Proceedings of MultilingualBIO 2020, pages 1–8
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Detecting Adverse Drug Events from Swedish Electronic Health Records Using Text Mining

Maria Bampa, Hercules Dalianis
Department of Computer and System Science
Stockholm University
{maria.bampa, hercules}@dsv.su.se

Abstract

Electronic Health Records are a valuable source of patient information which can be leveraged to detect Adverse Drug Events (ADEs) and aid post-mark drug-surveillance. The overall aim of this study is to scrutinize text written by clinicians in Swedish Electronic Health Records (EHR) and build a model for ADE detection that produces medically relevant predictions. Natural Language Processing techniques are exploited to create important predictors and incorporate them into the learning process. The study focuses on the five most frequent ADE cases found in the electronic patient record corpus. The results indicate that considering textual features, can improve the classification performance by 15% in some ADE cases, compared to a method that used structured features. Additionally, variable patient history lengths are included in the models, demonstrating the importance of the above decision rather than using an arbitrary number for a history length. The experimental findings suggest the importance of the variable window sizes as well as the importance of incorporating clinical text in the learning process, as it is highly informative towards ADE prediction and can provide clinically relevant results.

1. Introduction

With the introduction of Electronic Health Records (EHRs) an abundant of information has become available. This provides unique opportunities not only for monitoring patients but also for the use of these data sources in secondary research. An EHR contains all the key information regarding a patient case over time, including demographics, medication, diagnoses and procedures, vital signs, laboratory results and hand-written text. Some of the aforementioned are captured in a structured format, for example, drug and diagnoses codes are represented in the ATC and ICD-10 format respectively. However, the vast majority of this information is captured in an unstructured and non-standardized format, i.e. clinical free text notes.

As EHRs are a vast source of patient medical history, they have enabled more efficient retrospective research in various domains, namely epidemiology, public health research, outcome research and drug surveillance (Weiskopf et al., 2013). Specifically, in drug surveillance, EHRs are an alternative method to evaluate drug risk and mitigate the problem of Adverse Drug Reactions (ADEs). ADEs refer to injuries caused by medication errors, allergic reactions or overdoses, and are related to drugs[1]. They can happen in different settings of patient care, from hospitals to outpatient settings, after a drug has been released to the market. In the United States alone, each year, they account for approximately 2 million hospital stays, more than 1 million emergency department visits and cause prolonged hospitalizations[2]. Due to several factors and barriers that come with ADE reporting, they are heavily under-reported in EHRs, causing in that way a long-term burden in the healthcare sector and in the individuals suffering an ADE. Nevertheless, it is estimated that about half of the ADEs are preventable[3], indicating the importance of directing research in post-market drug surveillance, to reduce withdrawal of drugs from the market and more importantly lessen human suffering.

EHRs are representative for a wide range of patients, specifically for patients with different diseases, in different age and gender distribution. Data and text mining methods can be employed to leverage this information and predict unwanted ADEs. In the side of structured data sources stemming from EHRs, previous research has mainly focused on utilizing specific predictors, for example ICD-10[4], ATC[5] or laboratory results, to predict ADEs. A recent work by Bamba and Papapetrou (2019) has utilized the temporal and hierarchical aspect of the previously mentioned data sources to predict ADEs and concluded in a framework with high classification performance. Additionally, they experimented with variable history lengths before the occurrence of an ADE and indicated its importance in the experiments. However, they only utilized features in a structured format and did not consider important information that can be found in the text that accompanies the majority of patients.

To meet the challenges posed by narrative data, text mining is commonly used to extract and retrieve relevant information by recognizing statistical patterns in the text. In previous research the use of Natural Language Processing (NLP) has been investigated for obtaining models that are able to predict unseen ADEs from EHRs. For example, Eriksson et al. (2013) constructed a dictionary from a Danish EHR and managed to identify 35,477 unique possible ADEs. Henriksson et al. (2015) have modeled Swedish EHR data in ensembles of semantic spaces and reported

[1] ADE, `https://health.gov/hcq/ade.asp`

[2] https://health.gov/our-work/health-care-quality/adverse-drug-events

[3] https://psnet.ahrq.gov/primer/medication-errors-and-adverse-drug-events

[4] ICD-10, `https://www.icd10data.com`

[5] ATC, `https://www.whocc.no/atc_ddd_index/`

improved performance in 27 datasets. Additionally, An NLP system named MedLEE, was used to create discharge summaries and outperformed traditional and previous automated adverse event detection methods (Melton and Hripcsak, 2005).

To the best of our knowledge existing data mining approaches for ADE prediction in Swedish EHRs, have been mainly focusing on utilizing specific structured data types. Moreover, many of the studies do not take into account the importance of considering variable window lengths depending on the ADE studied. Exploiting a very large patient history window length can add noise to the data and a very small window size can eliminate useful and informative predictors.

Contributions. This paper, follows the work of Bamba and Papapetrou (2019) utilizing variable window lengths, but instead incorporating in the machine learning process textual features, rather than structured, that can be highly informative predictors for the specific ADEs studied. Specifically, the state-of-the-art is extended by:

1. including textual features, using the n-gram model and tf*idf weighting,

2. exploring variable patient history trajectories for each of the ADEs

3. benchmarking the proposed approach in three classification models.

