Detection of Hospital Acquired Infections in sparse and noisy Swedish patient records

A machine learning approach using Naïve Bayes, Support Vector Machines and C4.5

Claudia Ehrentraut
Dep. of Linguistics and Philology,
Uppsala University
P.O. Box 635
Uppsala, Sweden
Claudia.Ehrentraut.6386@student.uu.se

Hideyuki Tanushi
Dep. of Computer and Systems Sciences (DSV),
Stockholm University
Forum 100
Kista, Sweden
hide-tan@dsv.su.se

Hercules Dalianis
Dep. of Computer and Systems Sciences (DSV),
Stockholm University
Forum 100
Kista, Sweden
hercules@dsv.su.se

Jörg Tiedemann
Dep. of Linguistics and Philology,
Uppsala University
P.O. Box 635
Uppsala, Sweden
jorg.tiedemann@lingfil.uu.se

ABSTRACT

Hospital Acquired Infections (HAI) pose a significant risk on patients' health while their surveillance is an additional work load for hospital medical staff and hospital management. Our overall aim is to build a system which reliably retrieves all patient records which potentially include HAI, to reduce the burden of manually checking patient records by the hospital staff. In other words, we emphasize recall when detecting HAI (aiming at 100%) with the highest precision possible. The present study is of experimental nature, focusing on the application of Naïve Bayes (NB), Support Vector Machines (SVM) and a C4.5 Decision Tree to the problem and the evaluation of the efficiency of this approach. The three classifiers showed an overall similar performance. SVM yielded the best recall value, 89.8%, for records that contain HAI. We present a machine learning approach as an alternative to rule-based systems which are more common in this task. The classifiers were applied on a small and noisy dataset, generating results which pinpoint the potentials of using learning algorithms for detecting HAI. Further research will have to focus on optimizing the performance of the classifiers and to test them on larger datasets.

AND '12 Mumbai, India

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning—*Induction*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Text analysis*; I.5.1 [**Pattern Recognition**]: Models—*Statistical*; J.3 [**Life and medical sciences**]: Health

General Terms

Algorithms, Experimentation, Theory

Keywords

Hospital Acquired Infection, Noisy data, Swedish patient records, Naïve Bayes, Support Vector Machines, C4.5 Decision Tree

1. INTRODUCTION

Ten percent of all in-patients suffer from Hospital Acquired Infections (HAI) [13]. In Europe, this estimates to three million affected patients per year of which about 50,000 die. HAI is defined as "[a]n infection occurring in a patient in a hospital or other health care facility in whom the infection was not present or incubating at the time of admission. This includes infections acquired in the hospital but appearing after discharge, and also occupational infections among staff of the facility" by [9]. HAI can for example be caused by the use of catheter (during medical treatment), develop in wounds after surgery, pneumonia obtained during medical treatment, by spreading of norovirus as in the winter vomiting disease, etc. By law, the health-care process must be documented in patient records, but can also be regarded as a support and mnemonic for health-care professional. Computerization of patient records have made it much easier to

document health-care processes and monitor how the treatment of patients is proceeding.

Many approaches to detect HAI using such electronic patient records have been carried out so far. The majority of them either uses the unstructured or structured information of those records in order to build first and foremost rule-based systems. In the context of patient records, unstructured or free-text data refers to for instance discharge summaries[1]. By contrast, structured data is associated with data retrieved from hospital databases, such as microbiological data or temperatures. To our knowledge, however, only a few machine learning approaches have been carried out (see section 2.2).

Our study is of experimental nature, complementing a rule-based system for detecting HAI from Swedish patient records which is currently developed in a collaborative project between Karolinska University Hospital and the Department of Computer and System Science (DSV) at Stockholm University. For our task of applying machine learning techniques to patient records, of which we used the unstructured and structured part, we chose three well known learning algorithms: Naïve Bayes (NB), Support Vector Machine (SVM) and the C4.5 decision tree. The focus of our study lies on the recall values obtained by the different classifiers. By means of preprocessing the data and performing feature selection, we try different settings in order to increase recall. This serves our overall aim of approaching 100% recall with the highest precision possible. Algorithms with high recall are, according to [16], suitable for the screening of infections. Thus, this study is an important step towards implementing a system which is expected to constantly screen patient records, determining whether they contain HAI or not. We will focus on answering the question (1) which recall is possible for our approach and (2) does any of the classifier outperform another, making it (them) more applicable for our task. Automatic HAI screening is especially valuable for medical staff and hospital management, since it would significantly reduce the burden for physicians or nurses to manually check patient records for HAI. Instead of analyzing all, they would only have to check those patient records which were preselected by the system to either contain HAI or not.

