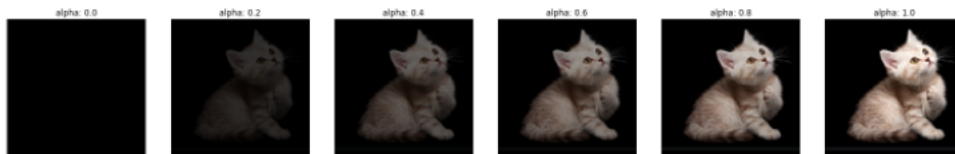


Figure 2.2.1: The images show the steps α between the baseline (leftmost image) and the original image (rightmost image) along a straight line path between them [17].

The baseline can be a black image for computer vision tasks or zero embedding vector for NLP problems. Equation 2.3 is the only attribution method that satisfies a set of desirable axioms for explanations. The desirable explanations suggested by IG are:

- Sensitivity. If a model predicts different classes for a baseline and an input that have the same features but one, that feature can not have a zero attribution value.
- Implementation invariance. The attribution does not depend on the model, if two models output the same predictions from the same inputs, they will have the same attributions because the method depends on the outputs and inputs only.
- Completeness. The sum of the attributions are equal to $F(x) - F(x')$, the difference between the prediction of the model with input x and the baseline (x').
- Linearity. If a deep network is a linear combination of two different networks, e.g., $axf_1 + bxf_2$, the attributions have to preserve the linearity within the network.

2.2.2 SHapley Additive exPlanations

Similarly to IG, SHAP, tries to explain a prediction by computing the importance of each input feature. It uses a game theoretic approach [38], i.e., the features of the data act as players and Shapley values are a distribution of the prediction among its features [25]. SHAP, like IG also uses baselines (explained in Section 2.2.1), if a feature pushes the prediction to a specific class more than the baseline, this feature gets a larger

attribution value. SHAP represents the explanations as an additive feature attribution by using a linear model as shown in Equation 2.4.

$$f(x) = g(\hat{x}) = \phi_0 + \sum_{i=1}^M \phi_i \hat{x}_i \quad (2.4)$$

where $f(x)$ is the original model with input x . $g(\hat{x})$ is the explanation model and \hat{x} is a simplified feature vector (where some of the features are set to zero, meaning that feature is absent). M is the maximum size of the features and ϕ_i is the feature attribution of feature i , the SHAP values.

SHAP satisfies the following properties:

- Local accuracy. The explanation model and the original model's outputs have to be the same.
- Missingness. An absent feature gets a zero attribution.
- Consistency. If there is a change in a model, making the value of a feature increase or decrease, the Shapley value has to change in the same way.

These properties have a unique solution, however, an approximation is needed to be computationally tractable. One of the approximation methods is Kernel SHAP. Kernel SHAP is a model agnostic method, it obtains explanations using the following algorithm:

1. Samples simplified feature vectors \hat{x} and sets some features randomly to zero, which means absent.
2. Obtains predictions of the simplified features vectors \hat{x} by feeding the simplified feature vectors to the model.
3. Computes a weight for each \hat{x} depending on how close a sample is to the original x .
4. Fits a weighted linear model to \hat{x} and the predictions made by the model.
5. Returns the SHAP values ϕ (coefficients of linear model in Equation 2.4).

2.3 Domain specific NLP

Researchers have demonstrated in different publications that NLP models are capable of performing domain-specific tasks, applications where the language used is not common and it is unlikely that the pre-training dataset used for the models contained the same vocabulary as the domain-specific data. In [1], they use the MIMIC dataset [16] a digital health record dataset for text classification with the BERT model. Moreover, [11] uses RoBERTa [23] for four domain-specific areas; biomedical, computer science publications, reviews and news. Furthermore, BioBERT [22] is a

BERT model used for the biomedical text mining. These publications are focused on showing how by further pre-training with unlabeled domain data, the NLP model performs better than using the model with its regular pre-training.

2.4 XAI for NLP

As mentioned in Section 1, building trust from users for AI techniques is crucial to many specialized fields like medicine. A survey of explainability methods for NLP models have been explored in [5]. Publications like [43], have worked to support or refute scientific arguments and provide rationales to justify the model's decision. In the medical domain, [30] presents a CNN with an attention mechanism that predicts medical codes from clinical text and also provides explanations for the model's predictions. Finally, in [15], they use different models with an attention mechanism, the models are fed with electronic medical records (one of the datasets is the MIMIC dataset) and from those records, the model predicts different outcomes with explanations from the attention mechanism, although, the rationales obtained from this mechanism are questioned and the authors claim that sometimes they are misleading [10].

2.5 Antibiotics prescription

For the task of antibiotics prescription, [36] uses random forests with laboratory and administrative data to predict urinary tract infections, based on their results, they created policies that decide if the patient should be immediately prescribed antibiotics or if they should delay the prescription. Moreover, [8] uses decision trees to select the best antibiotic to be prescribed which can effectively cure the disease avoiding bacteria resistance, the data they use consists of patient demographic, clinical history and previous antibiotic use.

Chapter 3

Methodology

In this chapter, the overall work process is explained starting with a description of the dataset, followed by the modifications done to be used for the medical text classification task. Furthermore, the way the NLP models were trained is explained and how the explainability methods were used to obtain individual and global explanations.

3.1 Dataset

The antibiotics dataset was provided by Folktandvården Västra Götaland, the largest dental care organization in Sweden. They collected data from the years 2012, 2014, 2017 and 2018, with pseudo anonymized patients information from all ages, the data collected consists of daily notes from patients medical records and recipe texts [2]. The dataset was handled with care to avoid leaking or sending it outside the Peltarion servers and any identifiable characteristic was removed from it to protect sensitive information. It contains four columns: Klinik, Personnr, Datum, Anteckning. Klinik states the name of the clinic visited, Personnr contains a unique pseudo anonymized number for each patient, Datum is the date of the visit and Anteckning is free text, it is where the medical record is, what was found by the dentist, procedures performed, information about the patient, etc. The last version of the dataset has 10,203 samples and it is almost balanced, 56% of the samples were from class 1 (antibiotics were prescribed) and 44% of the samples were from class 0 (no antibiotics were prescribed).

Patients' information is spread along many rows. To prepare the dataset for classification, the text of Anteckning of patients with the same patient number and same date were joined into one cell so that it contained all the information of the patients per visit. Moreover, the rest of the columns were removed and a "Label" column was added. The "Label" column has a value of 1 if antibiotics were prescribed and a value of 0 if they were not prescribed. For mT5, since it is a sequence-to-sequence model, the labels are instead "Positive" and "Negative".

The labels used are the decisions that dentists took with the medical record of each

patient, nevertheless, not all of those decisions were correct, there were mistakes when prescribing antibiotics. In [2], the antibiotics prescriptions of Folkhälsan Västmanland were evaluated in three occasions; during the month of September of years 2012, 2015 and 2018 and the results obtained are displayed in Table 3.1.1. There is no information about years 2014 and 2017. However they state that antibiotics prescription was reduced after 2012 and correct prescriptions increased as a consequence of conducting dental training on antibiotics treatment.

Table 3.1.1: Evaluation of antibiotics prescription in the dataset of this thesis. [2]

	2012 Share(%)	2015 Share(%)	2018 Share(%)
Correct	87.6	85.1	89.3
Inaccurate	12.4	14.9	10.7

The antibiotics dataset was not ready for ML tasks, it included different automatically generated messages, html tags, Not a Number (NaN) cells and data leakage was occurring in it. For example, the antibiotics prescribed were written in the medical records and it was easy for the models to "take a shortcut" and focus on them instead of focusing on the patient's condition. Thus, some modifications had to be done to be able to use the data for classification and train an NLP model to predict if antibiotics should be prescribed or not from the patients' medical records.

Data cleaning

The process of cleaning the data was iterative as depicted in Figure 3.1.1. First, the NLP models were trained with the dataset without any modifications. Then, explanations were obtained using IG and SHAP from individual samples and depending on those results (what the models were focusing on), if the models were not focusing on the relevant information (medical condition of the patient), the dataset was modified.

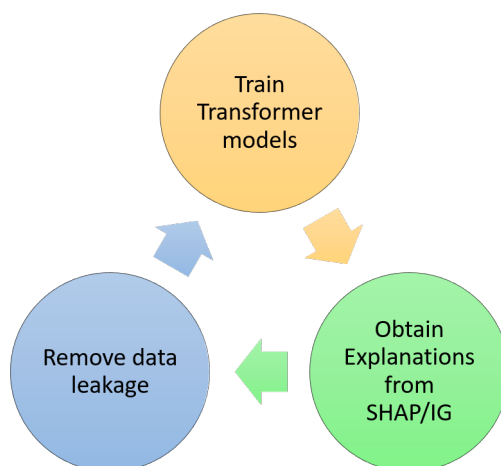


Figure 3.1.1: Data cleaning process.

The initial significant modifications to the dataset were:

- **Html tags removal.** Html tags were not important features that the models were focusing on to make predictions. Although the tags significantly increased the size of the input, as depicted in Figure 3.1.2 and given that the NLP models have a size limit of 512 tokens, the input was truncated, probably leaving out some important information.
- **Electronic recipe.** These are some of the automatically generated messages included in the medical record of the dataset and contained the antibiotics prescribed and the doses of it. The models were focusing on the prescribed medicine to make a prediction, so this was a source of data leakage. After removing these messages, the accuracy of the NLP models decreased by approximately 6%, these type of messages were found more than 5000 times inside the data. An example of this message is shown in Figure 3.1.4.

```

Us akut Pat söker för värk hö uk. Kraftig värk även på natten sedan 2 dagar. 46 omfattande kavitet, otätt
tfb. Mkt perk-, o kron öm. Ingen synlig svullnad 1 ap rtg: 46 ap radiolucence dist roten, bifynd 47 karies
do. Diagnos symptomgivande ap parodontit Rek ex 46, pat accept Beh <SPAN style='FONT-FAMILY:
"Arial",sans-serif'>Inj 3 x 1,7ml Septocaine Forte Artikain 40 mg/ml + adrenalin 10 mikrog/ ml.</SPAN>
<SPAN style='FONT-FAMILY: "Arial",sans-serif'>Seperation 2 rötter med Zekrya</SPAN> <SPAN
style='FONT-FAMILY: "Arial",sans-serif'>46 extraktion, mes roten uthävläs lätt, dist roten kräver marg
friläggning luxering m luxator, <SPAN style="mso-spacerun: yes"> </SPAN>tångextraheras</SPAN>
<SPAN style='FONT-FAMILY: "Arial",sans-serif'>e recept lbumetin 600 mg dosering enl nedan.</SPAN>
<SPAN style='FONT-FAMILY: "Arial",sans-serif'> <SPAN style="mso-tab-count: 1"> </SPAN></SPAN>
<SPAN style='FONT-SIZE: 12pt; FONT-FAMILY: "Arial",sans-serif; mso-fareast-font-family: "Times New
Roman"; mso-ansi-language: SV; mso-fareast-language: SV; mso-bidi-language: AR-SA'>Post ex info
om smärta och blödning</SPAN>

```

Figure 3.1.2: Examples of the Html tags inside patients' medical records are shown.