As shown in Section 5 the incorporation of text features in the learning process, combined with the different window lengths for each ADE can provide improvements in the classification performance while providing medical sound predictions.

2. Related Work

EHRs contain a wealth of longitudinal patient history which can be leveraged to create models for personalized-care and provide clinically relevant insights. Post-market drug surveillance based on EHRs can lead to further investigation and regulatory warnings about drugs (Karimi et al., 2015) and a decrease in drug withdrawal from the market. However, EHRs suffer from several disadvantages such as under-reporting, absence of protocols and reporting bias (Karimi et al., 2015), and in that way, the prevalence of an ADE cannot be estimated with full confidence. Previous research on EHRs tried to tackle problems like the aforementioned, utilizing a wide range of predictors to identify ADEs. This section summarizes research conducted towards ADE prediction from EHRs. The first paragraph presents research that utilized the structured data founds in EHRs; the rest of this section describes works that have focused on exploiting the textual features of EHRs to predict ADEs.

Studies that use structured clinical codes (diagnoses and/or drug codes) focus on different ways of representing them by internationally defined standards (ICD diagnosis and ATC drug codes respectively) and conclude that predictive performance was significantly improved when using the concept of hierarchies (Zhao et al., 2014; Zhao et al., 2015b). Other related work utilizes clinical codes and clinical measurements while taking their temporal aspect into account, for identifying ADEs (Zhao et al., 2015c). Studies in this area typically exploit logistic regression (Harpaz et al., 2010) or Random Forests (Zhao et al., 2015a) applied in clinical codes to identify ADEs. Using only laboratory abnormalities Park et al. (2011) used the underlying temporality of these events to predict ADEs. Finally, (Bagattini et al., 2019) focused on lab results extracted from EHRs and proposed a framework that transforms sparse and multivariate time series to single valued presentations that can be used be any classifier to identify ADEs; they conclude that taking into account the sparsity of the feature space can positively affect the predictive performance and be effectively utilized to predict ADEs.

The unstructured sections of EHRs, i.e. free-text, have also been used to detect ADEs. The main approach in this line of research is to employ NLP techniques to transform the text in some form of structured features in order to feed machine learning classifiers (Karimi et al., 2015). For example, (Wang et al., 2009) and (Melton and Hripcsak, 2005) have both used the MedLee NLP system to identify adverse drug event signals and they outperformed traditional and previous automated adverse event detection methods. MedLee is a natural language processor that extracts information from text employing a vocabulary and grammar and has been extended to cover a spectrum of applications (Melton and Hripcsak, 2005). LePendu et al. (2013) proposed a method to annotate the clinical notes found in EHRs and using medical terminologies, transformed them to a de-identified matrix. Eriksson et al. (2013) identified a wide range of drugs by creating an ADE dictionary from a Danish EHR. Furthermore, Henriksson et al. (2015) focused on Swedish EHR data and reported improvement in ADE detection by exploiting multiple semantic spaces built on different sizes, as opposed to a single semantic space. Finally, the combination of local and global representation of words and entities has proven to yield better accuracy than using them in isolation for ADE prediction according to Henriksson (2015).

3. Data

The clinical dataset Stockholm EPR Structured and Unstructured ADE corpus, (SU-ADE Corpus)[6] used in this study consists of information representing more than a million patients from Karolinska University Hospital in Sweden. The SU-ADE Corpus is an extract from the research infrastructure Health Bank (the Swedish Health Record Research Bank) at DSV/Stockholm University that contain patients from 512 clinical units encompassing the years 2007-2014 originally from the TakeCare CGM electronic patient record system at Karolinska University Hospital (Dalianis et al., 2015).

[6] Ethical approval was granted by the Stockholm Regional Ethical Review Board under permission no. 2012/834-31/5.

Both structured and unstructured data are part of the database and are also timestamped. Structured data are labelled using common encoding systems such as the Anatomical Therapeutic Chemical Classification System (ATC) for medications, the International Statistical Classification of Diseases and Related Health Problems, 10th Edition (ICD-10) for diagnoses as well as the Nomenclature, Properties and Units (NPU) coding system[7] for clinical laboratory measurements. Regarding unstructured format, each patient is described by a text that is written in free format by clinicians; the SU-ADE Corpus contains more than 500 million tokens, in total.

The ADE groups of this study were selected as they are some of the most frequent in the SU-ADE Corpus. The specific 5 ADE cases were chosen for comparison reasons to the paper by Bamba and Papapetrou (2019)(see section 5). The experiments are formulated as a binary classification task; according to patients' ICD-10 codes, labels are assigned to each of them. More concretely, each patient in a dataset is described by a class label that denotes if that patient has an ADE or not. Negative examples are denoted as 0, while positive examples are denoted as 1. The following procedure was adopted: Patients that are assigned a specific ADE code are considered positive to that ADE (Stausberg and Hasford, 2011), while patients that are not assigned that specific ADE code but have been given a code that belongs to the same disease taxonomy are considered ADE negative. For example, patients that are given the ADE code D61.1 (drug induced aplastic anaemia) are positive to that specific ADE, on the other side, patients that are given codes that belong to D61.x with x $\neq$ 0 are considered ADE negative. A list and an explanation for each dataset can be seen in Table 1.