Handling patient records may appear problematic since they can be considered noisy. Just like texts which are considered to be noisy, they contain misspellings, non-standard abbreviations, acronyms or domain specific terms [1]. Rule-based system easily become complex since they have to handle such 'errors', mostly by removing or resolving them, whereas machine learning (ML) is expected to handle the data by inducing knowledge from the data despite the underlying noise [29]. Besides assigning a greater objectivity to machine learning systems in contrast to rule-based systems, this is yet another fact which favors the approach of applying ML to HAI-detection.

The remaining parts of the paper are organized as follows: In section 2 we motivate why we consider patient records to be noisy before surveying related work on HAI detection. In section 3 we present the data used. Section 4 briefly describes the choice of machine learning algorithms as well as the preprocessing techniques deployed. In section 5 we will present the results. Section 6 finalizes the paper by discussing the results.

[1]Alternatively referred to as discharge diagnoses by for instance [20].

2. BACKGROUND

2.1 Noisy data

Online chat, SMS, email, blogs, wikis, web pages and historical texts are some examples of noisy texts which are given by [29] and [17]. More formally, [17] define "noise in text as any kind of difference between the surface form of a coded representation of the text and the intended, correct, or original text." In accordance with this definition, the authors regard for example spelling errors, abbreviations, non-standard terminology or missing punctuation as noise. The kind and frequency of noise varies thereby depending on which sort of text one refers to. See [29] for a more detailed apportionment.

Looking at [23] who present papers that reflect the current state-of-the-art in noisy text analytics, none of the papers cited study patient records in this regard. Based on the definitions and examples given by [17] we came to the conclusion to consider patient records as noisy. The remainder of this section focuses therefore on substantiating our assumption, with the overall aim of pointing out patient records as noisy to other researchers.

2.1.1 Spelling errors

Spelling errors are listed as the first example of noise by [23]. In the analysis of spelling errors, newspaper or journal texts are often considered as main reference objects [22], [27]. The authors in [22] for instance expect web texts to contain a significantly larger number of misspellings or typographical errors than newspaper articles, 0.8%[2] compared to 0.44%[3], something they motivate based on the following (1) authors of web texts differ substantially in their command of English, and (2) web pages are not as carefully edited as newspaper articles.

The latter does even apply to patient records. Those are, as reported by [27], primarily for hospital internal use, e.g., as a support and mnemonic for health-care professionals. This can be explained by the fact that, on behalf of the medical staff, there is neither the demand nor time to edit or correct eventual mistakes. Accordingly, [27] find spelling errors to be much more frequent in patient records than in other corpora, i.e., around 10%. Even [1], who conducted a study where they comprehensively analyzed Finnish and Swedish clinical texts mention the "numerous linguistic and grammatical mistakes made". In addition to [22], the author in [18] refers to a number of studies where spelling errors have been observed. She clearly states that spelling errors in text depend on various aspects, for example the text entry mode, e.g., if the text was handwritten, a typed textual conversation or written on a computer and automatically edited. According to this distinction she cites error rates of 1.5% to 2.5% for handwritten texts, 5% to 6% for typed textual conversation, 0.2% for texts written on the computer using a text-editing program and 0.05% for Associated Press news wire text. Table 1 gives an overview of the percentage of misspellings in the various kinds of texts cited by [22] as well as [18] and patient records.

[2]Based on Web corpus designed by [22].

[3]Based on the Gigaword corpus. `http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2011T07`.

Table 1: Spelling error rates (in %) in various texts.

	Misspellings
Texts written in e.g. Word	0.2
Newspaper texts	0.05 - 0.44
Web texts	0.8
Handwritten texts	1.5 - 2.5
Typed textual conversations	5.0 - 6.0
Patient records	10.0

2.1.2 Abbreviations

Secondly, [23] list abbreviations as an indication of noise. Even in this respect we could find a high correlation of the number of abbreviations in patient records and in for instance SMS. The author in [14] found that 10.6% of all tokens in his dataset[4] were abbreviations. Studies by [1] and [28] confirm the prevalent occurrence of abbreviations and even acronyms in clinical texts. Abbreviations used in Swedish patient records range from *rtg* - röntgen (Engl.: x-ray), *ul* - ultraljud (Engl.: ultrasound) or underläkare (Engl.: assistant physician) to *p5* - pertrokantär femurfraktur (Engl.: pertrochanteric femur fracture). In contrast, the author in [29] found that SMS contain around 5% abbreviations, i.e., not even half as much as in patient records. Table 2 depicts these differences.

Table 2: Abbreviations (in %) in SMS and patient records.

	Abbreviations
SMS	5.0
Patient records	10.6

2.1.3 Non-standard terminology

Further, a non-standard terminology is listed as an indication of noise by [23]. Even in this regard do patient records apply to the definition of noisy texts. In their study, [1] pinpoint the vast amount of domain specific terms which occur in medical texts. They exemplify abbreviations and terms which are specific to Intensive Care Unit nursing narratives. When analyzing those narratives, the authors found that some terms were already unclear to medical staff with less domain expertise, making them hard to understand if not unintelligible for patients.