Usually a drop in accuracy is undesirable, however, in this case, since the electronic recipe was a source of data leakage, it means that the problem was removed and the underlying task (predicting if antibiotics should be prescribed from the medical condition of the patient) becomes harder.

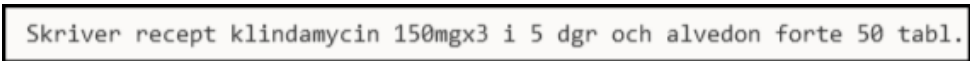
Replacing words

A set of words were occluded from the NLP models because, with the help of the explanations obtained it was possible to see that they were other sources of data leakage. These words were from different categories; antibiotics, years and names, all of them were replaced by PAD tokens (a token used in NLP to fill small text inputs so that every sample in the dataset has the same size). By being replaced by PAD tokens, the models no longer learned them during training so they could focus on the patients' conditions.

- **Antibiotics.** The antibiotics prescribed not only appeared in the Electronic recipe messages but sometimes inside the medical notes. Removing the "antibiotics" resulted in a decrease in accuracy of approximately 3% for every NLP model.

- Years. The antibiotics class (antibiotics were prescribed), consists of data collected in years 2012, 2014, 2017 and 2018 whereas the no antibiotics class (antibiotics were not prescribed), consists of data collected from the year 2018. The explainability methods showed that the models were focusing on the year that appeared in the medical notes to make predictions. Since there is a correlation between the years and the class of the dataset, years were also replaced by PAD tokens. After replacing the years, the models had approximately a 2% decrease in accuracy. Again, this decrease in this case is not negative, since it means that the source of data leakage has been removed.
- Names. Inside the dataset, some names were found, by manually inspecting the data. After discussing this with the dataset owners, we concluded that those names were dentist's names, usually specialists and when the names of the doctors were in the text, most of the time antibiotics were prescribed. The accuracy of the models decreased around 2% without names.

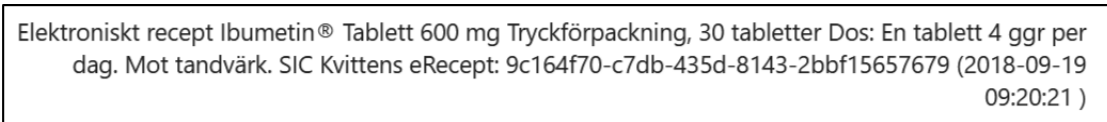
The dataset owners provided a list of antibiotics to remove them from the data. Despite having the list, it was challenging to get rid of all the antibiotics, since some medications were misspelled. In Figure 3.1.3, the antibiotic "klindamycin" is written, despite also being written sometimes as "clindamycin". Furthermore, some antibiotics were abbreviated, for example, penicillin was written as pc sometimes in the dataset, making the task of removing them more complex.



Skriver recept klindamycin 150mgx3 i 5 dgr och alvedon forte 50 tabl.

Figure 3.1.3: Antibiotics prescribed inside the medical records.

The years inside the data were detected by searching for number patterns similar to dates, e.g., 2012-08-11, as in the bottom right of Figure 3.1.4. Nonetheless, sometimes dates included only the month and the year. Moreover, the years consisted of two digits and other times they consisted of four digits so different patterns had to be used to replace all years from the dataset.



Elektroniskt recept lbumetin® Tablett 600 mg Tryckförpackning, 30 tabletter Dos: En tablett 4 ggr per dag. Mot tandvärk. SIC Kvittens eReceipt: 9c164f70-c7db-435d-8143-2bbf15657679 (2018-09-19 09:20:21)

Figure 3.1.4: Automatically generated message: Electronic recipes.

Finally, detecting names was the most difficult task, to do it, a BERT Name Entity Recognition (NER) model [26] was used, which is a pre-trained version of BERT that recognizes entities in the Swedish language such as persons (e.g., Engelbert), objects (e.g., Volvo), time (e.g., today), location (e.g., Djurgården), etc. The model was trained for NER using the SUC dataset [40]. The output of the model is the word detected with the category predicted and a score. The BERT-NER model detected many words as names and not all of them were actual names, however, by setting a threshold

of .95, most of the names were found and the number of false-positive names was reduced.

Visits

In the dataset, 23.55% of the patients that were prescribed antibiotics visited the dentist just once, as depicted in Figure 3.1.5. Out of all the patients that were prescribed antibiotics, 15.58% had their second visit within the next 14 days after the first visit, shown in Figure 3.1.6. After talking to the domain experts (owners of the dataset), since a patient that was prescribed antibiotics seldom comes back a few days later, we concluded that the reason for this is that sometimes if a patient goes to the doctor his or her medical notes start that day but if the dentist did some examination or procedure, comments could be added in the coming days to the dataset, making it look as if the patient had two visits when it was only one. To address this, the visits for each patient were joined when their second visit was within the next 14 days to count them as only one visit. By joining the visits within 14 days, the accuracy of the models improved around 3%.

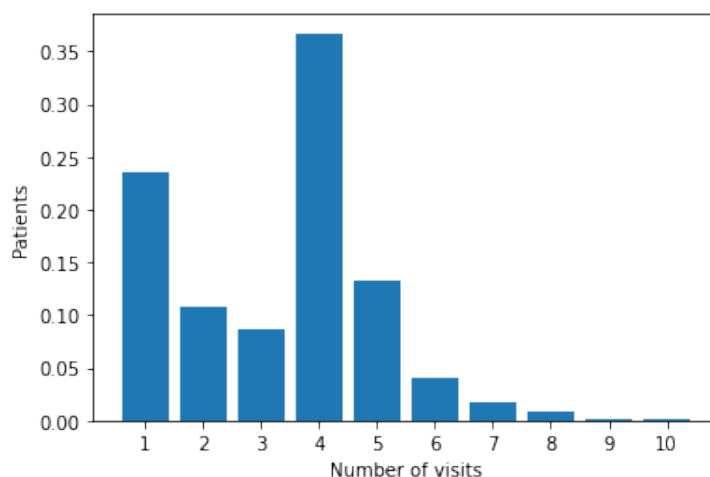


Figure 3.1.5: Number of visits of patients that were prescribed antibiotics.

Other changes that modified the accuracy of the NLP models were: using only lower cases for the medical notes and removing other automatically generated messages. Although, these changes did not impact the model performance significantly as previously mentioned.

3.2 Evaluation of Transformer models

The models were trained with the patients' dental records to predict if antibiotics should be prescribed or not. First, the input text was tokenized, an example of tokenization is:

In tokenization, text is splitted.

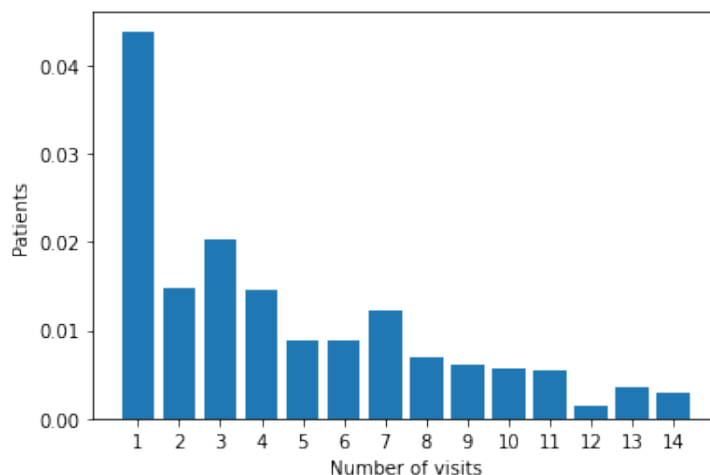


Figure 3.1.6: Number of days between first and second visits out of the patients that were prescribed antibiotics.

in token #ization , text is split #ted .

Then, the tokens are fed into the models and are transformed to word embeddings and fine-tuned to perform a binary classification task, predict whether or not to prescribe antibiotics. From the results, monolingual and multilingual models were evaluated.

The monolingual models selected were the models trained with data from the National Library of Sweden (KB), obtaining better performance than previous Swedish models with a larger amount of text [26]. The multilingual models were chosen because they obtained state-of-the-art results in multiple tasks. The smaller version of XLM-R and mT5 were used due to computational resources. Two different models of XLM-R were used, the original and the model with sentence transformers presented in [34]. The sentence transformers XLM-R was used since the embeddings of all languages are aligned and we thought it could obtain better results than the common version of the model.

The models were fine-tuned starting with the parameters obtained from their respective pre-training, moreover, the original word embeddings for each model were used. During training, each NLP model was trained with 80% of the dataset and the rest was used for validation.

For every model, some coarse hyperparameter search was performed to find a good learning rate by choosing the model with the highest accuracy in the validation set. Four different learning rates were tested starting with the learning rate proposed in each of the models' publications and then testing smaller and larger values.

The optimizer used was Adam and the metrics used to compare models were accuracy and F1. To define the F1 score, two other metrics have to be explained first, precision and then recall. Precision is a measure of how many positives were correct out of all

the predicted, i.e.,

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad (3.1)$$

On the other hand, recall measures how many positives were predicted out of all the actual positives.

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \quad (3.2)$$

F1 score then, is a metric often used to take into account missclassifications even when classes are not balanced.

$$\text{F1} = 2 \times \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3.3)$$

3.3 Explanations

IG and SHAP were used to obtain explanations to understand the Transformer models predictions and see what the models were focusing on. Additionally, a comparison between the individual explanations of both explainability methods was done and finally, the individual explanations were used to obtain global explanations from each NLP model to have a sense of what words were contributing the most to each class in the dataset.

- Individual explanations. These explanations consisted of only one patient's visit, the explanations showed the words that contributed most to predicting each class in one sample.
- Global explanations. The explanations consisted of multiple patients' visits, the explanations showed which words had a higher weight when predicting each of the classes in all samples.

The only model that was not used to obtain explanations is mT5, because its nature (sequence to sequence model) complicates the computation of the explanations. However, other work such as the model Why T5? (WT5) [31] is capable of explaining the predictions of T5 but those generative explanations are out of the scope of this thesis.

3.3.1 Local explanations

To obtain explanations, the input text was tokenized with a maximum length of 512 tokens (the input was truncated if it required more than 512 tokens), however, if the inputs were smaller, they were not filled with PAD tokens to fit the maximum length.