The ICD-10 codes serve only as reference for the extraction of the sub-datasets and the subsequent class labeling. In that way, from the original corpus we extract all the patients that have at least one reference of the following codes in their history: D61.*, E27.*, G62.*, L27.*,T80.* (* denotes every possible digit from one to nine). Following that, we create the sub-datasets according to the ADE codes and assign the class labels as described above. The patients are then described by text written by a healthcare expert. The main methodology is described in section 4.

4. Methods

The following section provides a description of the methods used in this work. In Figure 1 the process of the method that was used is depicted.

4.1. Text Preprocessing

Following the class labeling, text assigned to each patient is then pre-processed, so as to bring it in a format that is analyzable and predictable. Swedish is different from English thus the techniques used are different, for example Swedish is a highly inflected language as well as a compounding language, similar to German.

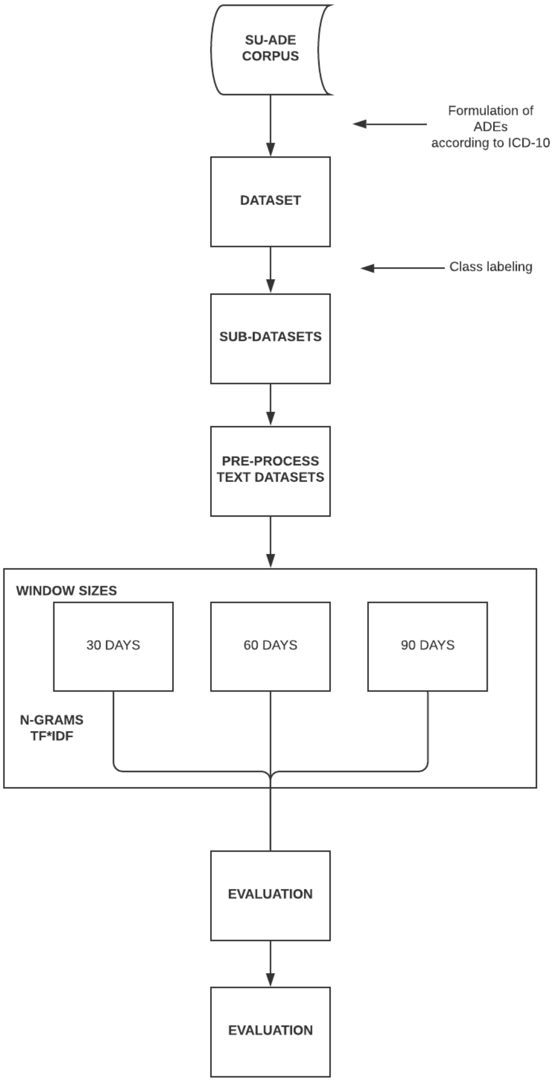

Figure 1: Depiction of the method flow. Starting from the SU-ADE corpus to the creation of the 5 ADE datasets and finally evaluation of the model.

Since the datasets that are handled in this work are very large, to help with the consistency of the expected output, all the words were lower-cased. Also, noise and Swedish stop word removal was carried out to help reduce the number of features before classification and produce consistent results. Stop-words do not convey any significant semantics in the output result, consequently they were discarded. Finally, lemmatisation was performed to transform the Swedish words into their dictionary form, a procedure highly important in our study as the Swedish language is highly inflectional (Carlberger et al., 2001). The library used for stop word removal and lemmatisation was NLTK[8].

[7] NPU, http://www.ifcc.org/ifcc-scientific-division/sd-committees/c-npu/

[8] Natural Language Toolkit, https://www.nltk.org/

Dataset	Description	Positive	Negative
D61.*	Aplastic Anaemia	557 (D61.1)	94($D61.x$)
E27.*	Adrenocortical insufficiency	55($E27.3$)	219 (E27.x)
G62.*	Polyneuropathy	79($G62.0$)	672 (G62.x)
L27.*	Generalized skin eruption	391 (L27.0)	172 (L27.x)
T80.*	Infusion complications	502 (T80.8)	135 (T80.x)

Table 1: The 5 most common ADE cases studied. The * under the Dataset column: denotes every possible code under the specific category included in the dataset; column *Positive* depicts the number of ADE positive patients and the specific ADE code is in parentheses; column *Negative* depicts the number of ADE negative patients where x is any number besides the last digit of the ADE depicted in each row and the examples are in parentheses.

4.2. Window Sizes

Furthermore, as this study focuses on events that occur over various time periods, the sub-datasets are created on different window sizes, in order to investigate potentially informative patient trajectories for specific ADEs. 30, 60 and 90 days window sizes are investigated. In those cases, the day when the ADE was registered was excluded from the learning process, as we are interested in predicting patients with a potential ADE.