Since patient records contain 10% spelling error, around 10% abbreviations as well as a non-standard terminology, we can consider that they fulfill the conditions to be noisy, as stated in [17]. As the author in [29] states, most systems are trained on clean text. This implies that, whenever noise occurs in unseen data which the systems are tested upon, performance can be adversely affected. Thus, the author concludes that noise needs to be explicitly handled by either removing noise or making the system more robust to it. While the former especially applies to rule-based systems, the challenge for machine learning approaches is to learn the underlying models even in the presence of noise [17].

[4]300 medical records provided by Karolinska University Hospital, Stockholm.

2.1.4 Noisy patient records

A study by [20] is the only paper that was found in which clinical data is referred to as noisy, containing typographical errors and misinterpretations of ambiguous terms and phrases. The authors pinpoint the fact that "the efficacy and usability of a statistical machine learning technique could be strongly affected by the quality of data available to a healthcare organization." The authors' main focus lies on analyzing the predictive power and robustness of a shrinkage-based classifier used for automatic ICD9-CM coding. For comparative reasons, the authors also use SVM and multinomial Naïve Bayes to classify their data. They trained and tested all three classifiers on a set of free-text discharge diagnoses, to which they incrementally added noise, that is by simulating typographical errors, randomly selected among a list of the most common errors in Spanish. Their results show that all three algorithms prove to be remarkably robust when subjected to training data with an increasing amount of errors, i.e., the impact of noise in the training data on the performance of the classifiers was minimal. Their findings lead us to conclude that applying machine learning to patient records in order to detect HAI can indeed be feasible although we consider our dataset to be noisy.

2.2 Detection of HAI

Research on detecting Hospital Acquired Infections from patient records has emerged throughout the last years.

A number of studies focus on an exclusive or primary use of the structured data of patient records, i.e., data which is derived directly from a hospital database. Klompas et al. [16] for example cite a number of studies where mainly microbiological data, antimicrobial criteria and diagnosis codes where used in automatic surveillance tools for identifying central line-associated bloodstream infections, surgical site infections, and VAP. The authors even pinpoint the fact that data, which is recorded as free text in patient records, they exemplify radiographic reports, eludes easy analysis by computers.

Bouzbid et al. [2] focus on comparing different strategies for detecting HAIs in intensive care units (ICUs). These strategies differ by whether data from a single or a combination of hospital databases is considered. Of all their strategies tested, the 'Drug prescription database or Microbiological database combined strategy' provided the highest recall[5] (99.3%), however at the expense of a low specificity (56.8%). The best ratio between recall and specificity was obtained for the 'Microbiological database strategy' and the 'Electronic hospital discharge summaries database'. The latter is however based on a manual analysis of the discharge summaries and no automatic detection tool yet [2].

The study described in [11] on the other hand developed a surveillance system for pneumonia in neonatal intensive care units (NICU). The system processes free text chest x-ray re-

[5]Bouzbid et al. [2] originally use the term sensitivity in their paper. Sensitivity as well as specificity are fundamental terms in medicine, used to evaluate clinical test [19]. Recall and precision are frequently used in computer science to measure the performance of a system. Sensitivity and recall refer to the exact same measure. To allow comprehensibility, we decided to consistently use recall in this study. Specificity and precision define slightly different measures and are used accordingly. See [19] for a comprehensive description of sensitivity and specificity.

ports with help of the MedLEE-tool[6] which extracts clinical information related to pneumonia. Subsequently, rules are applied in order to detect hospital acquired pneumonia. The authors obtain a recall value of 71%, and specificity value of 99.8%. Just like the authors in [11], the researchers in [26][7] focus in their latest ALADIN project[8] on utilizing unstructured data, in their case discharge summaries from several hospital units, to automatically detect HAI using Natural Language Processing [26], [2]. Their system yields a recall value of 87.6% and a specificity of 97.4% on their dataset.

Gerbier et al. [10] developed an application which automatically extracts and encodes medical concepts from emergency department narrative reports. This system is the first part within of a more comprehensive tool which shall analyze structured data and narrative reports in order to detect patients who may pose an epidemic risk. Their system so far obtains an overall recall of 85.8% and precision of 79.1%.

The above cited articles clearly illustrate the vast amount of research which focuses on automatic detection of HAI by analyzing structured or unstructured data in patient records. All of these approaches are rule-based, where the rules are handcrafted and not created by machine learning techniques. By contrast, studies which describe machine learning approaches in order to detect HAI are relatively rare.

The study by [4] is one of the few where a machine learning approach is chosen. The authors try to detect HAI by applying a SVM to patient records provided by the University Hospital of Geneva. Their focus lies on handling the imbalanced data, i.e., 11% of the patients records used contain HAI while 89% did not, and optimizing SVM in this regard. This is done by deploying a novel resampling strategy and asymmetrical soft-margin SVM respectively. When testing their resampling method, they used a variety of learning algorithms besides SVM. Without deploying any resampling methods, NB yielded a recall of 19% and a specificity of 96%, while SVM yielded a recall of 43% and a specificity of 92%. Deploying Combined AHC oversampling and K-means subsampling, Naïve Bayes yielded maximum performance level, 87% recall and 74% specificity of all classifiers tested as well as for all resampling preprocessing experiments. However, the asymmetrical soft-margin SVM obtained a recall of 92% and a specificity of 72%, thus clearly outperforming their resampling method.

