UTRECHT UNIVERSITY

MASTER THESIS

Domain-specific text classification: determining medical outcomes using free text in electronic patient records

Author:
Ilse GRIFFIOEN

Supervisor:
dr. Kees DE SCHEPPER

1st Examiner:
dr. Nico ROMEIJN

2nd Examiner:
dr. Christian JANSSEN

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in the Subject of Artificial Intelligence

in the

Graduate School of Natural Sciences

August 30, 2019

Abstract

Background In modern healthcare, outcome measures are considered of high value. They are used to make discharge decisions, to reflect on a specific treatment and to decide whether changes in the treatment plan need to be made. In psychiatry, these outcome measures are often only obtained at the start and end of a hospital admission, which does not cover the complete picture. However, it is also desirable to get a picture from the time between admission an discharge.

Method A support vector machine model was created that classifies daily written nurse reports in either being written at the start of admission or in the last days before discharge. In order to attain a measurement of patient well-being we make the assumption that patients suffer from serious mental problems at the beginning of admission and that their symptoms are reduced or at least have stabilized when they are discharged, thereby linking time of the report to mental state of well-being. For unseen nurse reports written in the days between admission and discharge the model predicts whether the report has been written in the last days before discharge with a certain probability. From these probability rates a line chart is created that follows the course of an admission. Higher probability rates are possibly able to show patient improvement, while lower probability rates may indicate patient impairment. Model results were compared with findings in the literature and with reflections made by human annotators who rated patient improvement during admission.

Results The model was able to predict whether a report was written at the start or at the end of admission with a 92% accuracy. The mean line chart shows a decelerating curve that follows the findings in the literature. Moments where the model found exceptionally high or low probability rates were indicated as apparent patient improvement or impairment four out of eight times by annotators. Overall, annotators observed more of these moments than the model did. Words that the model found indicating either improvement (high probability rates) or impairment (low probability rates) were indicated by annotators for 14.8% of the words.

Discussion The results suggest that the method we tested to find patterns of patient improvement and impairment has potential, although the model's explanations did not comply with human explanations. Since this project was a first attempt at getting insight in outcome measures through nurse reports, improvements should be made to the model and other methods should be explored.

Acknowledgements

During this project I have received a great deal of support from several different actors. Therefore, I would like to take a moment to show my gratitude towards anyone that contributed in the preparation and/or completion of this thesis.

First and foremost, I would like to thank my daily supervisor dr. Kees de Schepper. During the thesis project he has been of great value to me. Not only was he there for me during our weekly meetings for which he always took the time to go through the core of the project and discuss the considerations needed to be made, but he was also there when I got stuck coding where he did not mind to assist until the error was found. The way in which he guided me through this, often times, difficult thesis trajectory deserves a great amount of respect, gratitude and appreciation.

Second, I would like to thank my first examiner dr. Nico Romeijn for his insights, feedback and encouraging words that gave me the drive necessary to go the extra mile. Third, I would like to thank dr. Christian Janssen for bringing me in contact with Nico and for his critical notes on the scope of the project.

Finally, this project would not have come to an end without the help of my colleagues at the department who participated as annotators and I would like to thank them for all the fun and support they provided me during this project.

Contents

1 Introduction

In modern healthcare, outcome measures[1] are considered of high value [2, 3]. Firstly, it is important for practitioners to know whether their patients are improving and whether their treatment plan has the desired impact [4]. Secondly, patients themselves have the desire to obtain an indication of the measure of improvement, in order to know where they are within their care process. Thirdly, health insurers are interested in the quality-to-price ratio, to be able to compare treatments and healthcare providers [5]. Lastly, outcome measures are important for research. When changes in the patient's care process are reported within a standard format, statistical tests can be performed on this data. This allows us to analyze the efficacy of different treatment approaches and improve personalized care [6, 5].

Within the Electronic Health Record (EHR), care providers register health outcomes of their patients. Therefore, outcome measures can be obtained from structured data that is documented conforming to a predefined standard [7]. Examples of structured data are rating scales for symptom severity [8] and medication or lab measurements [9]. In order to use outcome measures for analysis it is important that enough measurements have been conducted. For instance, without pre- and post-measurements it is difficult to ascertain whether a patient's condition has improved after receiving care [10]. However, the use of measuring instruments for patient evaluation is time consuming and often requires well-trained raters [11, 12].

The EHR does not entirely consist out of structured data. A large part within the EHR documentation consists of free text. Research has shown that healthcare professionals prefer the flexibility and efficiency associated with documenting using free text [7, 13]. These free text fields contain more detailed information about a patient's health status, that can not be obtained from structured data [14]. Analyzing free text in medicine is much harder than analyzing structured data, considering the heterogeneity of the data, the lack of canonical forms and the richness of spelling and typing errors [13, 15]. Nevertheless, during the past two decades, progress has been made in applying text-mining techniques to these unstructured clinical records [16, 17, 18].

At the Psychiatry Department of the University Medical Center Utrecht (UMCU) in The Netherlands, research is being conducted in extracting outcome measures from the free text fields in the EHR. The current study is part of the research project. The objective of this study is to examine the usefulness of applying text classification to written nurse reports of in-patients over the course of a psychiatric admission, to identify patterns of improvement and impairment during admission.

A comparable study was performed by Page, Cunningham, and Hooke [19]. In their study, patients were asked to complete the Five Item Daily Symptom Index [20] on a daily

[1] An outcome measure is defined as *"any characteristic or quality measured to assess a patient's status."* [1, p.163]

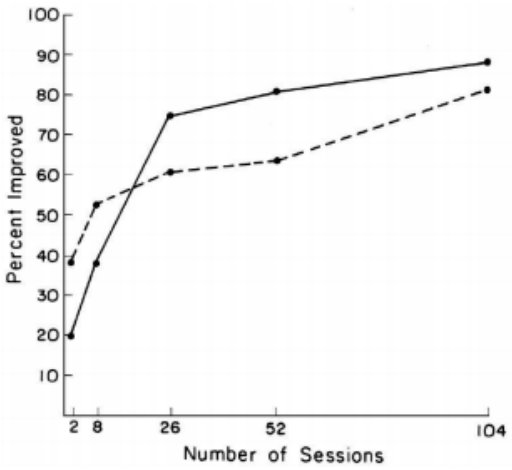

FIGURE 1.1: **Relation of number of sessions (dose) of psychotherapy and percentage of patients improved (effect).(Objective ratings at termination are shown by the solid line; subjective ratings during therapy are shown by the broken line.)** Reprinted from *"The Dose–Effect Relationship in Psychotherapy,"* , by K.I. Howard, S.M. Kopta, M.S. Krause, and D.E. Orlinsky, 1986, American Psychologist, 41, p. 160.

basis. Patterns of improvement and impairment were measured according to the level of symptom reduction or deterioration indicated by patients themselves. Although patient reported outcome measures is an increasing subject of attention [21, 22], we are facing a shifting paradigm of healthcare, where it is not just about remedying disease and symptoms, but rather about improving one's health [23]. Huber et al. [23] introduce a new concept for the definition of health: *"the ability to adapt and to self-manage, in the face of social, physical and emotional challenges"*. In the current study the choice was made to use nurse reports due to their holistic nature. Instead of only documenting the symptoms of a patient the overall improvement of a patient is documented within this report, which aligns with the shifting paradigm of healthcare.

Other studies with a focus on identifying patterns of improvement and impairment during treatment in psychiatry are found in the field of psychotherapy [24, 25, 26]. Data was collected from objective ratings by researchers on closed patient charts and subjective ratings on session reports filled in by patients themselves. Howard et al. [27] describe a model (Figure 1.1) that fits a three-phase healing process that goes beyond symptom reduction: *"a) remoralization, the enhancement of well-being, which is usually accomplished within a few sessions; (b) remediation, the attainment of symptomatic relief, which is accomplished more gradually; and (c) rehabilitation, the unlearning of troublesome, maladaptive, long-standing behaviors and the establishing of new ways of dealing with various aspects of life, which occurs even more gradually"*. The model follows a decelerating curve of patient improvement, which is in line with the increasing difficulty of achieving a, b and c during treatment [26]. Despite the differences in outpatients in psychotherapy and inpatients in a psychiatric care unit, it is hypothesized that the average patient included in our study will follow this model. The main drawback of the model in Figure 1.1 is addressed by Lutz, Martinovich, and Howard [25], arguing that patterns of improvement vary extensively between patients and that the expected course of treatment depends on the characteristics of the patient and their environment.