4.3. Word Vectors

The following representations of the text at word level are considered: *unigrams, bigrams, trigrams (n-grams)* and *tf*idf* (in a uni-gram word level) where *tf* stands for term frequency and *idf* for inverse document frequency (Van Rijsbergen, 1979).

The n-gram model predicts the occurrence of the n-th word based on the occurrence of n-1 words in the document. This follows the Markov assumption that only the previous words affect the next word (Manning et al., 1999). For instance, the bigram model (n=2) predicts the occurrence of a word given only its previous word, while the trigram model (n=3) predicts the occurrence of a word based on the previous two words. Even though the trigram model may be useful as it could predict sequences of words that have better medical meaning in terms of interpretability (ex. *500mg of paracetamol* (In English)), it may not be practical as it can increase the number of parameters and also the dimensionality of feature space.

The final approach is to assign a tf*idf weighting to all terms. *tf* is the simple term frequency; *idf* assigns in each term a weight such that it can compensate for the words that occur too often in the document, as they can be important to discriminate between others (Schütze et al., 2008). The number of features are reduced to a maximum of 500 terms, those with the highest tf*idf scores. A motivation for *tf*idf* was found in (Ehrentraut et al., 2016) where they utilized tf*idf, to classify healthcare associated infections and this method yielded the best results while extracting the most relevant results.

4.4. Classification

Three algorithms are benchmarked to evaluate the performance of the used methods :

- **RF**: Random Forests with 100 trees, gini impurity as the split criterion and the number of features considered at decision split criterion set to default$\sqrt{m}$ where m is the number of features in each dataset;
- **SVMlinear**: Support Vector Machines using a linear kernel and weighted class balance;
- **SVMrbf**: Support Vector Machines using the RBF kernel and weighted class balance;

4.5. Evaluation

All models were trained in the four different word sequences and for the three different window sizes. Stratified ten-fold cross validation was used, as described in (Kohavi, 1995) to ensure more efficient use of data. Since the datasets in this study are imbalanced, the Area Under the ROC Curve (AUC) was considered to be the most appropriate measure, as it has been proved to be an appropriate evaluation for skewed classification problems (Fawcett, 2006). Nevertheless, as in some cases the class imbalance favors the negative examples, the metrics precision, recall (as described in (Van Rijsbergen, 1979)) and F1 score were used to evaluate the results of each class independently.

5. Results

Table 2 presents our results in terms of predictive modelling. Five different types of ADEs expressed in the following ICD-10 codes D61.1, E27.3, G62.0, L27.0 and T80.8 are investigated. The table is separated in three sub-tables that present the investigation of variable window sizes for each ADE. The columns present the investigation of unigrams, bigrams, trigrams and tf*idf, for each ADE and window size. The results are depicted as mean AUC. Table 3 presents a classification report for the best performing window size and word sequence method for each ADE. Reported are: precision, recall, F1 score and support for both positive and negative classes, for the previously mentioned ADEs. Note that in binary classification, recall of the positive class is also known as sensitivity; recall of the negative class is specificity. Finally, Table 4 compares our classification results to the approach by Bamba and Papapetrou (2019).

Word vector representation. First, we investigate the importance of different kind of representations in a word level, for each ADE. We observe that although all n-gram approaches perform well they are almost always outperformed by the *tf*idf* approach. Specifically, ADEs

E27.3, G62.0, L27.0 and T80.8 had better classification performance when considering the inverse document frequency (idf) with the SVM linear classifier. Comparing the n-grams, the unigram was always performing better than the bigram and trigram approaches (sequence of two and three adjacent words), where the results are anging from 2% to 9% improvement. Unigram was always the second best performing after tf*idf.

Window Sizes. The aim in this section is to investigate variable window sizes in the patient trajectory following the work of Bamba and Papapetrou (2019). We can see that L27.0 acquired better results in a small window size of 30 days and E27.3, T80.8 gave an improvement of 1% to 3% in a window size of 60 days as compared to 30 and 90 days. For ADEs D61.1 and G62.0 the best results are obtained in a 90 days patient history length with AUC 0.9542 and 0.9045 respectively.

Classification Report. Furthermore, for each best performing size and word vector representation we provide a classification report for both negative and positive classes. In Table 3, we observe that for the ADEs D61.1 and T80.8 where the class imbalance favors the positive class, the precision and recall are high. However, for ADEs E27.3 and G62.0 we can see that the classifier is not performing well in the positive class, failing both to retrieve and correctly classify the cases, as the class distribution is skewed towards the negative class.

Comparison to LDM approach. Finally, we compare our approach to the LDM framework as described in (Bamba and Papapetrou, 2019). In this paper the authors studied the importance of incorporating three different types of structured predictors in the learning process, Lab measurements, Drug codes, Diagnoses codes (LDM) while using variable window sizes. In table 4, depicted are the best performing windows sizes and classifiers for each of the approaches. We observe that for 4 out for 5 classifiers, employing features found in the clinical text improves the classification task. Specifically, there is an improvement of 1% for D61.1, 13% for E27.3, 2% for G62.0 and 15% for L27.0.