The baselines, as explained in Section 2.2.1, are inputs that result in a neutral prediction. The baselines x' were different for each sample because their size has to be the same as the input sample. For both explainability methods, x' was the same, starting with a CLS token, ending with a SEP token and in the middle, having PAD tokens to fit the size of the input sample, as suggested by [18]. Following the example of Section 3.2, the format of the the baseline x' thus becomes:

In tokenization, text is splitted.

in token #ization , text is split ##ted .

$$x' = [\text{CLS}] [\text{PAD}] [\text{PAD}] [\text{PAD}] [\text{PAD}] [\text{PAD}] [\text{PAD}] [\text{PAD}] [\text{PAD}] [\text{PAD}] [\text{PAD}] [\text{SEP}]$$

As observed, the number of PAD tokens are the same as the number of elements after text is tokenized (splitted).

Local explanations of IG are obtained from computing the gradients of the output with respect to the input, for NLP problems, IG uses the embeddings space, where inputs have a continuous representation. The IG method is illustrated in Figure 3.3.1, the steps for this method are:

1. The input text is tokenized (the text is splitted according to the tokenization method used).
2. The tokens are mapped into the embeddings space (word representation that allows words with similar meaning to have a similar representation).
3. With these word embeddings x and the baseline x' , the path between them is computed as shown in Equation 3.4 where α are the number of steps taken.
4. The Transformer models make predictions with Equation 3.4 as their input.
5. Gradients from the output of the model are obtained from each step along the path.
6. All the gradients are integrated along the path between input and baseline.
7. With the integrated gradients the attribution values are obtained in the embeddings space.
8. Attributions in the token space are computed by performing a summation along the embeddings dimension.

$$x' + \alpha \times (x - x') \tag{3.4}$$

On the other hand, SHAP method is illustrated in Figure 3.3.2, the steps it takes

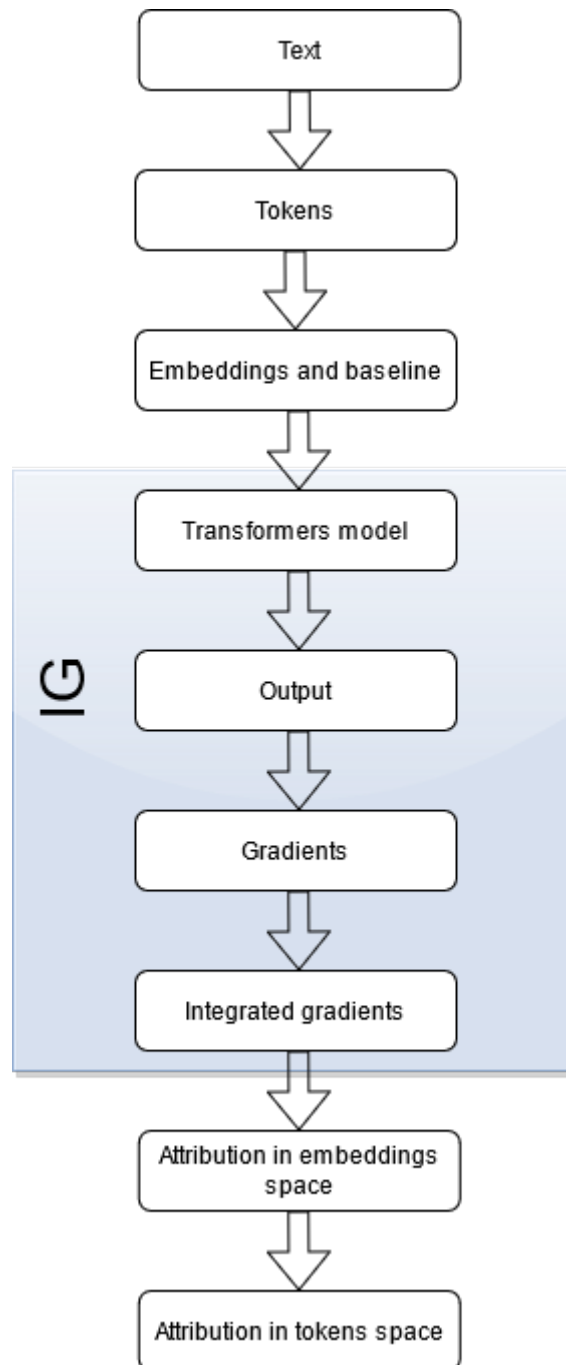


Figure 3.3.1: Diagram of IG explanations.

are:

1. The text is tokenized like in IG.
2. The tokens are perturbed (some parts of the tokens are set to absent randomly) those perturbed tokens are the simplified feature vectors \hat{x} .
3. The \hat{x} are the input of the Transformer models.
4. Using the resulting output of the model and the simplified feature vectors, an explainable model is trained to fit them.

5. Finally, the attributions are the coefficients of the weighted linear model used. As opposed to IG, SHAP works with the tokens while IG needs the embeddings.

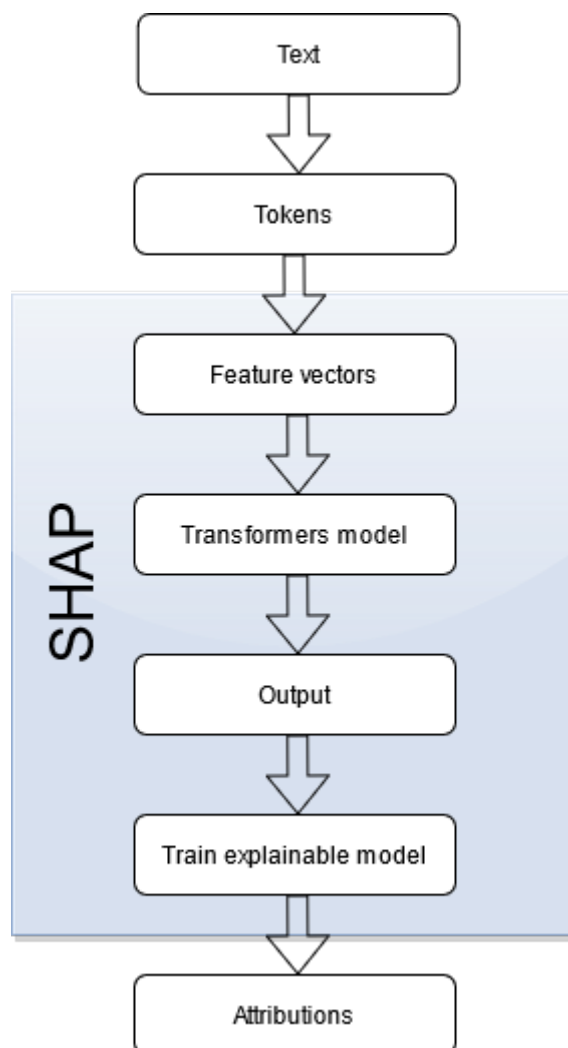


Figure 3.3.2: Diagram of SHAP explanations.

The number of steps α along the straight line between the input x and the baseline x' , from Equation 2.3 was chosen to 500 for IG. The number of samples of feature vectors \hat{x} to calculate their feature importance with SHAP was 500 as well.

The explainability methods assign a weight to each feature, in this case, the features most of the time are not words, they are parts of words (tokens). To improve the clarity of explanations, the tokens were joined to form the original words and the weights of each of their tokens were added. As a result, each word had a weight which is easier to interpret than having weights assigned to each token. For Transformer models using SentencePiece embedding, as explained in Section 2.1.2, the same process was also done but slightly different since SentencePiece embedding considers spaces between words a character too.

Local explanations were used to compare both explainability methods and analyze their results. Furthermore, they were also used to understand better what each model

is focusing on to make individual predictions.

3.3.2 Global explanations

To obtain global explanations, local explanations of every sample in the validation set were obtained. Thus, these were samples that were not seen before by the Transformer models. With the attribution values of each input sample, a dictionary was created, the keys were all the words inside the validation set and the keys are the attribution values aggregated. A toy example taken from Captum’s tutorials [18] of individual explanations from movie reviews is shown in Table 3.3.1 and the dictionary of this local explanations is shown in Table 3.3.2.

Table 3.3.1: Example of attributions of individual explanations using movie reviews.

it	was	the	best	film	ever
0.12	-0.05	0.15	0.25	0.04	0.04
it	was	a	fantastic	film	!
0.15	0.01	0.29	0.33	-0.20	-0.02

Table 3.3.2: Example of global explanations using movie reviews.

it	0.27
was	-0.04
a	0.29
the	0.15
best	0.25
fantastic	0.33
film	-0.16
ever	0.04
!	-0.02

For the value aggregation, two ways of computing the global attributions values were tried:

- Absolute values of each word, an example is shown in Table 3.3.2.
- Normalized values for each word (normalized by all the attribution values of the sample they belong to).

The absolute values of each word were used, since normalized values reduces the attribution values of each feature and the most frequent words become the top words without taking into account their importance.

With this procedure, the top words contributing to each class (from both explainability methods) were obtained. The global explanations were used to see what are the most important words in a set of samples. The top words found for prescribing antibiotics were compared to the correct guidelines of prescription presented in [2].

Chapter 4

Results

In this Chapter, the results obtained in the medical classification task from the Transformer models are presented and analyzed. Additionally, local explanations are used to compare the results of both explainability methods, the explanations of the Transformer models and to compare the explanations of cleaned dataset against an uncleaned version of it. Moreover, the correct criteria for prescribing antibiotics is presented and the top words obtained from both models are compared between them and against the correct guidelines. Finally, the global words obtained from different Transformer models are evaluated and the global explanations obtained with different versions of the dataset are analyzed.

4.1 Classification task

The classification task consisted of a binary classification where the model had to decide if antibiotics should be prescribed or not from individual visits of patients' medical records. The accuracy and F1 score results of the medical classification task are shown in Table 4.1.1. These metrics were calculated by training and evaluating the models multiple times (10) with random partition of the data, the standard deviation is included in the results.

Table 4.1.1: Results of the NLP models.

Model	Accuracy (%)	F1 (%)
KB-BERT	91.35 \pm 0.27	92.05 \pm 0.19
KB-ELECTRA	90.30 \pm 1.54	90.82 \pm 1.17
KB-ALBERT	88.09 \pm 0.68	88.49 \pm 1.10
mBERT	90.72 \pm 0.43	91.35 \pm 0.79
XLM-R	90.52 \pm 0.57	91.61 \pm 0.51
st-XLM-R	89.89 \pm 0.47	91.08 \pm 0.48
mT5	88.75 \pm 0.80	89.93 \pm 0.70

As observed in Table 4.1.1, the best model was KB-BERT, it had the highest accuracy

and F1 score. Although KB-ELECTRA has the same number of parameters (Table 2.1.1) and it is supposed to have a better pre-training algorithm that allows ELECTRA to improve BERT's results in other downstream tasks [3], this did not show any improved performance on our medical classification task.

The best multilingual models were mBERT and the original version of XLM-R. mBERT has higher accuracy, meaning that it made more correct predictions than XLM-R, even though, XLM-R has a better F1 score, so it accounts better the slight imbalance of the dataset.