The use of text-classification to identify patterns of improvement or impairment over the course of a psychiatric admission has, to the best of our knowledge, not been explored

before. In order to attain a measurement of patient well-being we make the assumption that patients suffer from serious mental problems at the beginning of admission and that their symptoms are reduced or at least have stabilized when they are sent home. This assumption fits the baseline model in Figure 1.1. A support vector machine (SVM) model is trained on nurse reports written in the first three days of admission and the last three days before discharge. Previously unseen nurse reports written in the days between are subsequently fed to the model. The classification task is to predict for these nurse reports if they are more likely to be written at the start or at the end of admission, which would coincide with an improved mental health status. A line chart can be created from the probability rates for likeliness that a report was written at the end of admission. The aim is to link increased probability rates with patient improvement and decreased probability rates with patient impairment. The decision not to use structured data from questionnaires as a validity measure is made since the data was only available for a short amount of patients. In order to obtain a first impression of the goodness of the model, six admission reports are double reviewed by humans and compared with the model's outcomes.

The research question we propose is: 'Can assumed mental health improvement be predicted by text classification on nurse reports, where the model is trained solely on text written at the beginning and end of a hospital admission?'. Since nurse reports are written with the purpose of informing colleagues about a patient's status, it is expected that a machine learning model is able to detect patterns in the reports.

1.1 Clinical relevance

Outcome measures concerning a patient's status are often obtained only at admission and discharge. By obtaining outcome measures over the course of admission, besides determining whether patients are responding to a treatment at all during their stay, we are also informed about the moment in time this response happened and whether there was a relapse or not. This information can aid clinicians in making discharge decisions and changes in the treatment plan. In addition, this information can be used when treating new patients with a similar background. [28, 29].

1.2 Relevance to the field of Artificial Intelligence

In the past decades, there have been an increase in the adoption of EHRs among healthcare institutions. This shift in information registration from paper-based to electronic medical record allows us to reuse medical data [30] and to explore the benefits of machine learning [31]. AI offers a lot of value to the medical field, such as supporting physicians in decision making [32], reducing diagnostic errors [33] and detecting unusual structures in medical images [34].

On the other side, the accessibility of clinical data has shown new perspectives for the field of AI as well. As was explained in the introduction, analyzing free text is more difficult than analyzing structured data. Clinical text in particular offers a challenge for AI. Unlike biomedical text[2], clinical text often contains spelling and grammatical errors, abbreviations and acronyms, and institution-specific templates with normative structures. Therefore, preprocessing clinical text is very important [35]. In English, there are several online databases

[2]We define biomedical text as the text we find in books, articles, papers, etcetera [35].

and knowledge bases available containing clinical narratives and dictionaries [36]. For other languages, including Dutch, these databases are not widely accessible and often may be used for in-house research purposes only.

The current study in particular offers a new method of applying machine learning to medical data. Thereby extending the application of AI to new fields that possibly are in great need of such applications.

2 Materials

This section describes the data used in this research, the pre-processing steps that are taken and the way the nurse reports are binned.

2.1 Data selection

For this research we had access to de-identified admission data and nurse reports from the Psychiatry Department of the UMCU [37]. The department is divided in six units, each with the focus on one of the four care pathways: affective and psychotic disorders, acute and intensive care, risk and prevention, and developmental disorders. Both adult and child patients are treated here. Adult patients who are admitted to the department are mainly treated for psychotic disorders, like Schizophrenia and Depression. The department is known for delivering care to patients with severe or complex psychiatric symptoms.

In order to obtain a relevant dataset for the classification task several exclusion criteria are formulated. Following the advise of psychiatrists assisting this research patients younger than 18 years during admission are excluded. To reduce the occurrence of self-determined premature discharge, we excluded patients who left the hospital against recommendation of the treating physician. Furthermore, patients who, after hospital admission, were transferred to psychiatric or nursing facilities are also excluded. Simultaneously, we excluded patients who had an admission duration of less than 14 days, since we expect for these patients that it will be difficult to find significant changes during their stay. In 2013 the department initiated a reorganization, which meant a change in sub-departments and administration. Therefore, we excluded patients who were admitted before 2013. Lastly, all patients with ongoing treatment at the moment of data retrieval (February 2019) are excluded.

The resulting dataset consists of 151,891 nurse reports of 1,104 in-hospital stays of 921 patients. Since the research objective is to find changes in well-being during a hospital admission, readmissions and transfers within the department are included. In Table 2.1 descriptive statistics of the included patient population are shown.

TABLE 2.1: **Descriptive Statistics of the patient population ($N = 1104$).**

Categorical variable	**n(%)**	
Gender		
Male	582	(52.7)
Female	522	(47.3)
Main diagnosis		
Schizophrenia and other psychotic disorders	430	(38.9)
Depressive disorders	255	(23.1)
Bipolar disorders	179	(16.2)
Personality disorders	51	(4.6)
Autism spectrum disorder	49	(4.4)
Addictive disorders	25	(2.3)
Anxiety disorders	19	(1.7)
Trauma- and stressor-related disorders	15	(1.4)
Psychotic disorder due to another medical condition	13	(1.2)
Other or unknown disorders	68	(6.2)
Continuous variable	**mean±SD**	
Age (years)	38	±15
Length of stay (days)	50	±49

2.2 Preprocessing

We used the pretrained word2vec architecture, that will be explained in Section 3, of Menger, Scheepers, and Spruit [18] to create text representations. Therefore, the preprocessing steps taken are similar to those of Menger et al, except for the expanded list of stop words. All computations are performed with Python version 3.7.1.

The nurse reports are preprocessed by transforming all capital letters to lower case letters. Thereafter, all non-ASCII characters are replaced by their ASCII counterparts (e.g. ë → e), using the Python package `Unidecode` [38]. In addition, a list of strings is created that are contained in the nurse report template (e.g. '1. psychische problemen', '2. sociale vaardigheden'). These strings are removed from the text so that they can not skew the result of the analysis. Furthermore, all double spaces, tabs, numbers and non-alphanumeric characters are removed. Periods are removed as long as they are not part of an abbreviation. The Natural Language Toolkit (NLTK) [39] is used to tokenize (splitting sentences into words) the text while simultaneously removing stop words. The list of stop words is expanded with words that directly indicate the beginning or end of a hospital admission (Table 2.2), to avoid the model of becoming biased against these words.

2.3 Binning

From the complete set of 151,891 nurse reports, only the reports written in the first three days after admission and the last three days before discharge are used for model training and testing. Nurse reports are written three times a day on average, which implies that for each admission we included approximately 18 nurse reports. The resulting dataset consists

TABLE 2.2: **List of removed stop words that directly indicate the beginning or end of a hospital admission**

Dutch	**English translation**
opname	admission
opgenomen	hospitalized
rondleiding	tour
gewend	got used to
opnamegesprek	admission interview
anamnese	anamnesis
ontslag	discharge
ontslaggesprek	discharge interview
ontslagbrief	discharge letter
ontslagmedicatie	discharge medication
ontslagdatum	discharge date
huis	house
afscheid	goodbye
nazorg	aftercare
inpakken	pack
spullen	stuff
gaan	going
recept	recipe
meegegeven	handed over
maandag - zondag	monday - sunday[1]
opgehaald	picked up

[1]Planned day for discharge

of 25,152 nurse reports. The three day span is chosen as a middle-ground solution. By choosing a smaller amount of days the chance of creating a model that is biased towards words that are likely to be written on the day of admission/discharge is higher. By choosing a larger amount of days the chance of patients already getting better in the first days and are still in need of treatment in the last days is higher [19]. Nurse reports written in the first three days after admission received label '0'. Nurse reports written in the last three days before discharge received label '1'. Thereafter, the nurse reports are aggregated on admission ID and label, resulting in documents containing nurse reports of one patient of three consecutive days with a mean of 1,232 words. In the remaining of the study, aggregated nurse reports of three consecutive days will be addressed as a document. All documents are subsequently divided in a 80% training- (1,766 documents) and 20% testset (442 documents). The model's classification task is to predict the right label for unseen documents in the testset.

After the model is trained, the complete set of 151,891 nurse reports is needed to identify the patterns of improvement or impairment during a hospital admission. For this task, previously assigned labels are removed and nurse reports are binned according to the following function:

$$Bin = [(n + 2)/3]$$

Where n is defined as the number of days after admission the report was written. This function is chosen to assign nurse reports written on day 0, 1 and 2 to the first bin, reports written on day 3, 4 and 5 to the second bin, and so forth. Again, nurse reports are aggregated on admission ID and bin. This results in an evaluation set of 19,129 documents.

3 Methods

This section explains the choices made regarding the methods and models used to perform the classification task.

3.1 Text representation

Classification models are unable to use unstructured text. Therefore, the text needs to be converted into structured data. This enables the models to perform calculations. Text representation methods are used to convert text into vectors. This conversion can work in different ways, for instance, text can be converted on word level, sentence level or even on the document level. Text representation methods are built on the premise that texts found in close proximity in the vector space are similar to one another [40]. By measuring the cosine or Euclidean distance between two text vectors, the similarity between these vectors can be expressed as a value.