5.1. Important Features and Medical Relevance

In this section top textual features are provided that were found important by the SVM classifier, for two of the studied ADEs. We are interested in investigating the features that the classifier based its decision upon and additionally, see if the are medically relevant. We only consider the results from SVM linear classifier and use the weights obtained from the coefficients of the vector which are orthogonal to the hyperplane and their direction indicates the predicted class. The absolute size of the coefficients in relation to each other can then be used to determine feature importance for the data separation task.

In figures 2, 3 we observe the most important features for both the negative and positive classes as decided by the SVM linear classifier for ADEs D61.1 (drug induced aplastic anaemia) and L27.0 (drug induced generalized skin eruption). Among the most import features for D61.1 are the words (In Swedish but also translated to English in parenthesis) *thrombocyter (platelets), sandimmun (a drug), blodtransfusion (blood transfusion), cytostatica (cytostatics), lymfom (lymphoma)* and *crp (a protein in blood made by the liver).* For example, according to the literature, irregular levels of plateles in the blood are indicators of aplastic anaemia and a way to treat is by blood transfusions (NIDDK, 2019). For L27.0 among the most important features are *svullnad (swelling), mjölk (milk), ägg (egg), övre (upper), hudutslag (rash), nötter (nuts), andningsljud* (noises heard on auscultation over any part of the respiratory tract), *mildison* (cream prescribed to relieve skin inflammation and itching), *reagera (react), akuten (emergency unit), hb (hemoglobin), ser (look), stor (big)* and *remiss (referral).*

These words are highly relevant for each ADE studied, thus indicating that the model is not performing at random. Nevertheless, we can observe that words such as the abbreviation *pga (because of)* or the numerical value *14*, are considered important features but cannot be related to the ADEs at a first glance. In the future, it would be of great importance to incorporate a medical expert in the process in order to validate the procedure and results, so as to create a safe and interpretable prediction model. Additionally, we observe that in some of the ADEs, the top important features include drugs or diagnoses that are administered and registered after the manifestation of the ADE. This indicates that the adverse events might be registered in the record at a later point in time, thus capturing both the treatment and the diagnosis of the ADE.

6. Analysis

The increased adoption of Electronic EHRs has brought a tremendous increase in the quantities of health care data. They contain records that offer a holistic overview of a patient's medical history, rendering them a valuable tool source for drug safety surveillance. Machine learning methods can be employed to uncover clinical and medical insights stemming from both structured and unstructured data to detect ADEs. Existing approaches on ADE prediction from EHRs have been mainly focusing on utilizing structured data types, on the other hand, text mining techniques have focused on identifying ADEs globally rather than focusing on specific types that occur frequently. This paper followed the work of Bamba and Papapetrou (2019) and incorporated in the learning process textual features while considering variable window lengths, for the five most frequent ADEs found in the SU-ADE corpus.

The experimental findings suggest that the textual features contain information that is highly important for ADE prediction. We observe that in many cases the word predictors outperformed the framework by Bamba and Papapetrou (2019) where the only utilized structured lab measurements, diagnoses and medication codes. In section 5.1 we included a number of important predictors as found by the SVM linear classifier, indicating that the model is not performing at random. We observed that some of

						30 DAYS						
		unigram			**bigram**			**trigram**			**tf*idf**	
	RF	**svmLin**	**svmRbf**	**RF**	**svmLin**	**svmRbf**	**RF**	**svmLin**	**svmRbf**	**RF**	**svmLin**	**svmRbf**
D61.1	0.9330	0.8707	0.8780	0.9159	0.8504	0.9133	0.8630	0.7604	0.8179	**0.9408**	0.9432	0.9249
E27.3	0.8109	0.7466	0.7113	0.6970	0.6928	0.7384	0.6947	0.7481	0.7272	0.7985	**0.8700**	0.8630
G62.0	0.7875	0.7436	**0.8805**	0.6879	0.6871	0.7796	0.6777	0.6833	0.7002	0.8235	0.8268	0.8666
L27.0	**0.9272**	0.8491	0.9118	0.8863	0.8328	0.8811	0.8113	0.8145	0.7920	0.9226	0.9109	0.9031
T80.8	0.8929	0.8173	0.8871	0.8829	0.8168	0.8621	0.8221	0.8105	0.8339	0.8863	**0.9060**	0.8962
						60DAYS						
D61.1	0.9354	0.8632	0.8572	0.9369	0.8882	0.9025	0.8597	0.7585	0.8477	0.9415	**0.9502**	0.9275
E27.3	0.8224	0.7941	0.7567	0.7574	0.766	0.7674	0.7040	0.7294	0.7194	0.8626	**0.8822**	0.8677
G62.0	0.8382	0.7603	0.828	0.7660	0.7301	0.8467	0.7490	0.721	0.7956	0.8547	0.8742	**0.8829**
L27.0	**0.9206**	0.8408	0.9089	0.8882	0.8303	0.8862	0.8134	0.8083	0.7997	0.9120	0.9108	0.8936
T80.8	0.9076	0.8274	0.8865	0.9042	0.8304	0.8887	0.8581	0.8609	0.8691	0.9185	**0.9207**	0.9171
						90DAYS						
D61.1	0.9403	0.8933	0.8503	0.9245	0.8836	0.8980	0.8651	0.8128	0.8425	0.9424	**0.9542**	0.9352
E27.3	0.7833	0.7647	0.7367	0.7219	0.6886	0.7259	0.6748	0.6848	0.6912	0.8078	**0.8420**	0.8337
G62.0	0.8357	0.7902	0.8454	0.8000	0.7639	0.8666	0.7749	0.7341	0.8048	0.8788	**0.9045**	0.8892
L27.0	**0.9216**	0.8427	0.9085	0.8811	0.824	0.8834	0.8248	0.7995	0.8036	0.9165	0.9142	0.8941
T80.8	0.8984	0.7936	0.8565	0.8801	0.8172	0.8856	0.8623	0.8447	0.8654	0.8836	**0.9005**	0.8944