The worst model was KB-ALBERT, however, the results were pretty close to the rest of the models considering that the model has significantly fewer parameters (Table 2.1.1). Furthermore, the worst multilingual model was mT5, a possible explanation is that the best performance from mT5 is obtained with the largest version of the model and for this work, the smallest version was used.

The original version of XLM-R had better results than the sentence Transformers XLM-R. However, we hypothesise that this model capabilities could be better exploited by further pre-training with English domain-specific data and then fine-tuning to perform the medical classification task making use of its aligned language embeddings.

In general, the monolingual models had a better performance than the multilingual models. A direct comparison could be done between KB-BERT and mBERT since they have similar number of parameters and same architectures and pre-training procedures. Their results are very close, the difference could be attributed to the extra Swedish data KB-BERT had for pre-training.

Overall the results were very accurate, especially taking into account that the dataset is not perfect and it is estimated that around 10-15% of the data was misclassified [2] which is almost the same error the Transformer models had. Moreover, there is still room for improving since most publications from domain-specific fields [11, 22] like the medical [1], suggest that great gains in performance are obtained when further pre-training the models.

4.2 Explanations

Random samples were obtained to compare explanations between Transformer models and to compare the explainability methods. Moreover, some explanations, local and global are shown to demonstrate the differences between using a model trained on a dataset without removing names, years and antibiotics as described in Section 3.1.

4.2.1 Local explanations

To visualize each method's explanation, Captum's text visualization functionality [18] was used. For this thesis report, the visualization highlights in green the words that positively contributed towards prescribing antibiotics, the more intense the color is, the more important that feature was for the classification. On the contrary, the words highlighted in red are the features that were contributing against the prescribing antibiotics class (negative attribution values).

Comparison between IG and SHAP

Figures 4.2.1 and 4.2.2 are explaining the same sample but in Figure 4.2.1 SHAP was used and in Figure 4.2.2 IG was used. As observed, both explanations have many words in common but the biggest difference is that SHAP highlights words like "akut" (acute), "svullnad" (swelling), "pcv" (Phenoxymethylpenicillin) and "infektion" with larger attribution values (the words have a more intense color). On the contrary, the explanation from IG, highlights multiple words but most of them have a similar attribution value, except only for "25".

Furthermore, IG highlights "ej allmänpåverkan" ("not general impact") as reason for not prescribing antibiotics which, is in alignment with the guidelines in [2] for not prescribing antibiotics.

The code for SHAP uses an L_1 regularization, as a result, the attribution values are sparser since it sets some of the variables to zero resulting in larger and usually fewer non-zero attribution values. On the other hand, IG does not use regularization which is more mathematically correct but the interpretation of the explanations is harder since many words have similar attribution values.

"PCV" should no longer be inside the dataset since it is an antibiotic but this demonstrates that there are still some medicines inside the dataset. Nevertheless, since there are very few antibiotics remaining, the model does not consider them as important as before.

[CLS] status ändrad till akut akut - 2 apikalbilder pat är tandvårdsrädd . söker för tandvärk , svullnad vä sida ök . ej feber / allmänpåverkan . pågått sedan igår , pat har tagit värktabletter mot besvären . klin : svullnad noteras extraoralt upp vä sida upp mot zygomaticus , rtg visar 24rr med apikal destruktion . 25 och 27 är apikalt ud . förskriver [PAD] pcv pga infektion med spridningsrisk (svullnad upp emot ögat) , går ej att erhålla dränage i omslagsvecket . [PAD] 2 tabletter 3 ggr / dag i 7 dgr . pat åter för ex 24rr i början av nästa v . info att höra av sig tidigare vid tilltagande svullnad , feber / allmänpåverkan . undersökning - kompletterande undersökning eller utredning , enstaka tand - (fk tillstånd 1001) : utförd [SEP]

Figure 4.2.1: Local explanation of KB-BERT using SHAP.

[CLS] status ändrad till akut akut - 2 apikalbilder pat är tandvårdsrädd . söker för tandvärk , svullnad vä sida ök . ej feber / allmänpåverkan . pågått sedan igår , pat har tagit värktabletter mot besvären . klin : svullnad noteras extraoralt upp vä sida upp mot zygomaticus , rtg visar 24rr med apikal destruktion . 25 och 27 är apikalt ud . förskriver [PAD] pcv pga infektion med spridningsrisk (svullnad upp emot ögat) , går ej att erhålla dränage i omslagsvecket . [PAD] 2 tabletter 3 ggr / dag i 7 dgr . pat åter för ex 24rr i början av nästa v . info att höra av sig tidigare vid tilltagande svullnad , feber / allmänpåverkan . undersökning - kompletterande undersökning eller utredning , enstaka tand - (fk tillstånd 1001) : utförd [SEP]

Figure 4.2.2: Local explanation of KB-BERT using IG.

Comparison between Transformer models

The comparison between explanation of different models using SHAP is shown in Figure 4.2.3. As observed there are similar words highlighted in the three models. First, KB-BERT (Figure 4.2.3, Subfigure a) has a large attribution value for words like "akut" (acute), "symtomfri" (symptom free), "symtom" (symptom), "svullnad" (swelling) and "utförd" (performed). Which intuitively, just "symtomfri" (symptom free) would be a mistake, however, since this method is using a WordPiece embedding, it is likely that it learned that the subword "symtom" (symptom) contributes toward prescribing antibiotics despite having a postfix in this input text.

Secondly, according to SHAP, mBERT is focusing on words like "akut" (acute), "trauma" (trauma), "svullnad" (swelling) and the numbers "41" and "1001" which

could be medical codes correlated with prescribing antibiotics (Figure 4.2.3, Subfigure b).

Lastly, for the XLM-R model (Figure 4.2.3, Subfigure c), a model that has a different embedding (SentencePiece), gives high importance to words like "efter" (after), "symtomfri" (symptom free), "åker" (goes), "smärta" (pain), "svullnad" (swelling) and "enstaka" (single). Most words make sense probably except for "efter" (after) and "enstaka" (single). However, the difference might be due to the embedding used.

These local explanations show what each of the top three models are focusing on to make predictions. The explanations have some differences but overall the models focus on similar words. In most cases evaluated, all the Transformer models have words in common to make predictions of individual samples.

[CLS] akut us , pat 41 är mobil efter att pat fastnat med sportdryckaflaska och det knäckte till . i övr helt symtomfri . pat hade trauma ukfront för många år sedan 41 , och 31 rf . klin synes 41 mob gr 2 - 3 . ej perköm . 2 aprtg - uppvisar ap ostit samt ser ut som tvärfraktur på 41 . info tanden ej går att rädda utan måste x : as . pat vill avv då han åker till japan nästa vecka . vi besl pat får recept på [PAD] endast att ta om symtom uååstår - smärta , svullnad . pat har tid för us i okt - forts terapidisk då . undersökning - kompletterande undersökning eller utredning , enstaka tand - (fk tillstånd 1001) : utförd [SEP]

(a) Explanation using KB-BERT.

[CLS] akut us , pat 41 är mobil efter att pat fastnat med sportdryckaflaska och det knäckte till . i övr helt symtomfri . pat hade trauma ukfront för många år sedan 41 , och 31 rf . klin synes 41 mob gr 2 - 3 . ej perköm . 2 aprtg - uppvisar ap ostit samt ser ut som tvärfraktur på 41 . info tanden ej går att rädda utan måste x : as . pat vill avv då han åker till japan nästa vecka . vi besl pat får recept på [PAD] endast att ta om symtom uååstår - smärta , svullnad . pat har tid för us i okt - forts terapidisk då . undersökning - kompletterande undersökning eller utredning , enstaka tand - (fk tillstånd 1001) : utförd [SEP]

(b) Explanation using mBERT.

akut us , pat 41 är mobil efter att pat fastnat med sportdryckaflaska och det knäckte till.i övr helt symtomfri. pat hade trauma ukfront för många år sedan 41,och 31 rf. klin synes 41 mob gr 2-3. ej perköm. 2 aprtg- uppvisar ap ostit samt ser ut som tvärfraktur på 41. info tanden ej går att rädda utan måste x:as. pat vill avv då han åker till japan nästa vecka. vi besl pat får recept på [PAD] endast att ta om symtom uååstår- smärta,svullnad. pat har tid för us i okt- forts terapidisk då. undersökning - kompletterande undersökning eller utredning, enstaka tand - (fk tillstånd 1001) : utförd

(c) Explanation using XLM-R.

Figure 4.2.3: Comparison of the local explanations of different Transformer models using SHAP.

Explanations for data cleaning

In this sections, we compare explanations of a Transformer model trained on the dataset with names, years and antibiotics still in it against a model trained on the cleaned data. The explanations are obtained using IG and SHAP.

In the explanations obtained with IG the models were heavily relying in antibiotics and years to make predictions. In Figure 4.2.4, the word with the largest attribution value is "pc" (penicillin), as mentioned in Section 3.1, this is not desirable. In addition, "2017" is also highlighted in this explanation.

[CLS] pat söker akut pga av värk samt känner sig svullen vä ök . läser daganteckning 2017 - 08 - 29 - då har man ställt diagnos : apikal parodontit 26 . pat ringt förra tandläkaren för att få journalkopia , men nekats . vi lovar att ringa förra tandläkaren , maria på friskvåderstorget , och be om att få hit journalanteckningar . perkusjonstest : 24 - 27 : 25 , 26 ömma . i omslagsveckat i regio 26 palperas en lokal svullnad . tar 2 ytterligare rtg regio 26 : ingen tydlig apikal radiolucens - ses tydligare på tidigare rtg ? ! skriver ut pc mot värk och svullnad . informerar om fortsatt behandling - revidering rf alt ex - info kostnad . pat har bokad tid 28 / 9 hos ordinarie tandläkare för lagning av 23 . då får man ta ställning till fortsatt behandling . undersökning - kompletterande undersökning eller utredning , enstaka tand - (fk tillstånd 1001) : utförd [SEP]

Figure 4.2.4: Local explanation of KB-BERT and IG using a dataset with antibiotics, years and names in it.

After padding the antibiotics, years and names, the explanation shown in Figure 4.2.5 was obtained. As observed, the model shifts its focus to words like "svullen" (swollen) or "svullnad" (swelling) which is something the model should focus on since those words describe the condition of the patient.

Figure 4.2.6, shows an explanation of KB-BERT using SHAP with the dataset that still has names, years and antibiotics in it. As observed, the word with the most intense color is "kåvepenin" (antibiotic) so it is the word with largest attribution value of the input text. Since the model should focus on the patients' condition to predict a class, antibiotics had to be removed from the dataset.