Bag of Words (BOW) is the most commonly used method for representing text [41]. First, a vocabulary is created, containing all unique words in the text corpus, which is the complete training dataset. Hereafter, each document is converted into a vector with a length equal to the vocabulary size. A commonly used technique for creating these vectors is to assign a binary scoring of the presence or absence of a word in the document. BOW is a favorable method when the dataset is small or the domain is very specific. For large corpora of training data containing documents from different domains, the vocabulary size can grow largely. Since documents contain only a small part of the words in the vocabulary, their vector representations will be sparse. Sparse vectors can be computationally expensive to process and require a lot of memory. Nevertheless, in recent years algorithms have been developed to overcome these problems for sparse vectors [42].

However, the main drawback of BOW is that similarities between documents are measured based on the words they both contain, without taking the meaning of these words into account. Nurse reports contain a lot of spelling mistakes, since nurses do not have a lot of time for reporting. Besides, different abbreviations are used to define the same concept. As a consequence, the vocabulary contains a lot of words with similar meanings that were spelled differently. Table 3.1 shows an example of how this can be problematic for BOW. We presume that Document 1 and Document 2 belong to the same class, since they are both optimistic. Yet, when we look at the vector representations of these documents, Document 1 shows as little similarity with Document 3 as it does with Document 2.

In order to overcome the limitations of the traditional BOW, word embeddings are applied. Word embeddings can represent text in a way that words with similar meanings have similar representations in the vector space [43]. For the classification task at hand the trained

TABLE 3.1: **Bag of Words representation with binary scoring.**

Vocabulary:	[feels,	good,	today,	very,	happy,	bad,	day]
Doc 1: 'Feels good today'	[1,	1,	1,	0,	0,	0,	0]
Doc 2: 'Is very happy'	[0,	0,	0,	1,	1,	0,	0]
Doc 3: 'Has a bad day'	[0,	0,	0,	0,	0,	1,	1]

word2vec[1] with Continuous Bag of Words (CBOW) architecture of Menger et al. is used as vocabulary to build features [18]. CBOW[2] generates word vectors by taking the surrounding context of a word in the document as input for that word and feed them to a single hidden layer with linear neurons. The vector representation of a word is thus defined by its surrounding context words. CBOW can be regarded as both supervised and unsupervised. The supervised task is that a word is predicted from its context, where the word itself can be considered the label. On the other hand, CBOW can be considered unsupervised, since it learns independently from human labelling [45].

Pre-trained embedding architectures using the word2vec algorithm are available online. These architectures are trained on big corpora of text including a variety of topics [46]. The dataset used for the classification task at hand is domain specific and is therefore expected to perform badly when used as input to a generic word embedding architecture [47]. The pre-trained word2vec architecture of Menger et al. [18] is trained on nurse reports and doctor's notes generated by the healthcare professionals in the UMCU. Therefore, this model is well suited to be applied to the dataset.

Normalization is applied using term frequency - inverse document frequency (tf-idf). Tf-idf is a weighting factor that increases when the term is relatively important in the document and thus corrects for words that generally appear more frequently in all documents. The weighting factor is measured by multiplying the tf function (the number of times a word appears in a document) by the idf function (the log of the total number of documents divided by the number of documents the word appears in). Here the log causes a dampening effect on the idf function, since the nominator is likely to vary greatly for different words.

3.2 Model selection

When choosing an algorithm for a machine learning task, there is no way of ensuring that the chosen algorithm is the best one for that specific task [48]. As stated in the No Free Lunch Theorem, there is no universally best algorithm that fits all problems [49]. Nevertheless, based on the characteristics of popular machine learning algorithms, the data, and the type of problem, an estimation of potentially well performing algorithms can be made.

To start the search for an applicable machine learning algorithm, a flowchart offered by scikit-learn.org is used [50]. Scikit-learn provides an open source python library with a wide selection of supervised and unsupervised machine learning algorithms [51]. Following the flowchart: since the prediction task is 'text classification', the data is labeled, and the dataset contains less than 100,000 documents, a Support Vector Machine (SVM) with linear kernel is recommended [50]. The SVM is a supervised algorithm that aims at finding the best decision boundary in the vector space that maximizes the distance between the vectors and the decision boundary. For binary classification problems this means that the optimal

[1] https://code.google.com/archive/p/word2vec/
[2] A more extensive explanation of the CBOW architecture is provided by Mikolov et al. [44]

decision boundary divides the classes in a way that the distance to both classes is equal and as large as possible. With kernel parameters the type of decision boundary can be adjusted. A linear kernel will try to separate the data linearly, while the radial bias function (RBF) and polynomial kernels separate the data non-linearly.

Experiments performed by Joachims on the Reuters-21578 dataset, which is one of the most widely used databases for text categorization, have shown that SVMs are likely to outperform more simplistic classifiers as Naive Bayes and k-nearest neighbors [52]. SVM with linear kernel was able to outperform other models on medical text classification problems in preliminary studies [53, 54, 55]. However, when more kernel parameters are explored, SVM with RBF kernel is likely to outperform SVM with linear kernel [18, 56, 57]. Therefore, the SVM with RBF kernel was included in our experiments.

Further analyses of the characteristics of the SVM algorithm and the nurse reports indicate that the SVM is likely to be a well suited algorithm for the considered classification task. Considering that the SVM uses a penalty parameter C to prevent it from overfitting on the training data, it has the potential to handle large feature spaces. Moreover, the algorithm does not only have to deal with a lot of features, in text classification all of these features (the words) are important as well. More transparent algorithms, like Naive Bayes, use feature selection in an attempt to reduce the feature space. These algorithms are more likely to overfit on the training data when all features are included. As a consequence, there will be loss of information [52].

An initial SVM model is set up using the `svm` module of the scikit-learn library in Python. Hyper-parameter tuning is performed on the three parameters C, γ, and kernel (RBF or linear) to find a classifier that predicts labels for the test set by optimizing the Area Under the Curve (AUC) of the Receiver Operating Characteristics (ROC) curve [58]. The ROC is a probability curve that is used to present the balance between the fraction of documents correctly predicted to have label '1' and the fraction of documents wrongly predicted to have label '1'. AUC is the measure of separability. The penalty parameter C controls the trade-off between classifying training points correctly and maximizing the margin of the decision boundary. A larger value of C will classify more training points correctly, but therewith a smaller margin for the decision boundary is allowed. With γ one can control the influence of a single data point, where larger values will affect the decision boundary more for individual data points. This parameter is only used for SVMs with RBF kernel. In Figure 3.1 is shown how the decision boundary changes in a 2-dimensional space when adjusting C for a linear kernel or γ for an RBF kernel. Hsu, Chang, Lin, et al. [59] propose a method where a "better" region in the search space is found after searching in a large space. A logarithmic scale with tenfold increments between 10^{-5} and 10^{5} for both C and γ is chosen for the first training iterations. Hereafter, a finer search is conducted in the area that appeared to contain the best fit. Random sampling of 250 arbitrary parameter settings was performed during training. Experiments of Bergstra and Bengio [60] showed that random sampling is more efficient than grid search for hyper-parameter tuning, since not all hyper-parameters are equally important to tune. In grid search a model is build and evaluated for every combination of hyper-parameters, which makes it a computationally expensive method when over two or three hyper-parameters are to be tuned, as is the case with SVM. Furthermore, we used 5-fold cross-validation to enable the model of using the complete training dataset and prevent from overfitting during training [59]. After the first training iterations, the range is narrowed down to 10^{-3} – 10^{1} for C and 10^{-4} – 10^{-1} for γ. The best performing model is