Table 2: AUC obtained by 3 classifiers on 3 different patient history lengths for 5 different ADE cases and 4 different word weighting factor approaches. Each table presents the AUC obtained by stratified 10-fold cross validation on the different window sizes. In bold: best AUC for each ADE in the specific window size across all approaches, In red: Best AUC for the specific ADE across all window sizes, classifiers, and approaches;

	Class	Precision	Recall	F1 score
D61.1	Negative	0.78	0.80	0.79
	Positive	0.90	0.89	0.89
E27.3	Negative	0.95	0.85	0.89
	Positive	0.28	0.56	0.37
G62.0	Negative	0.98	0.91	0.94
	Positive	0.30	0.62	0.40
L27.0	Negative	0.89	0.83	0.86
	Positive	0.72	0.82	0.77
T80.8	Negative	0.70	0.79	0.74
	Positive	0.93	0.89	0.91

Table 3: Classification report of each ADE in the best performing classifier and window size for each of them (the ones reported as red in Table 2). Support: the number of occurrences of each class in the correct values

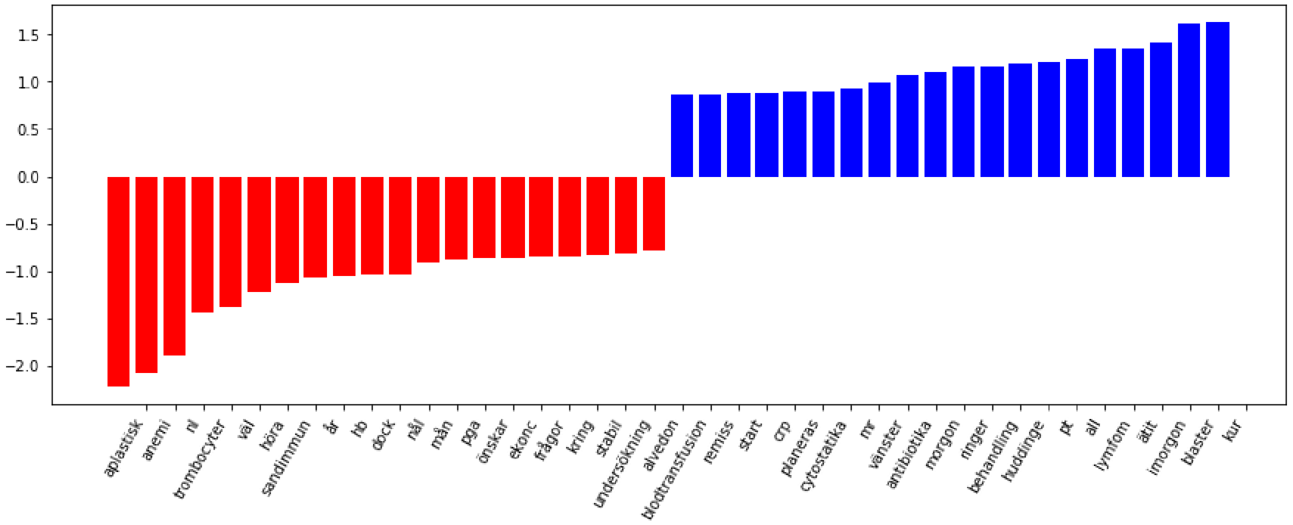

Figure 2: Top 20 feature importance for D61.1 in a 90 days window size using the tf*idf weighting and SVM linear. X-axis: Feature words in Swedish, Y-axis: Vector coefficients.