[CLS] pat söker akut pga av värk samt känner sig svullen vä ök . läser daganteckning [year] 08 - 29 - då har man ställt diagnos : apikal parodontit 26 . pat ringt förra tandläkaren för att få journalkopia , men nekats . vi lovar att ringa förra tandläkaren , [PAD] på friskvåderstorget , och be om att få hit journalanteckningar . perkussionstest : 24 - 27 : 25 , 26 ömma . i omslagsvecklet i regio 26 palperas en lokal svullnad . tar 2 ytterligare rtg regio 26 : ingen tydlig apikal radiolucens - ses tydligare på tidigare rtg ? ! skriver ut [PAD] mot värk och svullnad . informerar om fortsatt behandling - revidering rf alt ex - [PAD] kostnad . pat har bokad tid 28 / 9 hos ordinarie tandläkare för lagning av 23 . då får man ta ställning till fortsatt behandling . undersökning - kompletterande undersökning eller utredning , enstaka tand - (fk tillstånd 1001) : utförd [SEP]

Figure 4.2.5: Local explanation of KB-BERT and IG using a dataset without antibiotics, years and names in it.

[CLS] akut pat har besvär regio 48 . tandköttet regio 48 kräftig inflammerad . 48 under eruption . pat äter värktabletter . 48 perikoronit 47d 6mm ficka . 47 ngt perköm . pat febrig och allmänpåverkad . pat har gap svårighet . går inte ta rtg bilder . regio 48 spolning med natiumklorid 9mg / ml . ocklutions ktr . ingen påbittning på tandköttet . rek kåvepenin mot tandinfektion . skickar e - recep . pat ska höra av sig vid försämring . undersökning - kompletterande undersökning eller utredning , enstaka tand - (fk tillstånd 1001) : utförd tandvård - sjukdomsbehandlande åtgärder mindre omfattning - (fk tillstånd 3045) : utförd [SEP]

Figure 4.2.6: Local explanation of KB-BERT and SHAP using a dataset with antibiotics, years and names in it.

4.2.2 Global explanations

In this subsection, the criteria for prescribing antibiotics is introduced. Additionally, the top words found with the explainability methods are shown. In the graphs, the top words are displayed in the y-axis; the x-axis, represents the summation of the absolute attribution values of each word. When the bars are in green, the words contributed towards prescribing antibiotics and when the bars are in red, the words contributed to the opposite class (not prescribing antibiotics), same convention used for local explanations.

The correct criteria for prescribing antibiotics is shown in Table 4.2.1, on the other hand, the incorrect criteria is shown in Table 4.2.2.

Table 4.2.1: Description of criteria for correct indications when prescribing antibiotics [2].

Subgroups	Criteria
Affected general condition	General impact with feeling sick Fever
Increased risk of spreading	Trismus Extensive swelling submandibular up towards the eye or backwards in the pharynx Swollen, sore local lymph nodes
Other	Acute Necrotizing Ulcerative Gingivitis (ANUG) Prescription after specialist consultation Extensive trauma Dental sinusitis

Table 4.2.2: Description of criteria for incorrect indications when prescribing antibiotics [2].

Subgroups	Criteria
Local infection	Limited swelling / abscess, for example in connection with periapical periodontitis
In connection with endodontitis	Where endodontic causal therapy has been initiated, but that it has been supplemented with antibiotics
Pain	Pulpit Pain in postoperative inflammation (eg after extraction, alveolitis) Apical periodontitis where the primary cause of antibiotic treatment has been described as pain
In connection with extraction	Antibiotic treatment after complicated extraction or extraction of several teeth
Other	Prescription in connection with non-specific mucosal infection / inflammation In connection with anesthesia

Top words from SHAP and IG

Figure 4.2.7 shows the global explanations obtained by SHAP using KB-BERT, comparing the results obtained with Table 4.2.1. Words like "svullnad" (swelling) and "svullen" (swollen), depending on the dentists' criteria (if the swelling is considered limited or extensive), could fall in the subgroup of "Increased risk of spreading"; "akut" (acute) and "värk" (pain) would also require the dentist judgement (acute pain by infection or from postoperative inflammation) but they could belong to the subgroup

”Other” and finally, ”feber” (fever), is part of the ”Affected general condition” subgroup. Regardless of the dentists’ judgement, these words describe the medical condition of the patient, they are features that should be considered to make a prediction.

These results are very interesting since it appears that the model is focusing on some of the right words to predict that antibiotics should be prescribed. Nonetheless, words like utförandedatum (execution date) and utförd (performed) suggest that there are still some automatically generated messages that could have a strong correlation with the prediction of the model which is something undesirable since it would be another source of data leakage remaining in the dataset. Most of the rest of the words would require context to properly analyze their large aggregated attribution value but it is very difficult to analyze every sample they appeared on.

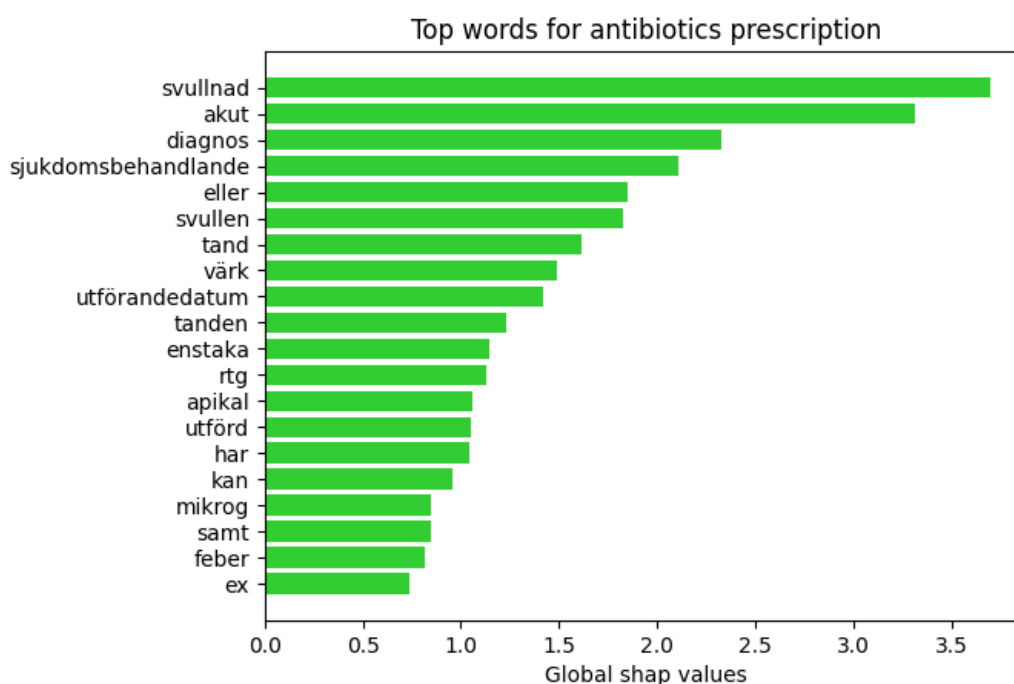


Figure 4.2.7: Global explanations for prescribing antibiotics using SHAP and KB-BERT.

In Figure 4.2.8, the top words for not prescribing antibiotics with SHAP and KB-BERT are shown, the aggregated attribution values are negative, not positive as shown in the graph, nevertheless, the absolute value is displayed to compare them with the words obtained for the antibiotics class. These words show that most of the patients that were not prescribed antibiotics went to the dentist for hygiene reasons; words like ”karies” (caries), ”fluorbehandling” (fluoride treatment), ”fluor” (fluorine), ”tandhygienist” (dental hygienist) and ”tandkräm” (toothpaste) belong to that category.

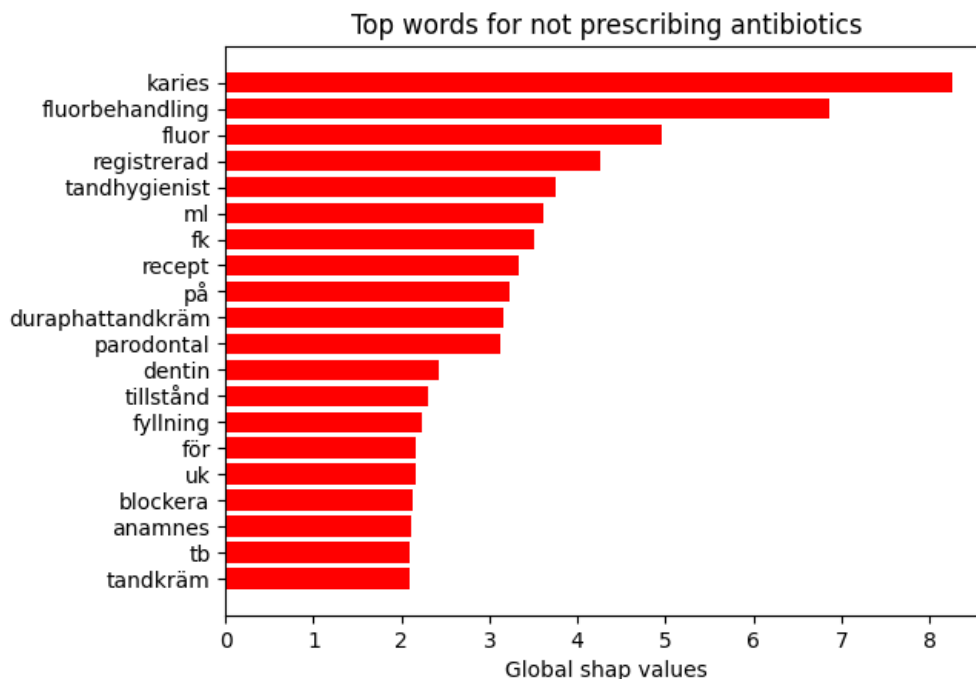


Figure 4.2.8: Global explanations for not prescribing antibiotics using SHAP and KB-BERT.

In Figure 4.2.9, the top words for prescribing antibiotics found with IG and KB-BERT are shown. Some of the top words are similar to the words found by SHAP. A list of the common top words found by both explainability methods for prescribing antibiotics are shown in Table 4.2.3.