Lastly, the authors in [3] conducted an observational study where they examined the performance of a Naïve Bayes classifier, CoCo, at "categorizing patients into 1 of 7 syndromes based on triage chief complaints.", which they obtained from the Emergency Department of the University of Pittsburgh's Medical Center. CoCo was trained on more than 10,000 chief complaints that were manually classified with any of the seven syndromic categories. The classifier's accuracy ranged from 92% to 99% while recall ranged from 30% to 75%.

[6] `http://www.cat.columbia.edu/medlee.htm`

[7] A large number of projects, focusing on automated rule-based HAI surveillance, are carried out in various constellations by researchers around Denys Proux, M-H. Metzger, S. Bouzbid and S. Gerbier in collaboration with Lyon University Hospital. They do, however, have different foci in there studies, some of which are cited in the course of this article.

[8] Collaboration of the Xerox Research Centre Europe in France and the Lyon University Hospital, `http://www.aladin-project.eu/index-en.html`.

3. PATIENT RECORDS

In our study we used patient records[9] provided by Karolinska University Hospital which were used as both training and test data. More specifically, we used PPM-records (Point Prevalence Measures) that encompasses all clinics. Generally, if a patient is admitted to a clinic, daily patient records (DPRs) are kept. Those records contain an unstructured part for notes, where general treatment and observations are written by nurses while more specific examination and treatment notes are written by physicians. When the patient is discharged, a discharge summary[10] is written by the physician, i.e., a summary of the treatment and also advice of how to care after discharge. Additionally, the patient record may contain structured parts, namely medication, microbiological data and body temperature, where the data is obtained from various hospital databases.

As cited above, various studies which utilized the unstructured part of the patient record, used discharge summaries for building their mainly rule-based systems. Our first attempt was to do likewise. However, based on an ocular analysis by our medical experts, it was found that, in some discharge summaries, no indications were given from which the medical experts could have inferred that an HAI occurred. Rather, information obtained from all (or numerous) daily patient records, which reflect the patients stay, gave information about whether an HAI occurred or not.

We thus decided to merge all daily patient records which belong to one hospitalization into one file, which we will call Hospitalization Records (HSR). Medical experts define one **hospitalization** as the stay of a patient at a health facility which is needed for one care process. In case the patient is discharged from one health facility and admitted to another one within 24 hours this is regarded as the same hospitalization. Moreover, even a noted event, which occurred 24 hours after discharge, is included in the hospitalization. Some patients, who were hospitalized for less than 48 hours, are not represented in our final dataset in order to minimize the risk of including infections which are not associated to the hospital. The time frame is based on international definitions of HAI and the incubation period of infections and is estimated to be less than 48 hours for a multitude of disorders. This procedure yielded a total of 213 PPM-HSRs of which 128 contain HAI and 85 do not. This is the data we use as input for the classifiers in our current study.

The number of daily patient records belonging to one hospitalization varies significantly, reflecting the number of days the patient has been hospitalized, i.e., from 2 to 144 DPRs for HSRs containing HAI and 3 to 93 DPRs for HSRs not containing HAI. Due to this variation, even the number of tokens per HSR varied. The number of tokens in those HSRs which contain HAI ranged from 172 to 48,150, yielding a total of 1,267,711 tokens. For those 85 HSRs which do not contain HAI the number of tokens for each file varied from 257 to 21,016, yielding a total of 282,197 tokens.

[9] This research has been approved by the Regional Ethical Review Board in Stockholm (Etikprövningsnämnden i Stockholm), permission number 2012/1838-31/3.

[10] Alternatively referred to as discharge diagnoses by for instance [20].

4. METHOD

4.1 Machine Learning

There is a large number of different learning algorithms and classifier models that could be applied in our task. A detailed discussion of classification models and their assumptions and properties is beyond the scope of this paper. We decided to apply three well-known techniques that have been shown to be very effective for text classification, namely Naïve Bayes (NB), Support Vector Machines (SVM) and decision trees (using C4.5). One important aspect which is common for all supervised learning techniques is feature selection. Both, [6] and [5], state that preprocessing, feature selection and parameter tuning have a large impact on performance, more than the actual choice of the classification model. For a more detailed description of Naïve Bayes and C4.5 see for instance [12] and for SVM see [25] or [6]. All three classifiers are part of the WEKA environment[11] and run in this context. For each algorithm the default parameters were used.