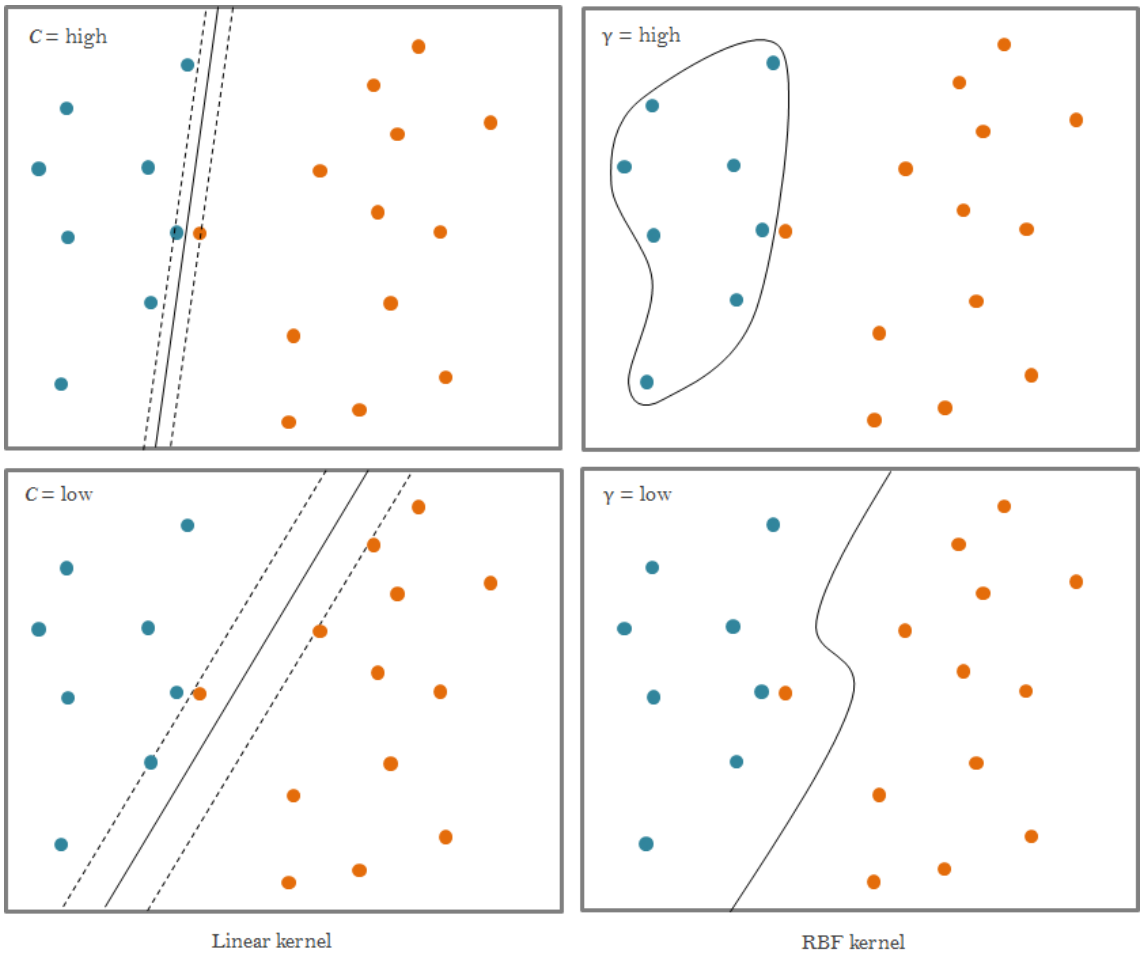

FIGURE 3.1: **Support vector machine parameter influences for binary classification.** On the left side is shown that a high value for C will classify more training points correctly, but uses a small margin. On the right side is shown that a high value for γ will classify more training points correctly, but depends greatly on the position of a single data point

tested against the testset for which model performance was evaluated. As recommended by Halevy, Norvig, and Pereira [61], to obtain a more generalizable model, the entire dataset (test and train) is trained again to generate the final model.

3.3 Model performance

In order to evaluate the performance of the SVM model, besides AUC, the performance metrics accuracy, precision, recall and F_1 score are used as well. These metrics are commonly used as metrics for binary text classification problems [18, 62, 63]. A good model performs well for all these metrics.

Accuracy is defined as the proportion of correctly predicted documents within the total number of documents. We define precision as the fraction of documents correctly predicted with label '1' within the total number of documents predicted with label '1'. Recall is defined as the fraction of documents correctly predicted with label '1' within the total number of documents that actually have label '1'. The F_1 score combines precision and recall, where a high F_1 score means that there are few documents wrongly predicted with label '1' and also

few documents wrongly predicted with label '0'. The score is defined as follows:

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall}$$

We ran a stratified dummy model from the `sklearn` package to evaluate the performance of the SVM model against. This model takes the class distributions of the training dataset as baseline and predicts the label from documents in the testset according to this distribution. McNemar's test is applied to test whether the error rate of the SVM model is significantly lower than the error rate of the dummy model [64].

Since model performance can only be tested for documents written in the first days of admission or in the last days before discharge, a different method is needed to evaluate the documents written in the days between. As a final step, the best performing model is applied to the complete dataset, as explained in Section 2.3. For each document the label is predicted with a certain probability. A line chart is created for each admission with on the x-axis the bin number and on the y-axis the probability that a document has label '1'. Since length of stay varies between admissions, the range of the x-axis differs as well. Therefore, we compare admissions with an equal number of documents.

3.4 Explainability

While SVM models have shown to be very effective in various domains in the past, they are generally unable to explain how they reached their conclusions. Especially in the medical domain a model's lack of explainability becomes problematic when it is designed for decision-making. Firstly, recent developments in regulation have made it more difficult to implement a black-box model in practice if they can not explain themselves [65]. According to the European Union's General Data Protection Regulation (GDPR) the end-user has the right to obtain clarification on a model's decision, to allow them to take their own decision on whether the model is trustworthy or not. Secondly, when introducing machine learning models, biases are likely to be built in as well [66]. These biases can more easily be found and fixed when the model is able to clarify its decisions. For instance, we might have missed out on some words that directly indicate the start or end of admission.

Holzinger et al. [67] describe two types of explainable systems, ante-hoc and post-hoc systems. Ante-hoc systems are the so-called glass-box algorithms that have explainability integrated in their design. Post-hoc systems explain the individual decisions of a model without explaining the internal structure of the model as a whole. Post-hoc systems have the ability to explain the decisions of a black box model and can thus be applied to the trained SVM model at hand.

In order to explain the decisions made by the SVM model in a human-friendly way, the post-hoc system LIME (Local Interpretable Model-agnostic Explanations)[3] is applied [68]. LIME creates an interpretable linear model for an instance to be explained. Therefore, it tries to approximate the original SVM model as close as possible while using less features. Since LIME explains only one instance at a time, it has the capability of drawing a local linear model. LIME explanations are easy to interpret, which makes them suitable in applications used by people with limited time, as is the case in a hospital [69].

[3] Further explanations of LIME are provided by Ribeiro, Singh, and Guestrin [68]

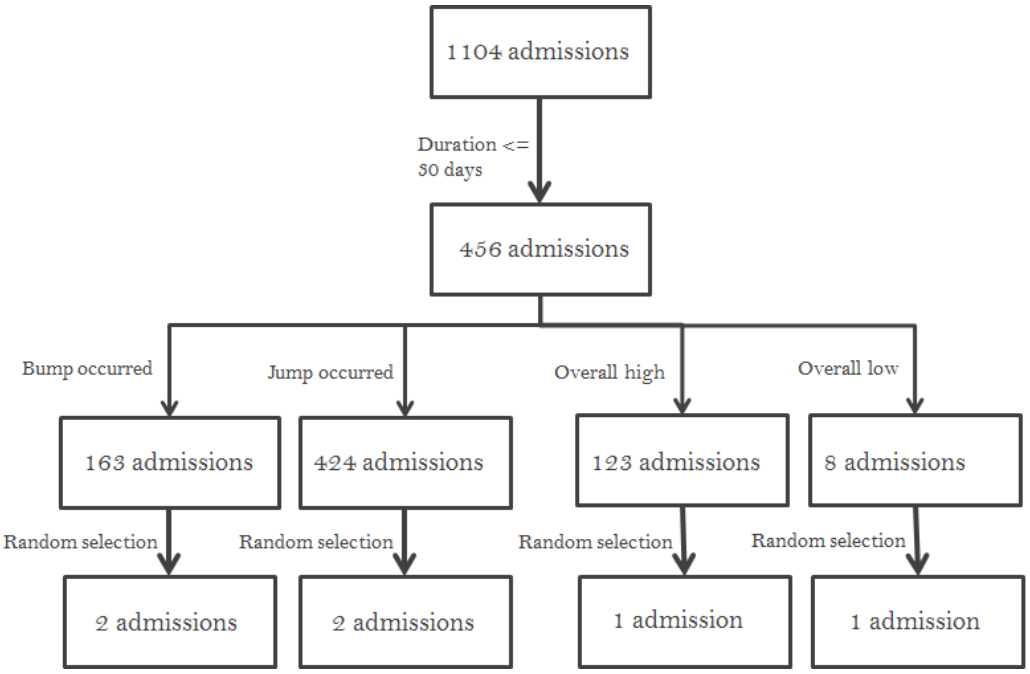

FIGURE 3.2: **Selection procedure of admission documents.** A bump is defined as the occurrence of a probability reduction of >= 30% between two consecutive documents written during admission. A jump is defined as the occurrence of a probability gain of >= 30% between two consecutive documents written during admission.

In the current study six complete admission reports (all documents written during an admission) are selected and LIME explanations are applied to the documents within these admission reports. Six annotators working within the Psychiatry Department of the UMCU have evaluated the admission reports. Basic knowledge of psychiatry and the data is required for participation. Annotators are asked to point out the moments during admission where they felt like the patient was improving or deteriorating and to underline the words that drew them to that particular conclusion. Each annotator evaluated two admission reports, therefore each admission report is reviewed double. For each admission report a first and a second annotator was designated. The first annotator wrote the words and general conclusions on a separate paper, while the second annotator underlined important words in the report and wrote general conclusions in the sideline.