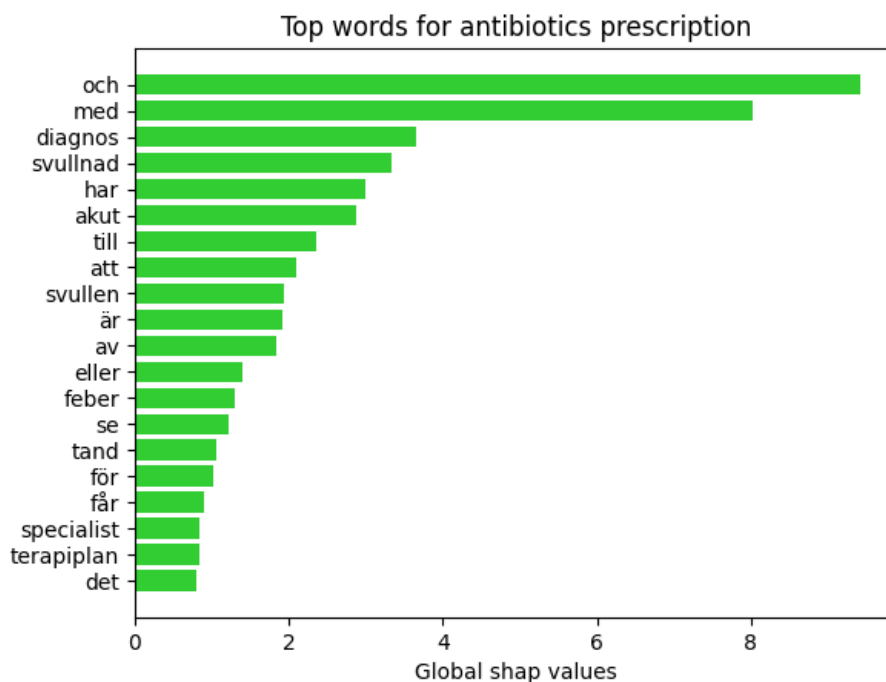


Figure 4.2.9: Global explanations for prescribing antibiotics using IG and KB-BERT.

Table 4.2.3: Common top words found by SHAP and IG with KB-BERT to prescribe antibiotics. The bold words describe the medical condition of the patient.

Common top words
svullnad (swelling)
akut (acute)
diagnos (diagnosis)
eller (or)
tand (tooth)
svullen (swollen)
har (has)
feber (fever)

Many of the words for prescribing antibiotics obtained by IG do not seem as important, at least intuitively, as the words obtained by SHAP. The reason could be, as mentioned in Section 4.2.1, the regularization used by SHAP. When calculating the attributions SHAP has sparser values, unlike IG that gives similar attribution values to multiple words. Since some of these words probably appear frequently in the data, they end being selected in this global explanations.

The top words for not predicting antibiotics of Figure 4.2.10 were very similar to the words obtained by SHAP in Figure 4.2.8. The most interesting feature in this list is the word utförd (performed), that for SHAP was contributing towards predicting antibiotics and for IG that word contributed to the opposite class.

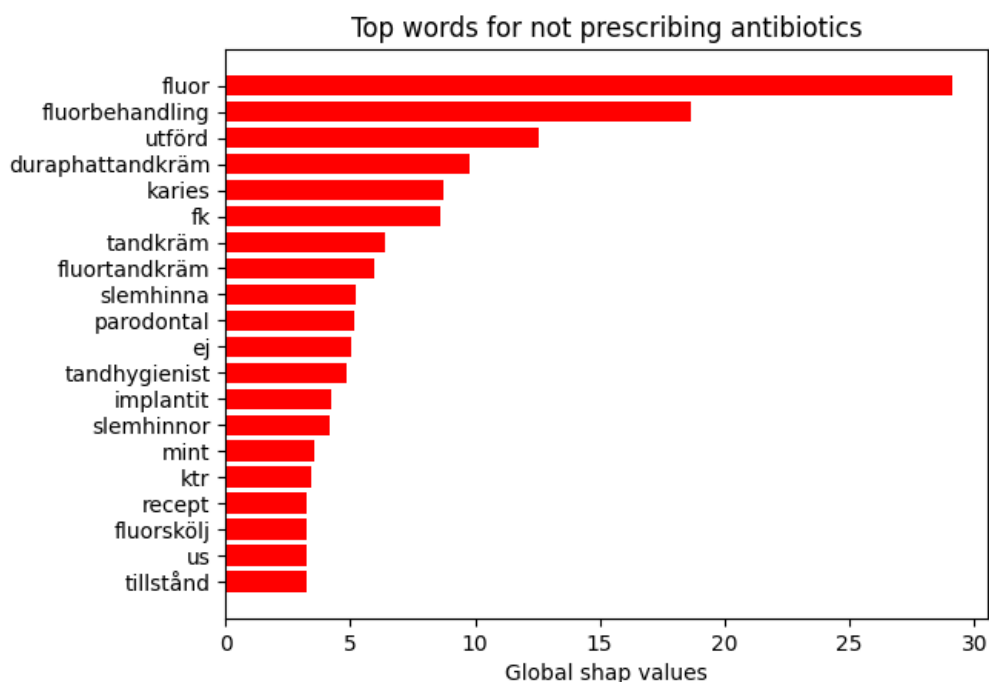


Figure 4.2.10: Global explanations for not prescribing antibiotics using IG and KB-BERT.

Comparison between Transformer models

Figures 4.2.11, 4.2.12 and 4.2.7 are compared in this section to see the different global explanations of different models. Table 4.2.4 shows the common words between models, however, in most cases if we see a bigger picture, e.g., top 100 words, most of the highest words are similar but in different places of importance. Moreover, KB-BERT and KB-ELECTRA have more words in common. This is expected since their architectures and embeddings are similar and both models are monolingual pre-trained with the same data. On the other hand, XLM-R since it is a SentencePiece model, considers spaces and punctuation marks as part of the characters, for example, Figure 4.2.12 has words "akut" and "akut." in the list as two different features. For WordPiece embeddings, on the contrary, punctuation marks are individual features, they are not joined with words.

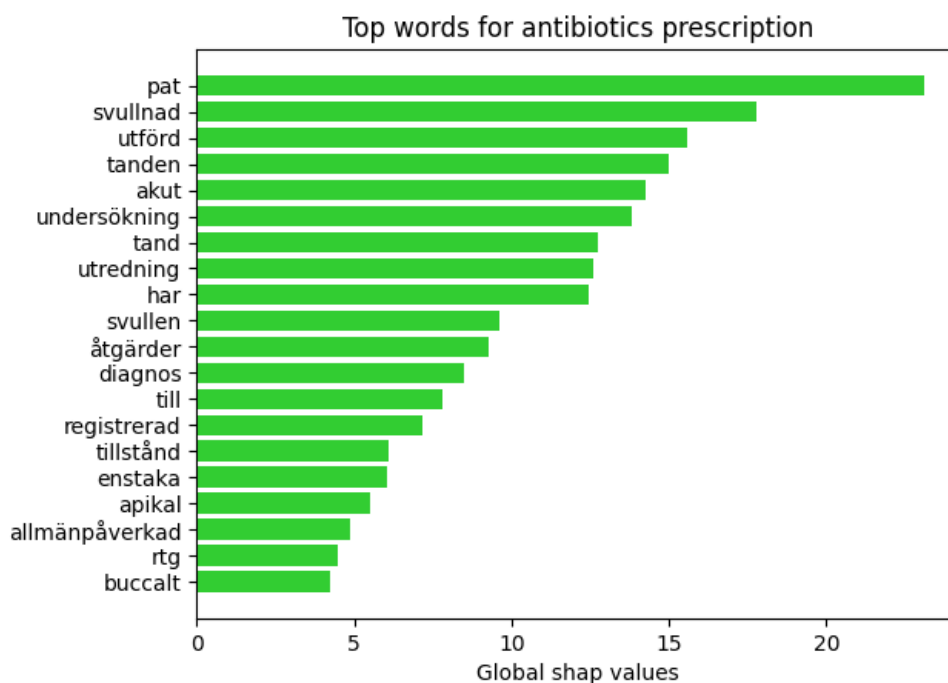


Figure 4.2.11: Global explanations for prescribing antibiotics using SHAP and KB-ELECTRA.

Table 4.2.4: Common top words found by SHAP with KB-BERT, KB-ELECTRA and XLM-R to prescribe antibiotics. The bold words describe the medical condition of the patient.

Common top words
svullnad (swelling)
akut (acute)
diagnos (diagnosis)
svullen (swollen)

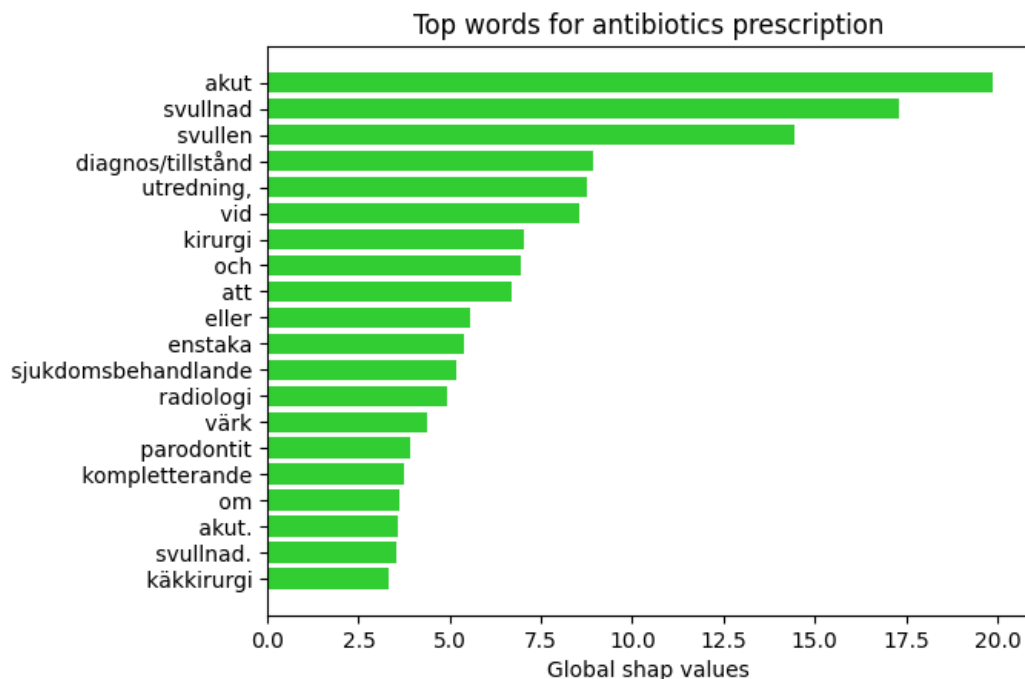


Figure 4.2.12: Global explanations for not prescribing antibiotics using SHAP and XLM-R.

Validation of data cleaning process

To further validate the process of data cleaning, the KB-BERT model was trained with the last version of the dataset without removing names, antibiotics and years. The global explanation for prescribing antibiotics with this dataset and using SHAP is shown in Figure 4.2.13.

The top words in this version of the dataset consist mainly of antibiotics which have a very large aggregated attribution value. The words "pc" (penicillin abbreviated), "kåvepenin", "antibiotika" (antibiotics), "dalacin", "ab" (antibiotics abbreviated), "amimox", "amimoxicillin", "penicillin", "pcv" (Phenoxymethylpenicillin) are all antibiotics related, i.e., the model was indeed heavily relying on them to predict a class.