4.2 Preprocessing and feature selection

To be able to perform binary text classification with the WEKA classifiers, it is necessary to convert the given hospitalization records from their text format into the ARFF-format, i.e., the Attribute-Relation File Format, which is required as an input format by WEKA. WEKA provides a *DirectoryLoader* and a *StringToWordVector* which perform this task.

According to [5] and [30], the high-dimensional feature space, i.e., the amount of unique terms (words or phrases) which occur in the text documents to be classified, marks a major characteristic and difficulty in text classification, making it a non-trivial task for automatic classifiers. It is thus desirable to reduce the dimensionality of the data to be processed by the classifier, and therewith reduce execution time and improve predictive accuracy. By doing so, irrelevant features which can introduce noise into the data and thus obscure possible relevant feature are filtered out [7]. Per default, the feature space comprised 1,000 features, i.e., given no further parameter specifications, WEKA chooses the 1,000 most frequent terms based on their term frequency (TF). TF refers to the simplest weighting scheme where the weight of a term is equal to the number of times the term occurs in a document [24].

In our study, we deployed well known preprocessing and filter methods in order to optimize and reduce the feature space, respectively. Filtering methods operate independently of any learning algorithm by reducing features from the data before learning begins [12]. For more examples of different preprocessing and filtering techniques see for example [30], [6] and [7].

4.2.1 Lemmatization

In machine learning, stop word removal and lemmatization are frequently used methods when preprocessing data [6]. In our study we used the CST lemmatizer[12] in order to perform lemmatization. Lemmatization describes the process of reducing a word to a common base form, normally its dictionary form (lemma). This is achieved by removing inflectional forms and sometimes derivationally related forms of the word, by means of vocabulary usage and morphological analysis. For instance: *am*, *are*, *is* ⇒ be, or *hospitals*, *hospital's* ⇒ *hospital* [24]. For Swedish that is highly inflectional, lemmatization is more important than for English. The patient records were lemmatized separately, converted to the required ARFF-format using the tools cited above and then given as input to the classifiers.

4.2.2 Stop word removal

Moreover, we used stop word removal as a preprocessing method. Using the WEKA environment, we were able to apply a filter which removed all stop words from the input text at the same time we ran the *StringToWordVector*. Stop words are terms which are regarded as not conveying any significant semantics to the texts or phrases they appear in and are consequently discarded [8]. The filter was configured to use the Swedish stop list which is available via Snowball[13] and which comprises 113 words, such as *och* (Engl.: and), *att* (Engl.: to) or *i* (Engl.: in).

4.2.3 Infection-specific terms (IST)

In the course of our project, a terminology which contains infection-specific terms was built in a semiautomatic approach. The medical experts involved in this project supplied a seed set of about 30 infection-specific terms; these were based on frequent observations in the data as well as the experts' knowledge about infections. The seed set was then extended by finding related terms, e.g., synonyms or misspellings of the input term, through the use of an automatic synonym generator. One medical expert then analyzed the proposed terms with respect to whether they could be regarded as applicable, infection-specific terms or not. All relevant terms were added to the terminology, yielding a total of 1,045 terms. Infection-specific terms, such as kateter (Engl.: catheter), ultraljud (Engl.: ultrasound), operation (Engl.: surgery), or feber (Engl.: fever), are expected to be contained in patient journals in case an infection occurs. When using the terminology as a feature reduction technique, we removed all terms from the HSRs except for those which occur in the terminology. By means of this procedure, the feature space was decreased to 374.

4.2.4 TF-IDF

In a final approach, we assigned a Term Frequency-Inverse Document Frequency weight to all terms. TF was defined above. IDF is, according to [24], a mechanism used in combination with TF, to attenuate the effect of words that occur too often in the set of documents as that they could be important in order to discriminate between those. IDF is calculated as follows: $idf_t = log\frac{N}{df_t}$ where N is the number of documents in a collection and df_t is the document frequency of term t, i.e. the number of documents in the collection that contain t. TF-IDF for a term is calculated by: $tfidf_{t,d} = tf_{t,d} \times idf_t$. Thus, TF-IDF for a t is highest if t occurs many times within a small number of documents. We reduced the number of features by keeping the 200, 100, 70 and 50 terms with the highest TF-IDF scores. For more information on TF-IDF and different weighting schemes see [24].

[11] Version 3.6.7 is used. WEKA is available via `http://www.cs.waikato.ac.nz/ml/weka/`.

[12] `http://cst.dk/online/lemmatiser/uk/`

[13] `http://snowball.tartarus.org/algorithms/swedish/stop.txt`

Table 3: Precision, Recall and F-score (in %) for detecting HAI using NB, SVM and C4.5 given the different preprocessing methods. In total, the material comprised 213 PPM-HSR of which 128 contained HAI.