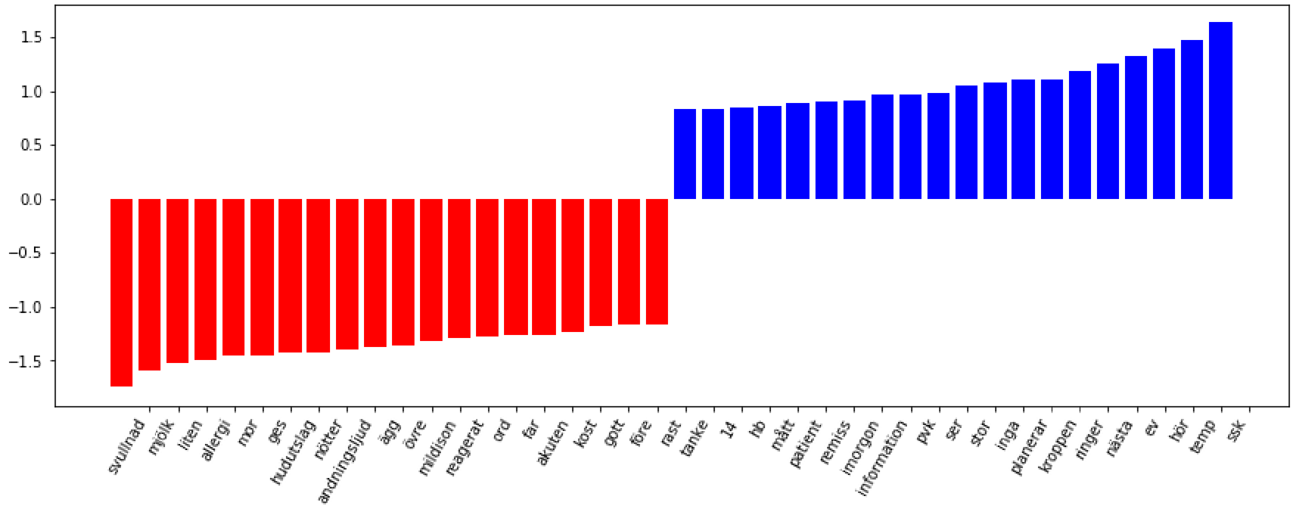

Figure 3: Top 20 feature importance for L27.0 in a 30 days window size using the tf*idf weighting and SVM linear. X-axis: Feature words in Swedish, Y-axis: Vector coefficients

	tf*idf	**LDM**	**WS**
D61.1	0.954	0.948	90
E27.3	0.882	0.756	30
G62.0	0.904	0.880	90
L27.0	0.927	0774	60
T80.8	0.920	0.946	30

Table 4: Comparison of textual tf*idf and LDM (Labs, Diagnoses, Medication) approach. WS: Best Performing Window Size for each ADE

the features are highly relevant with each ADE studied; for L27.0 (drug induced skin eruption) important features were swelling, egg and nuts or rash. This indicated that incorporating the clinical text in the learning process can provide medically sound predictions and provide a more interpretable model. Moreover, we observed that, as proposed by (Ehrentraut et al., 2016), tf*idf yields reasonably good results that can be clinically interpreted. Finally, the results indicate that considering different patient history lengths can increase the classification performance by 3%. A long patient history length could add noise to the dataset, while a short one could eradicate very important information. Carefully studying the appropriate window length depending on the ADE of interest is very important as it can provide medically relevant predictions.

A limitation of this study is the formulation of the ADE positive and negative groups. Although the positive groups are based on the study by (Stausberg and Hasford, 2011) the negative cases seem tightly close to the positive ones. Someone could argue that as some ADE codes are very similar to each other they can be used interchangeably by medical experts. Moreover, another limitation is the distribution of the positive and negative examples. In some datasets the distribution of the positive examples is far less than the one of the negative examples, causing lower predictive performance.

For future work we would like to investigate other ways of defining the control and test groups for the ADE examples. Furthermore, we would like to incorporate all structured and unstructured features in the learning process; we believe that not only it will improve the model performance but it will also shed light in ADE signalling. A natural extension of this paper would be to implement more recent NLP techniques as well as word-embeddings and evaluate them on the ADE problem. We plan to use decompounding of words to see if the performance of our algorithms will improve analysing the decompounded elements. Lastly, an extension would be to dynamically adjust the window sizes for each patient or ADE studied.

7. Conclusion

This paper focused on utilizing textual features using different word sequences and patient history lengths to predict ADEs from EHRs. We demonstrated the importance of incorporating in the machine learning process clinical text, as this textual source are very informative towards ADE prediction. NLP techniques can be utilized to meet the challenges posed by narrative data and provide meaningful predictions.

Acknowledgements

We would like to thank professor Panos Papapetrou for insightful advices regarding the machine learning set up.

Bibliographical References

Bagattini, F., Karlsson, I., Rebane, J., and Papapetrou, P. (2019). A classification framework for exploiting sparse multi-variate temporal features with application to adverse drug event detection in medical records. *BMC Medical Informatics and Decision Making*, 19(1):7, 12.

Bamba, M. and Papapetrou, P. (2019). Mining Adverse Drug Events Using Multiple Feature Hierarchies and Patient History Windows. In *Proceedings of the Workshop* *on Data Mining in Biomedical Informatics and Healthcare DMBIH'19 in conjunction with IEEE International Conference on Data Mining (ICDM'19), Beijing*.

Carlberger, J., Dalianis, H., Hassel, M., and Knutsson, O. (2001). Improving precision in information retrieval for swedish using stemming. In *Proceedings of the 13th Nordic Conference of Computational Linguistics (NODALIDA 2001)*.