The selection procedure of the admission reports can be found in Figure 3.2. As a first selection criteria we only selected admissions with a duration of 30 days or less. Admissions with a higher duration are considered too long for the annotator to keep focused. This resulted in a total of 456 admissions. In order to obtain a set of admissions that varies regarding the course of admission, we came up with some additional criteria. We selected two random admissions from a group of admissions where a *bump* occurred (Figure 3.3) and two random admission from a group of admissions where a *jump* occurred (Figure 3.4). Here a *bump* is defined as a probability reduction of >= 30% between two consecutive documents, which could indicate patient impairment. On the other hand, a *jump* is defined as a probability gain of >= 30% between two consecutive documents, which could indicate patient improvement. An extra criteria here was that the first document should have been classified with label '0' and the second with label '1'. For the admissions where a *bump* occurred this

had to be the other way around. Furthermore, we selected a random admission from a group of admissions that had all documents classified with label '1' and a group of admissions that had all documents classified with label '0' (Figure 3.5). Since the classifier is trained on the first and last document, we excluded these documents for the selection.

For the comparison of LIME explanations and human explanations we analyzed to what extend the *bumps* and *jumps* found by the model were found by annotators as well and which words lead to this decision. We calculated the percentage of words that were found by both the model and at least one annotator of the total of words the model indicated as important. Since annotators often indicate phrases as important instead of a single word, a word is included when it is part of a phrase (e.g. LIME: 'hoort' (*hears*); annotator: 'hoort stemmen' *hears voices*). When the model indicates two words of a phrase indicated by the annotator, both words are included (e.g. LIME: 'hoort' (*hears*) 'stemmen' (*voices*)).

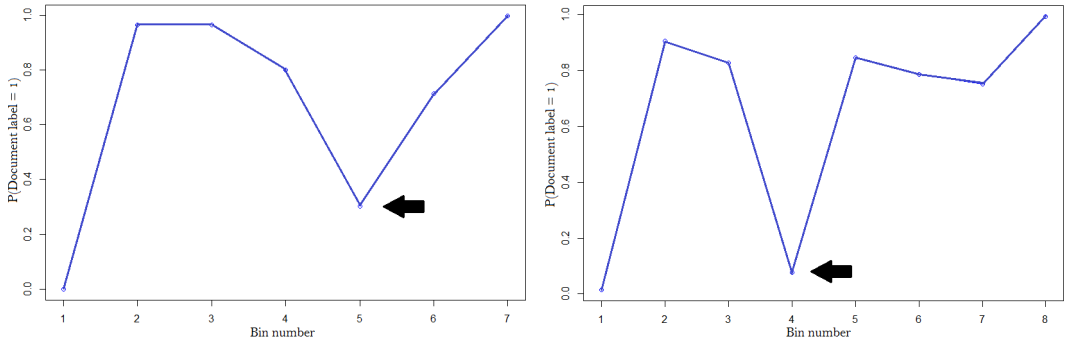

FIGURE 3.3: **Two admissions where a bump occurred.** A bump is defined as the occurrence of a probability (y-axis) reduction of >= 0.3 between two consecutive documents and is marked with a black arrow. The bin number on the x-axis represents a set of documents written over three days, where bin number '1' stands for the first three days, bin number '2' for the second three days, and so forth.

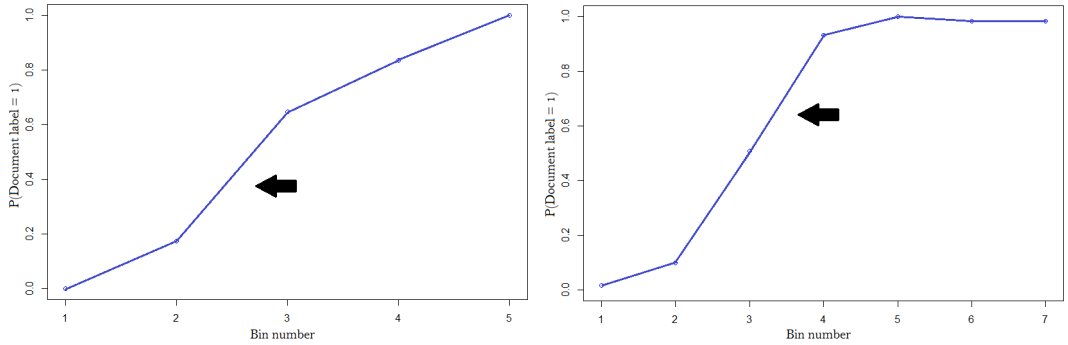

FIGURE 3.4: **Two admissions where a jump occurred.** A jump is defined as the occurrence of a probability (y-axis) gain of >= 0.3 between two consecutive documents and is marked with a black arrow. The bin number on the x-axis represents a set of documents written over three days, where bin number '1' stands for the first three days, bin number '2' for the second three days, and so forth.

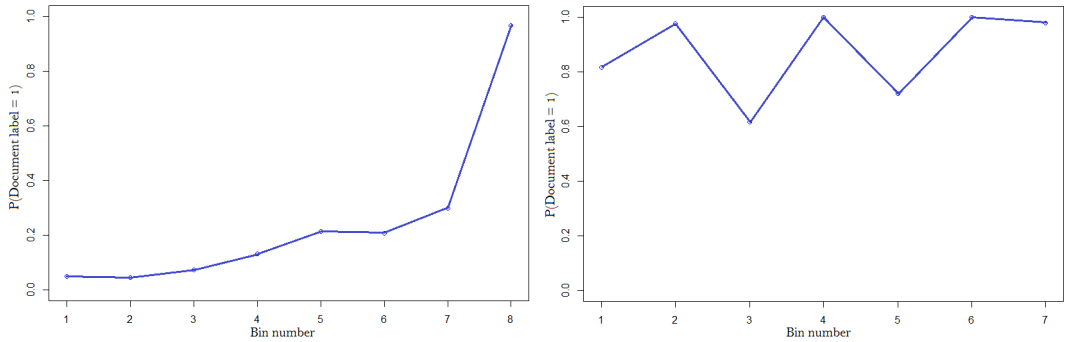

FIGURE 3.5: **Two stable admissions.** The admission on the left side has a overall low probability (P < 0.5) and the admission on the right side has an overall high probability (P > 0.5)The bin number on the x-axis represents a set of documents written over three days, where bin number '1' stands for the first three days, bin number '2' for the second three days, and so forth.

4 Results

In this section the results of the best performing SVM model are outlined. Furthermore, the main findings after the qualitative evaluation of the LIME explanations are described.

4.1 Model performance

The best performing SVM model had a RBF kernel and parameter settings C = 6.46748 and γ = 0.00387. In Table 4.1 the performance of both the SVM and the dummy classifier are shown for the different metrics. It was found that the SVM model significantly outperformed the dummy classifier on McNemar's test ($p < .001$). For each performance metric the results of the dummy classifier were around 0.5 (range between 0.49 - 0.52). For the SVM model an accuracy of 0.92 is found, implying that 92% of the documents in the test set are correctly predicted. The AUC of the SVM model was 0.97, implying that the expectation that a uniformly drawn document with label '1' is ranked before a uniformly drawn document with label '0' is 97%. A precision of 0.94 suggests that from all documents predicted to have label '1', 94% actually had label '1'. Subsequently, a recall of 0.91 implies that from all documents with label '1', 91% are correctly predicted to have label '1'. Since neither precision or recall shows bad results, the F_1 score had to show a good result. A score of 0.92 means that there is a good balance between precision and recall.

TABLE 4.1: **Results of the SVM[1] model and dummy classifier**

Classifier	**Accuracy**	**AUC[2]**	**Precision**	**Recall**	**F1**
SVM[1]	0.92	0.97	0.94	0.91	0.92
Dummy	0.51	0.51	0.49	0.51	0.52

[1]Support Vector Machine; [2]Area Under the Curve of the receiver operating characteristics curve

In Figure 4.1 the means and standard deviations are shown for patients with respectively seven, nine and twenty documents written during admission. The models for seven and nine documents, show a decelerating curve resembling the curve in Figure 1.1.