	Naïve Bayes			SVM			C4.5		
	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score
No preprocessing	76.2	60.2	67.2	76.9	**80.5**	78.6	71.4	78.1	74.6
Lemmatized	55.4	**72.9**	62.9	60.0	60.0	60.0	62.2	60.0	61.1
No stop words	78.1	58.6	67.0	74.8	**78.9**	76.8	74.0	75.8	74.9
IST	76.9	62.5	69.0	70.5	**67.2**	68.8	68.0	64.8	66.4
TF-IDF 50	66.4	78.9	72.1	66.9	**89.8**	76.7	68.3	85.9	76.1
LS-TFIDF	67.6	76.6	71.8	63.4	81.3	71.2	66.0	**82.0**	73.2

In an additional preproceesing step, lemmatization, stop word removal and TF-ID 50 were combined. This preprocessing step is named LS-TFIDF in Table 3.

4.3 Evaluation

4.3.1 10-fold cross-validation

For evaluating each classifier independently, stratified 10-fold cross-validation, one of the best known and most commonly used evaluation techniques, was used. When applied, the dataset of 213 HSRs is divided into 10 folds, the classifier is then trained on 9 folds and tested on the remaining one. The procedure is repeated until all folds have been used for testing once (compare [12]). Stratification ensures that each class, that is HAI and non-HAI, is properly represented in each fold with respect to the class distribution over the entire dataset, e.g., 128 or 60% of our HSRs belong to class HAI, thus each subset should consist of roughly 60% class HAI instances [21]. Cross-validation is especially useful if the dataset, as in our case, is small, as it maximizes the amount of training data [15].

4.3.2 Statistical tests

When comparing the classifiers' results, statistical testing is necessary in order to verify the significance of the results. In this study, the non-parametric sign test is used. The choice was motivated by the fact that the authors in [15] present this statistical test as being simple to calculate and yet appropriate when wanting to compare the performance of multiple classifiers on a single domain. Just like [15] do in their example calculations, the sign test was one-tailed and performed at 5% significance level. The null hypothesis states that the classifiers perform equally well. When applying the sign test to compare classifiers on one domain, multiple trials are made on the domain, i.e., by performing cross-validation. For each preprocessing technique, the performances of the classifiers are compared and statistically tested based on the classifiers' results on each fold. These results are not depicted due to space restrictions. Instead, the average performance measures are shown in Table 3.

5. RESULTS

Table 3 depicts the results of Naïve Bayes, SVM and C4.5 given the different preprocessing and feature selection methods. The best recall values for each preprocessing and feature selection method are highlighted. The focus of this study lies on obtaining high recall for HSRs which contain HAI. Thus, we only present the performance measures of the classifiers for those records. Precision, Recall and F-score for HSRs not containing HAI, which are neither depicted nor analyzed, were in all cases lower than the values obtained for HSRs containing HAI. This coherent result may be explained by the fact that the number of HSRs not containing HAI, which were available for training and testing, was less than the number of HSRs containing HAI, i.e., 85 compared to 128.

When considering recall, SVM yields the highest recall for four of six different preprocessing techniques, i.e., 80.5% (no preprocessing), 78.9% (stop word removal), 67.2% (IST) and 89.8% (TF-IDF 50). When comparing the results of SVM and NB, SVM performs significantly better than NB when no preprocessing, stop word removal and TF-IDF 50 are used for preprocessing. In case of IST as a preprocessing method, the results of SVM and NB, 67.2% compared to 62,5%, are not significantly different at a 5% significance level. When comparing SVM to C4.5, none of the results are significant. When lemmatization is applied, Naïve Bayes yields the highest recall at 72.9%, while C4.5 obtains the highest recall of 82.0% when lemmatization, stop word removal and TF-IDF 50 are combined (LS-TFIDF). However, compared to the results obtained by the other classifiers, those obtained recall values were found to be not statistically significant. The highest recall, 89.8%, of all three classifiers, when comparing all the different preprocessing techniques, is yielded by SVM when TF-IDF 50 is applied as a preprocessing technique. Compared to the recall obtained by NB, 78.9%, the recall is significantly better at a 5% significance level. When compared to the recall value obtained by C4.5, 85.9%, the result is not significantly different.

After pointing out the three highest recall values, all obtained by SVM, it is important to analyze the respective F-scores. This is because, as stated before, our overall aim is to obtain a recall value of 100%, with the highest precision possible. In other words, we want to achieve high recall without the precision being too low.

SVM obtains the highest F-score for three of six preprocessing techniques, that is 78.6% (no preprocessing), 76.8% (stop word removal) and 76.7% (TF-IDF 50). Compared to NB, SVM performs significantly better when no preprocessing and stop word removal are applied as preprocessing techniques. However, when TF-IDF 50 is applied, the difference between the F-scores obtained by SVM and NB are not statistically significant. When comparing the F-score of SVM and C4.5 for those three techniques, non of the results obtained by SVM are significantly better.