Dalianis, H., Henriksson, A., Kvist, M., Velupillai, S., and Weegar, R. (2015). HEALTH BANK–A Workbench for Data Science Applications in Healthcare. In *Proceedings of the CAiSE-2015 Industry Track co-located with 27th Conference on Advanced Information Systems Engineering (CAiSE 2015), J. Krogstie, G. Juel-Skielse and V. Kabilan, (Eds.), Stockholm, Sweden, June 11, 2015, CEUR, Vol-1381*, pages 1–18.

Ehrentraut, C., Ekholm, M., Tanushi, H., Tiedemann, J., and Dalianis, H. (2016). Detecting hospital-acquired infections: a document classification approach using support vector machines and gradient tree boosting. *Health informatics journal*, 24(1):24–42.

Eriksson, R., Jensen, P. B., Frankild, S., Jensen, L. J., and Brunak, S. (2013). Dictionary construction and identification of possible adverse drug events in Danish clinical narrative text. *Journal of the American Medical Informatics Association : JAMIA*, 20(5):947–53.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 6.

Harpaz, R., Haerian, K., Chase, H. S., and Friedman, C. (2010). Mining electronic health records for adverse drug effects using regression based methods. In *the 1st ACM International Health Informatics Symposium*, pages 100–107. ACM.

Henriksson, A., Zhao, J., Boström, H., and Dalianis, H. (2015). Modeling electronic health records in ensembles of semantic spaces for adverse drug event detection. In *Proceedings - 2015 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2015*, pages 343–350. Institute of Electrical and Electronics Engineers Inc., dec.

Henriksson, A. (2015). Representing Clinical Notes for Adverse Drug Event Detection. pages 152–158. Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis / [ed] Cyril Grouin, Thierry Hamon, Aurélie Névéol, Pierre Zweigenbaum, Association for Computational Linguistics, 2015, s. 152-158.

Karimi, S., Wang, C., Metke-Jimenez, A., Gaire, R., and Paris, C. (2015). Text and data mining techniques in adverse drug reaction detection. *ACM Computing Surveys (CSUR)*, 47(4):56.

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1137–1145.

LePendu, P., Iyer, S. V., Bauer-Mehren, A., Harpaz, R., Mortensen, J. M., Podchiyska, T., Ferris, T. A., and Shah, N. H. (2013). Pharmacovigilance using clinical notes. *Clinical Pharmacology and Therapeutics*, 93(6):547–555, jun.

Manning, C. D., Manning, C. D., and Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT press.

Melton, G. B. and Hripcsak, G. (2005). Automated detection of adverse events using natural language processing of discharge summaries. *Journal of the American Medical Informatics Association*, 12(4):448–457.

NIDDK. (2019). Aplastic Anemia & Myelodysplastic Syndromes. `https://www.niddk.nih.gov/health-information/blood-diseases/aplastic-anemia-myelodysplastic-syndromes`, Accessed: 2019-11-21.

Park, M. Y., Yoon, D., Lee, K., Kang, S. Y., Park, I., Lee, S.-H., Kim, W., Kam, H. J., Lee, Y.-H., Kim, J. H., and Park, R. W. (2011). A novel algorithm for detection of adverse drug reaction signals using a hospital electronic medical record database. *Pharmacoepidemiology and Drug Safety*, 20(6):598–607.

Schütze, H., Manning, C. D., and Raghavan, P. (2008). Introduction to information retrieval. In *Proceedings of the international communication of association for computing machinery conference*, page 260.

Stausberg, J. and Hasford, J. (2011). Drug-related admissions and hospital-acquired adverse drug events in Germany: A longitudinal analysis from 2003 to 2007 of ICD-10-coded routine data. *BMC Health Services Research*, 11.

Van Rijsbergen, C. (1979). *Information Retrieval*. Butterworth & Co. Accessed 2018-01-11.

Wang, X., Hripcsak, G., Markatou, M., and Friedman, C. (2009). Active Computerized Pharmacovigilance Using Natural Language Processing, Statistics, and Electronic Health Records: A Feasibility Study. *Journal of the American Medical Informatics Association*, 16(3):328–337, May.

Weiskopf, N. G., Hripcsak, G., Swaminathan, S., and Weng, C. (2013). Defining and measuring completeness of electronic health records for secondary use. *Journal of Biomedical Informatics*, 46(5):830–836, Oct.

Zhao, J., Henriksson, A., and Böstrom, H. (2014). Detecting Adverse Drug Events Using Concept Hierarchies of Clinical Codes. pages 285–293, 9.

Zhao, J., Henriksson, A., Asker, L., and Boström, H. (2015a). Predictive modeling of structured electronic health records for adverse drug event detection. *BMC Medical Informatics and Decision Making*, 15(Suppl 4):S1.

Zhao, J., Henriksson, A., and Boström, H. (2015b). Cascading adverse drug event detection in electronic health records. In *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–8. IEEE, 10.

Zhao, J., Henriksson, A., Kvist, M., Asker, L., and Boström, H. (2015c). Handling Temporality of Clinical Events for Drug Safety Surveillance. *Annual Symposium proceedings. AMIA Symposium*, 2015:1371–80.