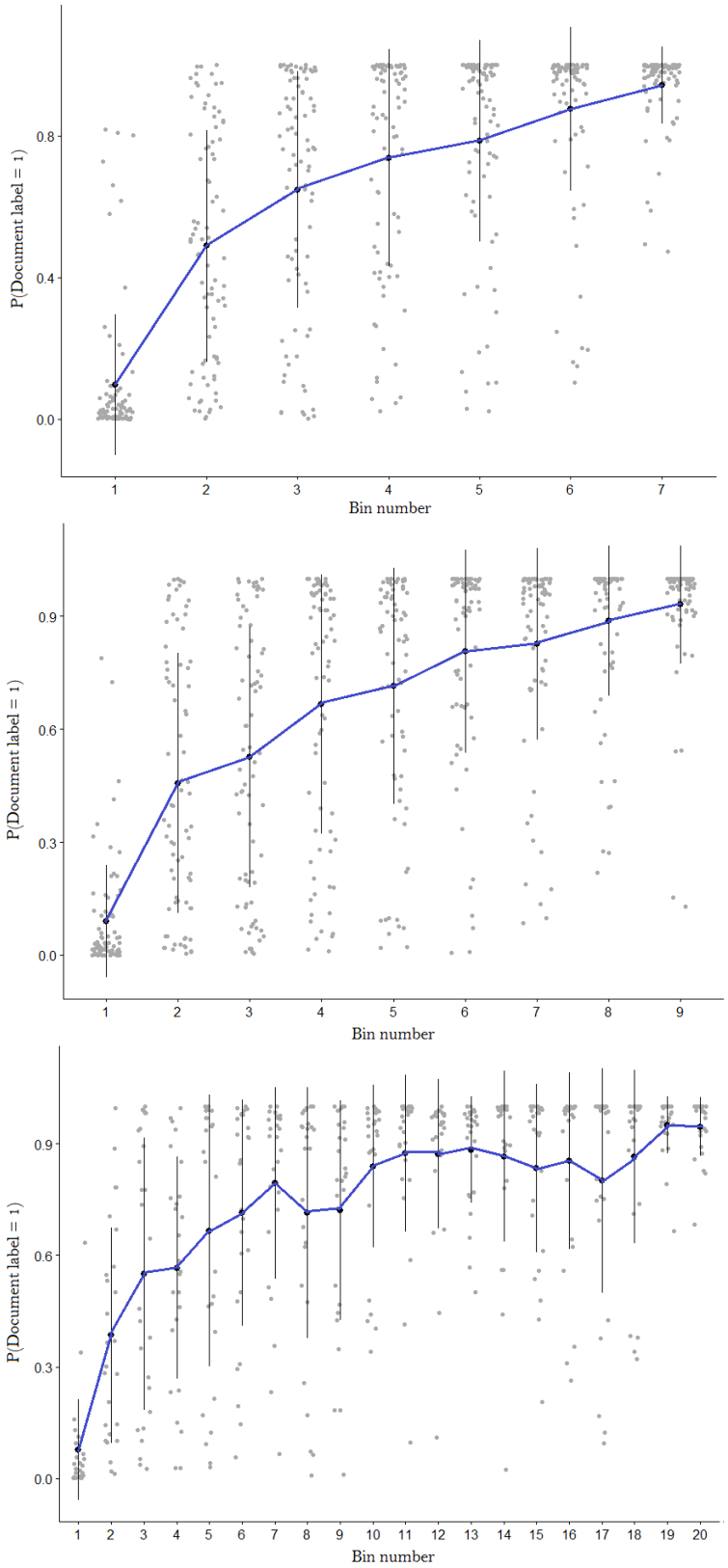

FIGURE 4.1: **Admissions for patients with seven (topmost), nine (center) and twenty (bottom) documents written during admission.** The blue line represents the mean for each bin. The grey dots show the probabilities (y-axis) for individual admissions. The bin number on the x-axis represents a set of documents written over three days, where bin number '1' stands for the first three days, bin number '2' for the second three days, and so forth.

4.2 Explainability

Annotators evaluated two admission reports where a *bump* was found. In both these admission reports the model found one *bump*. Three of the four annotators found the *bumps*, the fourth annotator argued that "the patient didn't seem to be really focused on his admission as a result of which the admission was not going well". Annotators who did find the *bump* found other moments they thought the patient was deteriorating as well.

Two admission reports are evaluated where a *jump* was found. In both these admission reports the model found one *jump*. One annotator found the *jump*, another annotator found the *jump* just before the model did and the third annotator found the *jump* just after the model did. The last annotator mentioned that 'the admission remained stable and that there were no moments standing out for improvement or impairment'.

From the annotators who evaluated the admission reports that were overall stable high or low, one annotator mentioned not to find any changes either. The other annotators found some moments of improvement or impairment, however, for the greater part of the admission, annotators and the model agreed upon each other.

In Table 4.2 the word count for both negative and positive words found by the model and by the first and second annotator are shown. It is found that the second annotator underlined the most words and the first annotator the least. In addition, the number of words is counted that is seen as positive or negative by both an annotator and the model, by both annotators or by both annotators and the model. The results can be found in Table 4.3. When adding up the numbers in Table 4.3, we found that the model makes predictions based on words that are indicated by annotators for 14.8% of the words. It is found that most overlap in words are indicated by the model and the second annotator. Negative words that were found by both annotators and the model were: 'Scheiding' (*divorce*) and 'Contact' (*contact*), where 'Scheiding' was mentioned two times in different documents. The positive word that is found by both annotators and the model was: 'Rustig' (*calm*). Overall, we found a high variation in the number of words indicated by annotators as either positive or negative. The first annotator often provided only a global thought about the course of admission while the second annotator pointed out multiple words per nurse report. Furthermore, it is shown that annotators often underlined multiple words together as indication for patient improvement or impairment, while the LIME explanations provide just one word (e.g. LIME: 'hoort' (*hears*); annotator: 'hoort stemmen' *hears voices*).

TABLE 4.2: **Words counts for positive and negative words.** Annotator 1 wrote important words on a separate paper. Annotator 2 underlined important words in the report.

Negative words

Model	Annotator 1	Annotator 2
219	42	309

Positive words

Model	Annotator 1	Annotator 2
201	40	241

TABLE 4.3: **Word agreements for positive and negative words.** The number of words is counted exclusively, hence word counts for Model and Annotator 1 do not include words indicated by Annotator 2 as well.

Negative words

Model & Annotator 1 [1]	Model & Annotator 2	Annotator 1 & Annotator 2	All
3	34	22	3

Positive words

Model & Annotator 1	Model & Annotator 2	Annotator 1 & Annotator 2	All
0	21	14	1

5 Discussion

Outcome measures in psychiatry are often obtained only at the start and end of a hospital admission. Information regarding the patient's status in the remaining time can be useful for making discharge decisions, to reflect on a specific treatment and to compare with other trajectories of patients with a similar history. This data can be retrieved from clinical notes. However, a proper method for obtaining an outcome measure from these notes is needed. In this study we proposed a machine learning model for obtaining such an outcome measure. We trained a text classification model on nurse reports written at the beginning and end of a hospital admission. In order to attain a measurement of patient well-being we made the assumption that patients suffer from serious mental problems at the beginning of admission and that their symptoms are reduced or stabilized when they are discharged, thereby linking time of the report to mental state of well-being. Subsequently, nurse reports written over the course of admission were fed to the model. The outcome measure was defined as the probability that a document was written in the last days before discharge.

The SVM classifier predicted whether a document was written in the first days of admission or in the last days before discharge with a high accuracy of 92 percent and thereby outperformed the dummy classifier. The model was applied to documents written over the course of admission without knowing the true probability rates. For admissions containing seven, nine and twenty documents respectively, a line chart was drawn indicating the mean course of a psychiatric admission in terms of improvement. For all three classes of admissions the line showed a decelerating curve, which is in line with findings in previous studies in psychotherapy. The similarity in mean trajectories found provides additional weight to the theory of the three-phase healing process of Howard et al. [27]. However, we cannot rule out the possibility that the similarity is found by coincidence. For instance, our model can be biased for the possibility that nurses write extensive summaries of the patient's situation and interests in the first days of admission and moderate this during the remainder of the admission. This might cause documents being classified to be written in the last days soon after the first days of admission. A similar decelerating curve can be drawn from this theory. Furthermore, when comparing different admission trajectories a lot of variance is found in course of admission, possibly indicating that the course of admission depends on the characteristics of the patient and their environment. If a model as the current one will be used to make predictions on the course of admission for a specific patient, we have to keep in mind that the mean course of admission is not very informative, as argued by Lutz et al. [25], since there are a lot of determining factors for patient improvement.

In order to evaluate the explainability of our model, we provided LIME explanations for the model and compared them to human explanations. Six admission reports were evaluated by two annotators each. It was found that the model makes predictions based on words that were not indicated by annotators for 85.2% of the words. An explanation for this finding

is that some annotators did not underline many words, but rather provided an indication of whether they thought a patient was improving or deteriorating. As a result, for each admission report we had one extensive human explanation summary and one limited human explanation summary. Only 14.8% of the words indicated as important by the model were found by annotators as well. Annotators provided a general conclusion on the course of admission as well. A remarkable finding is that these general conclusions matched the model's line chart for the greater part of the time. One of the main reasons for physicians not to adopt a clinical decision support system, is when they question the validity of the system [70]. Our current model has potential to find patterns of patient improvement and impairment, although it cannot provide an explanation that is good enough to be fully trusted by humans yet.