To summarize our observations, SVM obtains the highest recall and F-score on the following preprocessing methods: no preprocessing, stop word removal, and TF-IDF 50. The difference between the best recall value, 89.8% (TF-IDF 50), and the second best, 80.5% (no preprocessing), is 9.4

percentage points. The difference to the third best recall value, 78.9% (stop word removal) amounts to 10.9 percentage points. Compared to this spread, the respective F-scores remain quite close: 78.6% (no preprocessing), 76.8% (stop word removal) and 76.7% (TF-IDF 50), indicating a comparable overall performance. The highest F-score, 78.6%, is obtained when no preprocessing is applied. When comparing SVM (TF-IDF 50) and SVM (no preprocessing) this means:

- SVM (TF-IDF 50) obtains the highest recall with a slightly lower overall performance while
- SVM (no preprocessing) shows the overall best performance, having yet a considerable lower recall.

Since we aim at the highest recall with the highest precision possible, i.e., a reasonable overall performance in terms of F-score, we conclude that the performance of SVM (TF-IDF 50) can be considered to come closest to our objectives.

6. CONCLUSIONS

This paper focuses on deploying three machine learning algorithms to hospitalization records. By means of applying different preprocessing as well as feature selection methods, we tried to increase the recall values. Against our expectations, the results of the three machine learning algorithms were all in all very encouraging. Especially the drastic feature reduction by only taking the top 50 features according to their TF-IDF scores yielded the most promising recall value of 89.8% for SVM, while yielding a reasonable overall performance of 76.7%. The recall value is close to the 92% which [4] obtained with their asymmetrical soft-margin SVM. It is especially interesting to look at the features which the classifiers base their decision upon. Terms such as *och* (Engl.: and), *det* (Engl.: it) or *hon* (Engl.: she), i.e., stop words, are among the 50 top features. Intuitively, we would not expect the classifier to consider such terms in order to detect patient records which contain HAI, but rather terms like *operation* (Engl.: surgery), *feber* (Engl.: fever) or *kad* (Engl.: abbreviation for catheter), i.e., terms which are considered to be infection-specific by the medical experts. This observation asks for a more thorough analysis of the terminological structure of patient records in order to optimize feature selection. The best F-score was 78.6%, obtained by SVM when no preprocessing is used. Moreover, none of the classifiers outperformed the others, indicating the applicability of Naïve Bayes, SVM well as C4.5 for our task.

Further, we are well aware of the fact that our dataset was small and that the differences in the results of Naïve Bayes, SVM and C4.5 are marginal and not in all cases significantly different. Yet, we are convinced that the results reveal the potential of applying machine learning techniques to patient records, including the structured as well as unstructured parts. This is further motivated by the fact that, so far, we have neither tuned parameters of the different classifiers nor used particularly elaborate preprocessing and feature reduction methods. Future research will thus have to focus upon improving the scores by, for instance, using wrapper techniques for feature reduction which are optimized on a specific learning algorithm and therefore, according to [12], yield better results.

Moreover, the medical experts involved will, in the course of the project, value 292 additional HSRs from the rheumatic clinic at Karolinska University Hospital. Thus, we will be able to train the classifiers on about twice as much data as we did now, leading us to expect an improvement in performance. In addition, we aim at training the classifiers on a more realistic dataset, i.e., a dataset which is less balanced than our current one with regard to the number of patient records containing HAI and not containing HAI, respectively.

Furthermore, the obtained results verify that applying machine learning techniques is a reasonable approach even though the dataset is noisy. Still, it will be interesting to analyze (1) if the classifiers turn out to be robust to noise when trained on a larger dataset or (2) if the deployment of noise reduction techniques in a preceding step to the actual learning task can reduce uncertainty in the data and improve the performance when classifying patient records. The authors in [20] tested the impact of noise on the performance of the classifiers by incrementally adding noise to the data. An alternative would be to reduce noise instead of adding it, e.g. through the correction of spelling errors or abbreviation resolution, and analyze how this effects the performance.

Finally, the overall goal will continue to be obtaining high recall (approaching 100%) with the highest precision possible for hospitalization records. This will enable us to implement a system which can screen all hospitalization records, and filter out all HSRs which contain HAI. This would reduce the workload for hospital staff tremendously as they only need to analyze those HSRs which were preselected by the system.

7. ACKNOWLEDGEMENTS

We would like to thank our excellent physicians Maria Kvist and Elda Sparrelid for their insightful classifications of Hospital Acquired Infections as well as Martin Hassel for suggestions regarding preprocessing and computing methods.

8. REFERENCES

[1] H. Allvin, E. Carlsson, et al. Characteristics of Finnish and Swedish intensive care nursing narratives: a comparative analysis to support the development of clinical language technologies. *Journal of Biomedical Semantics*, 2 Suppl 3:1–11, 2011.

[2] S. Bouzbid, Q. Gicquel, et al. Automated detection of nosocomial infections: evaluation of different strategies in an intensive care unit 2000-2006. *Journal of Hospital Infection*, 79:38–43, 2011.