After inspecting the effect of names and years that were also present in this version of the dataset, despite they contributed towards prescribing antibiotics, they were not as frequent inside the dataset as the antibiotics. As a result, names and years did not reach the top 25 most influential words. Year 2014 was the year with highest attribution value (.3046) for prescribing antibiotics.

Figure 4.2.14, shows the top words that contributed most towards not prescribing antibiotics with KB-BERT and SHAP. Those words are similar to the words of Figure 4.2.8 (the final version of the dataset). Some of the differences are words like "citodon", "ibumetin" and "duraphat", where the first two are pain killers and the last is a fluoride treatment and with the last version of the dataset, the Transformer models do not

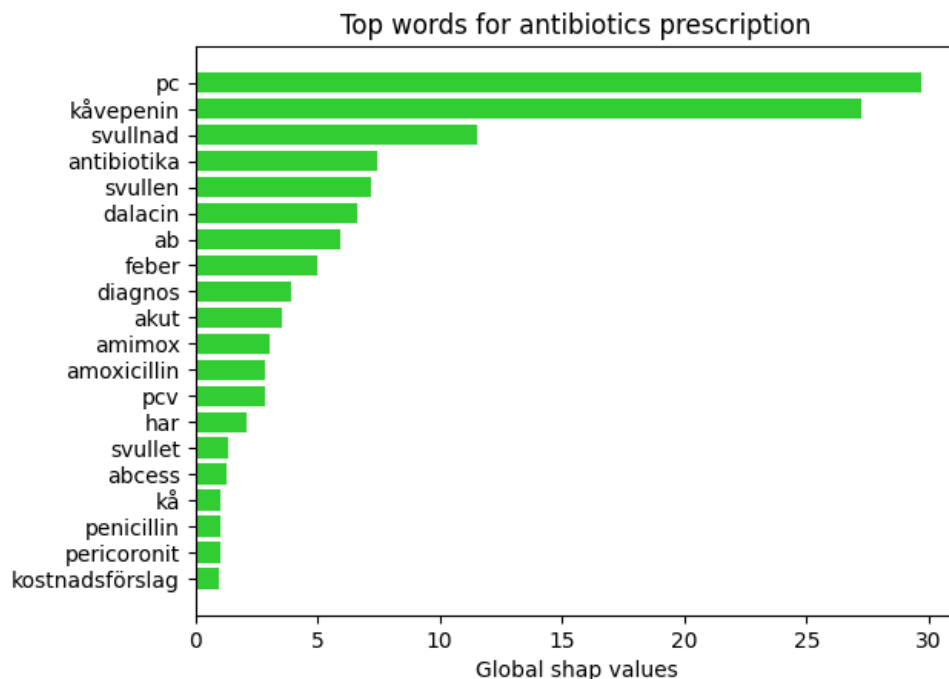


Figure 4.2.13: Global explanations for prescribing antibiotics using SHAP and KB-BERT without removing antibiotics from the dataset.

consider them as important.

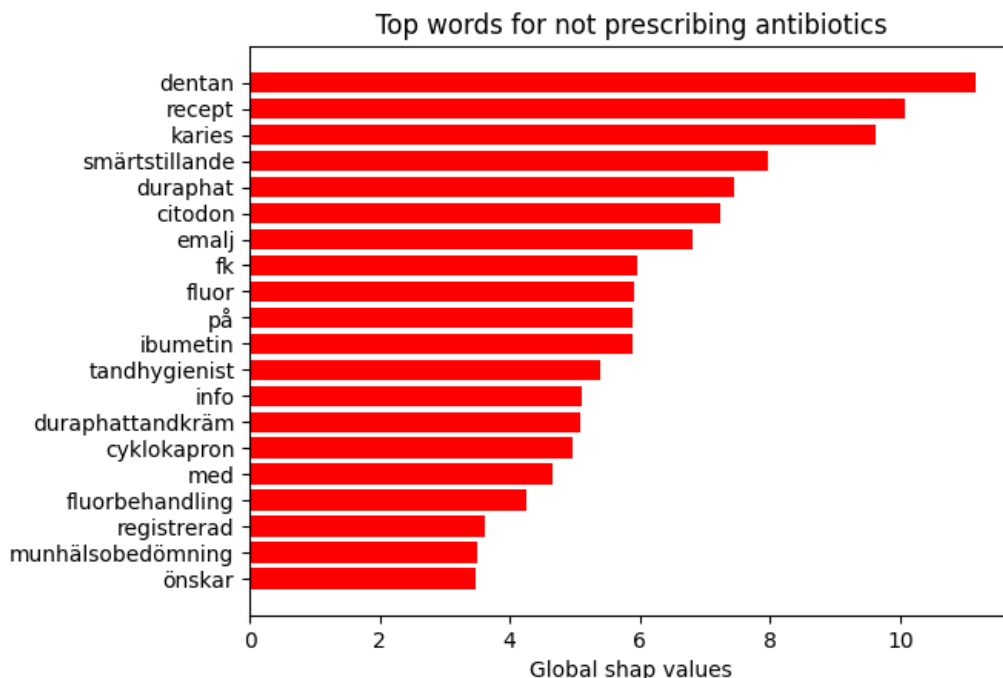


Figure 4.2.14: Global explanations for not prescribing antibiotics using SHAP and KB-BERT without removing antibiotics from the dataset.

In most local and global explanations, the explainability methods highlight words that, at least instinctively, seem like something the models should be focusing, however,

there are also cases where words are highlighted and they do not seem as relevant. XAI is still a research area in progress but the results presented have some positives that encourage further work in this area to keep improving explanations.

Chapter 5

Conclusions

In this work, we demonstrate that NLP models are capable of performing a downstream classification task with a medical domain dataset in Swedish even without further pre-training with unlabeled domain-specific data. Monolingual models prove to obtain better performance in accuracy and F1 score for this task using fewer parameters than the multilingual models.

The local explanations that IG and SHAP produce, show that attribution methods are capable of explaining predictions from NLP models trained with a medical dataset. Moreover, the explanations can be easier to interpret and understand by visualizing words instead of features (tokens). By comparing the attribution methods, the conclusion is that SHAP is easier to interpret since it has a sparser explanation, although, IG or SHAP without regularization are more mathematically correct. Nevertheless, it is important to stress that both methods output similar explanations. Furthermore, by aggregating the weights of individual explanations, global explanations can be obtained, resulting in the words that contributed the most to each class.

It is encouraging for research in the medical domain that global explanations, allow us to observe that NLP models are capable of learning from data some of the correct criteria for prescribing antibiotics. Additionally, it is interesting that the Transformer models tend to focus on similar words to make individual predictions and that the global explanations are also similar between models. Finally, it is demonstrated that explanations can not only be used to improve models by having a better understanding of them, they can be used as guidance for data cleaning as well.

Future Work

This thesis work classifies from patients' medical records if antibiotics should be prescribed or not and explains the reasons behind those decisions. Nevertheless, the classifier (model) is forced to decide between a predefined set of possible outcomes

even though it might have no clue or it could be very uncertain. Therefore, it is crucial to quantify the uncertainty in the models' predictions.

Moreover, as mentioned in Chapter 2.3, NLP models improve significantly when they are pre-trained with unlabeled domain-specific data. Although, medical datasets in Swedish are not common, multilingual models could be pre-trained with a medical English dataset to potentially improve the downstream task in Swedish despite it being a zero-shot learning task.

Another interesting thing to do in the future would be to train a Why mT5? (WmT5) model [31] to obtain generative explanations and compare them with the explanations of the post-hoc feature attribution methods. Additionally, further exploring the effects of the parameters of each explainability method would be interesting and could improve the quality of explanations. In addition, having domain experts (dentists) evaluating the explanations would also be very helpful to improve the model accordingly to their needs.

Finally, a more thorough hyperparameter search could be performed and the larger versions of models like mT5 and XLM-R could be trained to compare their prediction results with the metrics presented in Table 2.1.1.

References

- [1] Alsentzer, Emily, Murphy, John R., Boag, Willie, Weng, Wei-Hung, Jin, Di, Naumann, Tristan, and McDermott, Matthew B. A. “Publicly Available Clinical BERT Embeddings”. In: *CoRR* abs/1904.03323 (2019). arXiv: 1904 . 03323. URL: <http://arxiv.org/abs/1904.03323>.
- [2] Blomgren, Johan. “Förskrivning av antibiotika”. In: *Tandläkartidningen* 5 (2020), pp. 52–57. URL: <https://www.tandlakartidningen.se/wp-content/uploads/2020/04/Blomgren-Jacobsen.pdf>.
- [3] Clark, Kevin, Luong, Minh-Thang, Le, Quoc V., and Manning, Christopher D. “Pre-Training Transformers as Energy-Based Cloze Models”. In: *CoRR* abs/2012.08561 (2020). arXiv: 2012 . 08561. URL: <https://arxiv.org/abs/2012.08561>.
- [4] Conneau, Alexis, Khandelwal, Kartikay, Goyal, Naman, Chaudhary, Vishrav, Wenzek, Guillaume, Guzmán, Francisco, Grave, Edouard, Ott, Myle, Zettlemoyer, Luke, and Stoyanov, Veselin. “Unsupervised Cross-lingual Representation Learning at Scale”. In: *CoRR* abs/1911.02116 (2019). arXiv: 1911 . 02116. URL: <http://arxiv.org/abs/1911.02116>.
- [5] Danilevsky, Marina, Qian, Kun, Aharonov, Ranit, Katsis, Yannis, Kawas, Ban, and Sen, Prithviraj. “A Survey of the State of Explainable AI for Natural Language Processing”. In: *CoRR* abs/2010.00711 (2020). arXiv: 2010 . 00711. URL: <https://arxiv.org/abs/2010.00711>.
- [6] Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, and Toutanova, Kristina. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *CoRR* abs/1810.04805 (2018). arXiv: 1810 . 04805. URL: <http://arxiv.org/abs/1810.04805>.
- [7] Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, and Toutanova, Kristina. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *CoRR* abs/1810.04805 (2018). arXiv: 1810 . 04805. URL: <https://github.com/google-research/bert/blob/master/multilingual.md>.
- [8] Didelot Xavier, Koen B Pouwels. *Machine-learning-assisted selection of antibiotic prescription*. 2019. doi:10.1038/s41591-019-0517-0: NatureMedicine (25(7):1033-1034).