5.1 Limitations

Limitations in this study are found in the preprocessing of the data. For instance, we did not fully correct for spelling mistakes made. As stated in the introduction, nurses make spelling mistakes due to the lack of administration time. To control for these a spell corrector algorithm is needed, which is not freely available for Dutch medical text. In this study we used the word2vec algorithm as a first approach for this problem, since similar vectors will be created for words with similar meanings, as typos. Another preprocessing step that was missing is word negations. Banerjee et al. [71] used a Report Condenser to bind a negation with their dependencies (e.g. *'No_infarction'*). Such a negation searcher was not used by Menger, Scheepers, and Spruit [18] for creating the vocabulary, therefore in this study it could not be used either.

Another type of limitation was found in the patient population. The majority of patients who are admitted to the Psychiatry Department of the UMCU are suffering from severe mental illness. Some patients have chronic diseases, which means that they will never fully recover. When removing patients with chronic diseases from the database, there would not have been enough data left to create a working model. However, the current population might not be a well-suited population for the task. On the other hand, patients with chronic diseases have periods where they feel better or worse just like non-chronic patients. In addition, we want the model to be predictive for these patients as well.

The method that was used to evaluate the model's outcome and explanations should be revised. Since admission reports that were evaluated by human annotators covered seven pages on average, the time needed to review an admission report was between half an hour and an hour. Annotators who were not allowed to underline words within the report itself, but were asked to write those words on a separate paper, often got discouraged towards the end of the report. Therefore, we could not make a proper comparison between the explanations of Annotator 1 and those of Annotator 2 or the model. Furthermore, we had difficulties in comparing human reviews with the model's line chart directly. The model made predictions based on nurse reports written over three days and provided a probability rate, while annotators read nurse reports one by one, and provided an objective conclusion.

The biggest limitation is that we made the assumption that a patient's health status was bad at the beginning of admission and good or at least better at the end, but we do not know this for sure. It would be better if there was some way to check whether this is true, for instance by having patients filling out a form or letting nurses write a final conclusion.

5.2 Future research

In this study we aimed at answering the question 'Can assumed mental health improvement be predicted by text classification on nurse reports, where the model is trained solely on text written at the beginning and end of a hospital admission?'. Following the methods of the current study we did find some promising results. The model could predict whether documents were written at the start or end of admission with a very high accuracy. Furthermore, the mean line chart of probability rates followed the course of patient improvement found in the literature and indicated by human annotators. Unfortunately, for now, the model's explanations cannot be trusted enough. Nevertheless, since text classification on nurse reports with the purpose of finding patterns of improvement or deterioration has not been explored before, the door has been opened for further research in this area.

Besides SVM, neural networks (NN) have successfully been used for text classification tasks as well [72, 73]. NNs are more complex than SVMs, but can consequently be more flexible and efficient as well. The reason for this is that NNs use a reduced document size for text classification which can save computation time and costs.

Apart from testing other machine learning models, changes and improvements can be made to the word2vec model as well. The vocabulary used is build out of documents originating from the same source as the nurse reports that were fed to the model, therefore, the model is not very generalizable. A method that can be explored is transfer learning [74], which includes the reuse of an existing generalized model and tune that model for a new task. For instance, a well-performing model that is trained on a large medical dataset could be reused for the current classification task. The model could be enhanced for the current task by expanding the list of stop words and by adding extra weight to important words that were indicated by annotators.

Bibliography

[1] L. Fetters and J. Tilson. *Evidence based physical therapy*. Philadelphia: F.A. Davis Company, 2018, p. 163.

[2] E.C. Nelson et al. "Patient reported outcome measures in practice". In: *BMJ* (2015). [Online] Available from: https://www.bmj.com/content/350/bmj.g7818.long [Accessed 13 February 2019].

[3] J.F. Boswell et al. "The expanding relevance of routinely collected outcome data for mental health care decision making". In: *Administration and Policy in Mental Health and Mental Health Services Research* 43.4 (2016), pp. 482–491.

[4] D.R. Hatfield and B.M. Ogles. "The Use of Outcome Measures by Psychologists in Clinical Practice". In: *Professional Psychology: Research and Practice* 35.5 (2004), p. 485.

[5] C. Johnson. "Outcome measures for research and clinical practice". In: *Journal of Manipulative and Physiological Therapeutics* 31.5 (2008), pp. 329–330.

[6] H. Neuvirth et al. "Toward personalized care management of patients at risk: the diabetes case study". In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. San Diego, CA, USA, 2011, pp. 395–403.

[7] S.T. Rosenbloom et al. "Data from clinical notes: a perspective on the tension between structure and flexible documentation". In: *Journal of the American Medical Informatics Association* 18.2 (2011), pp. 181–186.

[8] M. Sajatovic et al. "Rating Scales for Psychiatric Disorders". In: *The Medical Basis of Psychiatry*. Ed. by Clayton P. Fatemi S. New York, NY: Springer, 2016, pp. 869–880.

[9] B.L. Handen et al. "A randomized, placebo-controlled trial of metformin for the treatment of overweight induced by antipsychotic medication in young people with autism spectrum disorder: open-label extension". In: *Journal of the American Academy of Child & Adolescent Psychiatry* 56.10 (2017), pp. 849–856.

[10] K.E. Roach. "Measurement of health outcomes: reliability, validity and responsiveness". In: *JPO: Journal of Prosthetics and Orthotics* 18.6 (2006), pp. 8–12.

[11] M.G.A. Opler, C. Yavorsky, and D.G. Daniel. "Positive and Negative Syndrome Scale (PANSS) Training: Challenges, Solutions, and Future Directions". In: *Innovations in clinical neuroscience* 14.11-12 (2017), p. 77.

[12] M.J. Müller and A. Dragicevic. "Standardized rater training for the Hamilton Depression Rating Scale (HAMD-17) in psychiatric novices". In: *Journal of affective disorders* 77.1 (2003), pp. 65–69.

[13] P.B. Jensen, L.J. Jensen, and S. Brunak. "Mining electronic health records: towards better research applications and clinical care". In: *Nature Reviews Genetics* 13.6 (2012), p. 395.

[14] H. Kharrazi et al. "The value of unstructured electronic health record data in geriatric syndrome case identification". In: *Journal of the American Geriatrics Society* 66.8 (2018), pp. 1499–1507.

[15] K.J. Cios and G.W. Moore. "Uniqueness of medical data mining". In: *Artificial intelligence in medicine* 26.1-2 (2002), pp. 1–24.

[16] D.T. Heinze, M.L. Morsch, and J. Holbrook. "Mining free-text medical records". In: *Proceedings of the AMIA Symposium*. American Medical Informatics Association. San Diego, CA, USA, 2001, p. 254.

[17] X. Zhou et al. "Approaches to text mining for clinical medical records". In: *Proceedings of the 2006 ACM symposium on Applied computing*. ACM. Dijon, France, 2006, pp. 235–239.

[18] V. Menger, F. Scheepers, and M. Spruit. "Comparing deep learning and classical machine learning approaches for predicting inpatient violence incidents from clinical text". In: *Applied Sciences* 8.6 (2018), p. 981.

[19] A.C. Page, N.K. Cunningham, and G.R. Hooke. "Using daily monitoring of psychiatric symptoms to evaluate hospital length of stay". In: *BJPsych open* 2.6 (2016), pp. 341–345.

[20] K. Dyer, G. Hooke, and A.C. Page. "Development and psychometrics of the five item daily index in a psychiatric sample". In: *Journal of affective disorders* 152 (2014), pp. 409–415.

[21] J.L. Donovan et al. "Patient-reported outcomes after monitoring, surgery, or radiotherapy for prostate cancer". In: *New England Journal of Medicine* 375.15 (2016), pp. 1425–1437.

[22] E. Basch et al. "Symptom monitoring with patient-reported outcomes during routine cancer treatment: a randomized controlled trial". In: *Journal of Clinical Oncology* 34.6 (2016), p. 557.

[23] M. Huber et al. "Towards a 'patient-centred' operationalisation of the new dynamic concept of health: a mixed methods study". In: *BMJ open* 6.1 (2016). [Online] Available from: `https://bmjopen.bmj.com/content/6/1/e010091.info` [Accessed 6 July 2019].

[24] W. Lutz et al. "Prediction of dose–response relations based on patient characteristics". In: *Journal of Clinical Psychology* 57.7 (2001), pp. 889–900.

[25] W. Lutz, Z. Martinovich, and K.I. Howard. "Patient profiling: An application of random coefficient regression models to depicting the response of a patient to outpatient psychotherapy." In: *Journal of consulting and clinical psychology* 67.4 (1999), p. 571.

[26] K.I. Howard, R.J. Lueger, and G.G. Kolden. "Measuring progress and outcome in the treatment of affective disorders". In: *Measuring patient change after treatment for mood, anxiety, and personality disorders: Toward a core battery*. Ed. by L.M. Horowitz, M.J. Lambert, and H.H. Strupp. Washington, DC: American Psychological Association, 1997, pp. 263–281.

[27] K.I. Howard et al. "The dose–effect relationship in psychotherapy." In: *American psychologist* 41.2 (1986), p. 159.