[3] W. W. Chapman, J. N. Dowling, and M. M. Wagner. Classification of emergency department chief complaints into 7 syndromes: A retrospective analysis of 527,228 patients. *Annals of Emergency Medicine*, 46(5):445–55, 2005.

[4] G. Cohen, M. Hilario, et al. Learning from imbalanced data in surveillance of nosocomial infection. *Artificial Intelligence in Medicine*, 37:7–18, 2006.

[5] F. Colas and P. Brazdil. Comparison of SVM and some older classification algorithms in text classification tasks. *Artificial Intelligence in Theory and Practice*, 217:169–178, 2006.

[6] M. K. Dalal and M. A. Zaveri. Automatic text classification: A technical review. *International*

Journal of Computer Applications, 28(2):37–40, 2011.

[7] S. Doraisamy, S. Golzari, et al. A study on feature selection and classification techniques for automatic genre classification of traditional malay music. In *Ninth International Conference on Music Information Retrieval*, pages 331–336, 2008.

[8] E. Dragut, F. Fang, et al. Stop word and related problems in web interface integration. In *Proceedings of the VLDB Endowment*, pages 349–360, 2009.

[9] G. Ducel, J. Fabry, and L. Nicolle, editors. *Prevention of Hospital Acquired Infections: A Practical Guide.* World Health Organization, 2 edition, 2002.

[10] S. Gerbier, O. Yarovaya, et al. Evaluation of natural language processing from emergency department computerized medical records for intra-hospital syndromic surveillance. *BMC Medical Informatics and Decision Making*, 10(1):50–62, 2011.

[11] J. P. Haas, E. A. Mendonca, et al. Use of computerized surveillance to detect nosocomial pneumonia in neonatal intensive care unit patients. *American Journal of Infection Control*, 33(8):439–443, 2005.

[12] M. A. Hall. *Correlation-based Feature Selection for Machine Learning.* PhD thesis, The University of Waikato, Hamilton, New Zealand, 1999.

[13] H. Humphreys and E. T. M. Smyth. Prevalence surveys of healthcare-associated infections: what do they tell us, if anything? *Clinical Microbiology and Infection*, 12(1):2–4, 2006.

[14] N. Isenius. Abbreviation detection in Swedish medical records. The development of SCAN, a Swedish clinical abbreviation normalizer. Master's thesis, Department of Computer and Systems Sciences, Stockholm University, 2012.

[15] N. Japkowicz and M. Shah. *Evaluating Learning Algorithms: A Classification Perspective.* Cambridge University Press, 2011.

[16] M. Klompas and D. S. Yokoe. Automated surveillance of health care-associated infections. *Clinical Infectious Diseases*, 48(9):1268–1275, 2009.

[17] C. Knoblock, D. Lopresti, et al. Special issue on noisy text analytics. *International Journal on Document Analysis and Recognition*, 10(3):127–128, 2007.

[18] K. Kukich. Techniques for automatically correcting words in text. *ACM Computing Surveys (CSUR)*, 24(4):377–439, 1992.

[19] A. G. Lalkhen and A. McCluskey. Clinical tests: sensitivity and specificity. *Continuing Education in Anaesthesia, Critical Care & Pain*, 8(6):221–223, 2008.

[20] E. J. Lauría and A. D. March. Combining bayesian text classification and shrinkage to automate healthcare coding: A data quality analysis. *Journal of Data and Information Quality (JDIQ)*, 2(3):1–22, 2011.

[21] N. Lavesson. *Evaluation and Analysis of Supervised Learning Algorithms and Classifiers.* PhD thesis, Department of Systems and Software Engineering. School of Engineering. Blekinge Institute of Technology, 2006.

[22] V. Liu and J. R. Curran. Web text corpus for natural language processing. In *Proceedings of the 11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 233–240, 2006.

[23] D. Lopresti, S. Roy, et al. Special issue on noisy text analytics. *International Journal on Document Analysis and Recognition*, 14(2):111–112, 2011.

[24] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval.* Cambridge University Press, 2008.

[25] W. S. Noble. What is a support vector machine? *Nature Biotechnology*, 24(12):1565–1567, 2006.

[26] D. Proux, C. Hagège, et al. Architecture and systems for monitoring hospital acquired infections inside a hospital information workflow. In *Proceedings of the Workshop on Biomedical Natural Language Processing*, pages 43–48, 2011.

[27] P. Ruch, R. Baud, and A. Geissbühler. Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record. *Artificial Intelligence in Medicine*, 29(1):169–184, 2003.

[28] M. Skeppstedt, M. Kvist, and H. Dalianis. Rule-based entity recognition and coverage of SNOMED CT in Swedish clinical text. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012*, pages 1250–1257, 2012.

[29] L. V. Subramaniam. Noisy text analytics. In *The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics - NAACL HLT 2010*, 2010.

[30] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *Proceedings of the 14th International Conference on Machine Learning*, pages 412–420, 1997.