- [9] Dzindolet, Mary T., Peterson, Scott A., Pomranky, Regina A., Pierce, Linda G., and Beck, Hall P. “The role of trust in automation reliance”. In: *International Journal of Human-Computer Studies* 58.6 (2003). Trust and Technology, pp. 697–718. ISSN: 1071-5819. DOI: [https://doi.org/10.1016/S1071-5819\(03\)00038-7](https://doi.org/10.1016/S1071-5819(03)00038-7). URL: <https://www.sciencedirect.com/science/article/pii/S1071581903000387>.
- [10] Erliksson, Karl Fredrik, Arpteg, Anders, Matskin, Mihhail, and Payberah, Amir H. “Cross-Domain Transfer of Generative Explanations using Text-to-Text Models”. In: *26th International Conference on Natural Language and Information Systems, NLDB 2021*. to appear. 2021.
- [11] Gururangan, Suchin, Marasovic, Ana, Swayamdipta, Swabha, Lo, Kyle, Beltagy, Iz, Downey, Doug, and Smith, Noah A. “Don’t Stop Pretraining: Adapt Language Models to Domains and Tasks”. In: *CoRR abs/2004.10964* (2020). arXiv: 2004.10964. URL: <https://arxiv.org/abs/2004.10964>.
- [12] He, Jianxing, Baxter, Sally, Xu, Jie, Xu, Jiming, Zhou, Xingtao, and Zhang, Kang. “The practical implementation of artificial intelligence technologies in medicine”. In: *Nature Medicine* 25,1 (Jan. 2019). DOI: 10.1038/s41591-018-0307-0.
- [13] He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification”. In: *CoRR abs/1502.01852* (2015). arXiv: 1502.01852. URL: <http://arxiv.org/abs/1502.01852>.
- [14] Hendrycks, Dan and Gimpel, Kevin. “Bridging Nonlinearities and Stochastic Regularizers with Gaussian Error Linear Units”. In: *CoRR abs/1606.08415* (2016). arXiv: 1606.08415. URL: <http://arxiv.org/abs/1606.08415>.
- [15] Jain, Sarthak, Mohammadi, Ramin, and Wallace, Byron C. “An Analysis of Attention over Clinical Notes for Predictive Tasks”. In: *CoRR abs/1904.03244* (2019). arXiv: 1904.03244. URL: <http://arxiv.org/abs/1904.03244>.
- [16] Johnson, Alistair, Pollard, Tom, Shen, Lu, Lehman, Li-wei, Feng, Mengling, Ghassemi, Mohammad, Moody, Benjamin, Szolovits, Peter, Celi, Leo, and Mark, Roger. “MIMIC-III, a freely accessible critical care database”. In: *Scientific Data* 3 (May 2016), p. 160035. DOI: 10.1038/sdata.2016.35.
- [17] Khandelwal, Renu. *Understanding Deep Learning Models with Integrated Gradients*. 2020. URL: <https://towardsdatascience.com/understanding-deep-learning-models-with-integrated-gradients-24ddce643dbf>.
- [18] Kokhlikyan, Narine, Miglani, Vivek, Martin, Miguel, Wang, Edward, Alsallakh, Bilal, Reynolds, Jonathan, Melnikov, Alexander, Kliushkina, Natalia, Araya, Carlos, Yan, Siqi, and Reblitz-Richardson, Orion. *Captum: A unified and generic model interpretability library for PyTorch*. 2020. arXiv: 2009.07896 [cs.LG].

REFERENCES

- [19] Kudo, Taku. “Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates”. In: *CoRR* abs/1804.10959 (2018). arXiv: 1804.10959. URL: <http://arxiv.org/abs/1804.10959>.
- [20] Kudo, Taku and Richardson, John. “SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing”. In: *CoRR* abs/1808.06226 (2018). arXiv: 1808.06226. URL: <http://arxiv.org/abs/1808.06226>.
- [21] Lan, Zhenzhong, Chen, Mingda, Goodman, Sebastian, Gimpel, Kevin, Sharma, Piyush, and Soricut, Radu. “ALBERT: A Lite BERT for Self-supervised Learning of Language Representations”. In: *CoRR* abs/1909.11942 (2019). arXiv: 1909.11942. URL: <http://arxiv.org/abs/1909.11942>.
- [22] Lee, Jinhyuk, Yoon, Wonjin, Kim, Sungdong, Kim, Donghyeon, Kim, Sunkyu, So, Chan Ho, and Kang, Jaewoo. “BioBERT: a pre-trained biomedical language representation model for biomedical text mining”. In: *CoRR* abs/1901.08746 (2019). arXiv: 1901.08746. URL: <http://arxiv.org/abs/1901.08746>.
- [23] Liu, Yinhan, Ott, Myle, Goyal, Naman, Du, Jingfei, Joshi, Mandar, Chen, Danqi, Levy, Omer, Lewis, Mike, Zettlemoyer, Luke, and Stoyanov, Veselin. “RoBERTa: A Robustly Optimized BERT Pretraining Approach”. In: *CoRR* abs/1907.11692 (2019). arXiv: 1907.11692. URL: <http://arxiv.org/abs/1907.11692>.
- [24] Lund, Bodil, Cederlund, Andreas, Hultin, Margareta, and Lundgren, Frida. “Effect of governmental strategies on antibiotic prescription in dentistry”. In: *Acta Odontologica Scandinavica* 78.7 (2020). PMID: 32293215, pp. 529–534. DOI: 10.1080/00016357.2020.1751273. eprint: <https://doi.org/10.1080/00016357.2020.1751273>. URL: <https://doi.org/10.1080/00016357.2020.1751273>.
- [25] Lundberg, Scott and Lee, Su-In. “A unified approach to interpreting model predictions”. In: *CoRR* abs/1705.07874 (2017). arXiv: 1705.07874. URL: <http://arxiv.org/abs/1705.07874>.
- [26] Malmsten, Martin, Börjeson, Love, and Haffenden, Chris. “Playing with Words at the National Library of Sweden - Making a Swedish BERT”. In: *CoRR* abs/2007.01658 (2020). arXiv: 2007.01658. URL: <https://arxiv.org/abs/2007.01658>.
- [27] Martin, Louis, Müller, Benjamin, Suárez, Pedro Javier Ortiz, Dupont, Yoann, Romary, Laurent, Clergerie, Éric Villemonte de la, Seddah, Djamé, and Sagot, Benoit. “CamemBERT: a Tasty French Language Model”. In: *CoRR* abs/1911.03894 (2019). arXiv: 1911.03894. URL: <http://arxiv.org/abs/1911.03894>.
- [28] Meta-Wiki. *List of Wikipedias*. URL: https://meta.wikimedia.org/wiki/List_of_Wikipedias.

- [29] Molnar, Christoph. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. <https://christophm.github.io/interpretable-ml-book/>. 2019.
- [30] Mullenbach, James, Wiegrefe, Sarah, Duke, Jon, Sun, Jimeng, and Eisenstein, Jacob. “Explainable Prediction of Medical Codes from Clinical Text”. In: *CoRR abs/1802.05695* (2018). arXiv: 1802.05695. URL: <http://arxiv.org/abs/1802.05695>.
- [31] Narang, Sharan, Raffel, Colin, Lee, Katherine, Roberts, Adam, Fiedel, Noah, and Malkan, Karishma. “WT5?! Training Text-to-Text Models to Explain their Predictions”. In: *CoRR abs/2004.14546* (2020). arXiv: 2004.14546. URL: <https://arxiv.org/abs/2004.14546>.
- [32] Organization, World Health. *Antibiotics Resistance*. 2020. URL: <https://www.who.int/news-room/fact-sheets/detail/antibiotic-resistance>.
- [33] Raffel, Colin, Shazeer, Noam, Roberts, Adam, Lee, Katherine, Narang, Sharan, Matena, Michael, Zhou, Yanqi, Li, Wei, and Liu, Peter J. “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer”. In: *CoRR abs/1910.10683* (2019). arXiv: 1910.10683. URL: <http://arxiv.org/abs/1910.10683>.
- [34] Reimers, Nils and Gurevych, Iryna. “Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation”. In: *CoRR abs/2004.09813* (2020). arXiv: 2004.09813. URL: <https://arxiv.org/abs/2004.09813>.
- [35] Reimers, Nils and Gurevych, Iryna. “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks”. In: *CoRR abs/1908.10084* (2019). arXiv: 1908.10084. URL: <http://arxiv.org/abs/1908.10084>.
- [36] Ribers, Michael Allan and Ullrich, Hannes. *Battling Antibiotic Resistance: Can Machine Learning Improve Prescribing?* 2019. arXiv: 1906.03044 [econ.GN].
- [37] Schuster, Mike and Nakajima, Kaisuke. “Japanese and Korean Voice Search”. In: *International Conference on Acoustics, Speech and Signal Processing*. 2012, pp. 5149–5152.
- [38] Shapley, L. S. “17. A Value for n-Person Games”. In: *Contributions to the Theory of Games (AM-28), Volume II*. Ed. by Harold William Kuhn and Albert William Tucker. Princeton University Press, 2016, pp. 307–318. DOI: doi:10.1515/9781400881970-018. URL: <https://doi.org/10.1515/9781400881970-018>.
- [39] Sundararajan, Mukund, Taly, Ankur, and Yan, Qiqi. “Axiomatic Attribution for Deep Networks”. In: *CoRR abs/1703.01365* (2017). arXiv: 1703.01365. URL: <http://arxiv.org/abs/1703.01365>.
- [40] University of Gothenburg Sweden, Department of Swedish. *Språkbanken text*. URL: <https://spraakbanken.gu.se/en/resources/suc3>.

REFERENCES

- [41] Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., Kaiser, Lukasz, and Polosukhin, Illia. “Attention Is All You Need”. In: *CoRR* abs/1706.03762 (2017). arXiv: 1706.03762. URL: <http://arxiv.org/abs/1706.03762>.
- [42] Virtanen, Antti, Kanerva, Jenna, Ilo, Rami, Luoma, Jouni, Luotolahti, Juhani, Salakoski, Tapio, Ginter, Filip, and Pyysalo, Sampo. “Multilingual is not enough: BERT for Finnish”. In: *CoRR* abs/1912.07076 (2019). arXiv: 1912.07076. URL: <http://arxiv.org/abs/1912.07076>.
- [43] Wadden, David, Lo, Kyle, Wang, Lucy Lu, Lin, Shanchuan, Zuylen, Madeleine van, Cohan, Arman, and Hajishirzi, Hannaneh. “Fact or Fiction: Verifying Scientific Claims”. In: *CoRR* abs/2004.14974 (2020). arXiv: 2004.14974. URL: <https://arxiv.org/abs/2004.14974>.
- [44] Wang, Alex, Pruksachatkun, Yada, Nangia, Nikita, Singh, Amanpreet, Michael, Julian, Hill, Felix, Levy, Omer, and Bowman, Samuel R. “SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems”. In: *CoRR* abs/1905.00537 (2019). arXiv: 1905.00537. URL: <http://arxiv.org/abs/1905.00537>.
- [45] Xue, Linting, Constant, Noah, Roberts, Adam, Kale, Mihir, Al-Rfou, Rami, Siddhant, Aditya, Barua, Aditya, and Raffel, Colin. “mT5: A massively multilingual pre-trained text-to-text transformer”. In: *CoRR* abs/2010.11934 (2020). arXiv: 2010.11934. URL: <https://arxiv.org/abs/2010.11934>.