[28] International Breast Cancer Study Group (IBCSG). "Endocrine responsiveness and tailoring adjuvant therapy for postmenopausal lymph node-negative breast cancer: a randomized trial". In: *Journal of the National Cancer Institute* 94.14 (2002), pp. 1054–1065.

[29] C. Sabine, Antiretroviral Therapy (ART) Cohort Collaboration, et al. "AIDS events among individuals initiating HAART: do some patients experience a greater benefit from HAART than others?" In: *AIDS* 19.17 (2005), pp. 1995–2000.

[30] T. Botsis et al. "Secondary use of EHR: data quality issues and informatics opportunities". In: *Summit on Translational Bioinformatics* 2010 (2010), pp. 1–5.

[31] T.B. Murdoch and A.S. Detsky. "The inevitable application of big data to health care". In: *JAMA* 309.13 (2013), pp. 1351–1352.

[32] X. Liu et al. "Privacy-preserving patient-centric clinical decision support system on naive Bayesian classification". In: *IEEE journal of biomedical and health informatics* 20.2 (2016), pp. 655–668.

[33] S.E. Dilsizian and E.L. Siegel. "Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment". In: *Current cardiology reports* 16.1 (2014), p. 441.

[34] Y. Xu et al. "Deep convolutional activation features for large scale brain tumor histopathology image classification and segmentation". In: *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. Brisbane, QLD, Australia, 2015, pp. 947–951.

[35] S.M. Meystre et al. "Extracting information from textual documents in the electronic health record: a review of recent research". In: *Yearbook of medical informatics* 17.1 (2008), pp. 128–144.

[36] A. Vatian et al. "Adaptation of Algorithms for Medical Information Retrieval for Working on Russian-Language Text Content". In: *Text, Speech, and Dialogue*. Ed. by P. Sojka et al. Cham, Switzerland: Springer, 2018, pp. 106–114.

[37] V. Menger et al. "DEDUCE: A pattern matching method for automatic de-identification of Dutch medical text". In: *Telematics and Informatics* 35.4 (2018), pp. 727–736.

[38] S.M. Burke. "Basics of Different Scripts". In: *The Perl Journal* (2001).

[39] S. Bird, E. Klein, and E. Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc., 2009.

[40] J. Yan. "Text Representation". In: *Encyclopedia of Database Systems*. Ed. by L. Liu and M.T. Özsu. Boston, MA: Springer US, 2009, pp. 3069–3072.

[41] B.S. Harish, D.S. Guru, and S. Manjunath. "Representation and classification of text documents: A brief review". In: *IJCA, Special Issue on RTIPPR (2)* (2010), pp. 110–119.

[42] W. Liu and B. Vinter. "CSR5: An efficient storage format for cross-platform sparse matrix-vector multiplication". In: *Proceedings of the 29th ACM on International Conference on Supercomputing*. ACM. 2015, pp. 339–350.

[43] Q. Le and T. Mikolov. "Distributed representations of sentences and documents". In: *International conference on machine learning*. Google Inc. Mountain View, CA, 2014, pp. 1188–1196.

[44] T. Mikolov et al. "Efficient estimation of word representations in vector space". In: *International Conference on Learning Representations: Workshops Track* (2013).

[45] J. Lilleberg, Y. Zhu, and Y. Zhang. "Support vector machines and word2vec for text classification with semantic features". In: *2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)*. IEEE. Beijing, China, 2015, pp. 136–140.

[46] J. Pennington, R. Socher, and C. Manning. "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. Association for Computational Linguistics. Doha, Qatar, 2014, pp. 1532–1543.

[47] P.K. Sarma, Y. Liang, and W.A. Sethares. "Domain Adapted Word Embeddings for Improved Sentiment Classification". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics. Melbourne, Australia, 2018, pp. 51–59.

[48] R. Caruana and A. Niculescu-Mizil. "An empirical comparison of supervised learning algorithms". In: *Proceedings of the 23rd international conference on Machine learning*. ACM. Pittsburgh, Pennsylvania, USA, 2006, pp. 161–168.

[49] D.H. Wolpert, W.G. Macready, et al. "No free lunch theorems for optimization". In: *IEEE transactions on evolutionary computation* 1.1 (1997), pp. 67–82.

[50] Scikit Learn developers. *Choosing the Right Estimator*. [Online] Available from: `http://scikit-learn.org/stable/tutorial/machine_learning_map/index.html` [Accessed 24 May 2019]. 2019.

[51] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[52] T. Joachims. "Text Categorization with Support Vector Machines: Learning with Many Relevant Features". In: *Proceedings of the 10th European Conference on Machine Learning*. Berlin, Heidelberg: Springer-Verlag, 1998, pp. 137–142.

[53] W. Weng et al. "Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach". In: *BMC medical informatics and decision making* 17.1 (2017), p. 155.

[54] L. McKnight and P. Srinivasan. "Categorization of sentence types in medical abstracts". In: *AMIA Annual Symposium Proceedings*. Vol. 2003. American Medical Informatics Association. Washington, DC, 2003, p. 440.

[55] J.J.G. Adeva et al. "Automatic text classification to support systematic reviews in medicine". In: *Expert Systems with Applications* 41.4 (2014), pp. 1498–1508.

[56] R. Swaminathan, A. Sharma, and H. Yang. "Opinion mining for biomedical text data: Feature space design and feature selection". In: *The Nineth International Workshop on Data Mining in Bioinformatics, BIOKDD*. 2010.

[57] M. Ong, F. Magrabi, and E. Coiera. "Automated categorisation of clinical incident reports using statistical text classification". In: *Qual Saf Health Care* 19.6 (2010), e55.

[58] A.K.S. Wong, J.W.T. Lee, and D.S. Yeung. "Improving text classifier performance based on AUC". In: *18th International Conference on Pattern Recognition (ICPR'06)*. Vol. 3. IEEE. Hong Kong, China, 2006, pp. 268–271.

[59] C. Hsu, C. Chang, C. Lin, et al. *A practical guide to support vector classification*. Tech. rep. Taipei 106, Taiwan: Department of Computer Science, National Taiwan University, 2003.

[60] J. Bergstra and Y. Bengio. "Random search for hyper-parameter optimization". In: *Journal of Machine Learning Research* 13.2 (2012), pp. 281–305.

[61] A. Halevy, P. Norvig, and F. Pereira. "The unreasonable effectiveness of data". In: *Intelligent Systems, IEEE* 24.2 (2009), pp. 8–12.

[62] A. Sun, E. Lim, and Y. Liu. "On strategies for imbalanced text classification using SVM: A comparative study". In: *Decision Support Systems* 48.1 (2009), pp. 191–201.

[63] M. Ikonomakis, S. Kotsiantis, and V. Tampakas. "Text classification using machine learning techniques." In: *WSEAS transactions on computers* 4.8 (2005), pp. 966–974.

[64] E. McCrum-Gardner. "Which is the correct statistical test to use?" In: *British Journal of Oral and Maxillofacial Surgery* 46.1 (2008), pp. 38–41.

[65] B. Goodman and S. Flaxman. "European Union regulations on algorithmic decision-making and a "right to explanation"". In: *AI Magazine* 38.3 (2017), pp. 50–57.

[66] B. Resnick. *Genetics has learned a ton — mostly about white people. That's a problem. Vox*. [Online] Available from: https://www.vox.com/science-and-health/2018/10/22/17983568/dna-tests-precision-medicine-genetics-gwas-diversity-all-of-us [Accessed 13 June 2019]. 2018.

[67] A. Holzinger et al. "What do we need to build explainable AI systems for the medical domain?" In: *arXiv preprint arXiv:1712.09923* (2017).

[68] M.T. Ribeiro, S. Singh, and C. Guestrin. "Why should i trust you?: Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM. San Francisco, California, USA, 2016, pp. 1135–1144.

[69] C. Molnar et al. "Interpretable machine learning: A guide for making black box models explainable". In: *Christoph Molnar, Leanpub* (2019). [Online] Available from: https://christophm.github.io/interpretable-ml-book/ [Accessed 3 April 2019].

[70] S. Khairat et al. "Reasons for physicians not adopting clinical decision support systems: critical analysis". In: *JMIR medical informatics* 6.2 (2018), e24.

[71] I. Banerjee et al. "Intelligent word embeddings of free-text radiology reports". In: *AMIA Annual Symposium Proceedings*. Vol. 2017. American Medical Informatics Association. 2017, p. 411.

[72] S. Lai et al. "Recurrent convolutional neural networks for text classification". In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. AAAI. Austin, Texas, 2015, pp. 2267–2273.

[73] A. Conneau et al. "Very deep convolutional networks for text classification". In: *arXiv preprint arXiv:1606.01781* (2016).

[74] W. Pan, E. Zhong, and Q. Yang. "Transfer learning for text mining". In: *Mining Text Data*. Ed. by C.C. Aggarwal and C.X. Zhai. Boston, LA: Springer, 2012, pp. 223–257.