Leaving no Stone Unturned: Using Machine Learning Based Approaches for Information Extraction from Full Texts of a Research Data Warehouse

Johanna Fiebeck[1[0000-0001-7068-1508]], Hans Laser[1[0000-0002-0958-8600]], Hinrich B. Winther[2[0000-0001-6283-8042]], Svetlana Gerbel[1[0000-0003-3430-5294]]

[1] Centre for Information Management, Hannover Medical School, Carl-Neuberg-Str. 1, 30625 Hannover, Germany
[2] Institute for Diagnostic and Interventional Radiology, Hannover Medical School, Carl-Neuberg-Str. 1, 30625 Hannover, Germany
fiebeck.johanna@mh-hannover.de, gerbel.svetlana@mh-hannover.de

Abstract. Data in healthcare and routine medical treatment is growing fast. Therefore and because of its variety, possible correlation within these are becoming even more complex. Popular tools for facilitating the daily routine for the clinical researchers are more often based on machine learning (ML) algorithms. Those tools might facilitate data management, data integration or even content classification. Besides commercial functionalities, there are many solutions which are developed by the user himself for his own, specific question of research or task. One of these tasks is described within this work: qualifying the Weber fracture, an ankle joint fracture, from radiological findings with the help of supervised machine learning algorithms. To do so, the findings were firstly processed with common natural language processing (NLP) methods. For the classifiying part, we used the bags-of-words-approach to bring together the medical findings on the one hand, and the metadata of the findings on the other hand, and compared several common classifier to have the best results. In order to conduct this study, we used the data and the technology of the Enterprise Clinical Research Data Warehouse (ECRDW) from Hannover Medical School. This paper shows the implementation of machine learning and NLP techniques into the data warehouse integration process in order to provide consolidated, processed and qualified data to be queried for teaching and research purposes.

Keywords: Clinical Research Data Warehouse, Machine Learning, Text Mining, Data Science, Unstructured Data, Secondary Use, Radiology, NLP

1 Introduction

Medical records, pathology and radiology findings or medication are often available in an unstructured form. Relevant information is therefore not always described in concrete fields, but mostly in free text form. Drawing inferences on disease progressions, processes or statistics for quality assurance are difficult to extract from this

information. The structure of the texts differs within departments and partly between findings. Machine Learning (ML) methods can be used to solve this problem. Frequent data mining tasks in radiology include [1]:

- Automated derivation of numbers for defined instances or from finding results (feature extraction) from the unstructured text [2]
- Information enrichment of structured data by feature extraction
- Text analysis using controlled terminologies, in radiology (mainly the terminology RadLex [1, 3])
- Classification and clustering, e.g. to identify patient cohorts (selection of patients with similar clinical pictures) [4]

At the Hannover Medical School (MHH) the radiological findings are captured in the Radiology Information System (RIS) in free text form but are divided into individual, predefined sections.

In this study, we used these semi-structured findings data integrated into the MHH Enterprise Clinical Research Data Warehouse (ECRDW). The ECRDW of the MHH is an interdisciplinary data integration and analysis platform for research-relevant issues. In the clinical-university sector, a data warehouse based technology, serves to consolidate data routinely generated in health care for secondary use purposes [5, 6]. The typical use cases of a clinical data warehouse (CDW) include:

- Patient screening for clinical trials
- Epidemiological estimations
- Validation of data in research databases and their data enrichment with the aim of quality improvement in research tasks
- Development of decision support approaches for specific research questions

In order to make these findings available for queries, a method for data cleansing and data processing is to be developed and implemented within the standard ETL (Extraction, Transformation, and Loading) process of the ECRDW.

Our research question is to locate radiology findings that refer to the so-called Weber fracture, an ankle fracture, in order to be able to analyze the corresponding X-ray images or to make them available for teaching courses. To do so, the findings are to be preprocessed in a structured manner with the aid of natural language processing (NLP) methods and to classify the records with ML algorithms. We decided to use ML methods because the diagnosis often is not exactly named in the text. Thus, a simple full text search will not find all relevant results or will also find negating results. Additionally search for possible synonyms is required. By using ML techniques, we also included some report metadata as features, such as radiology service group and department.

This paper is divided into the typical sections material and methods, results, and discussion and conclusion. In material and methods the necessary steps for the preprocessing of the relevant data and the structure of the ML pipeline are described. In the section results the particularities of the original data set are first summarized and then the resulting selection of a suitable algorithm for ML with corresponding metrics

explained. Subsequently, the result of the prediction of the Weber fracture and the implementation of the process into the ETL process are outlined.

2 Material and Methods

2.1 Accessing the Data via a Data Warehouse Plattform

In order to develop appropriate methods for information retrieval technology and data of the MHH ECRDW was used. The ECRDW is based on the Microsoft (MS) SQL Server Stack. The basis for the machine learning are radiological findings (with additional metadata), which were joined with ICD10-GM[1] diagnosis codes. The metadata and diagnosis codes were used to select necessary features and annotate the training data on the one hand and to access the medical records for prediction on the other. A total of 2,000 medical findings were identified for training and 17,354 medical findings were predicted on the basis of the following diagnosis codes:

- S82 as fracture of the lower leg, including the upper ankle joint
- S82.5 as fracture of the inner ankle
- S82.6 as fracture of the outer ankle
- S92 as fracture of the foot (except upper ankle)

The findings from the RIS were integrated into the ECRDW via HL7 by using the MS integration services. By doing so, the whole semi-structured finding text is integrated as a full text, while the findings are split into four separated fields: “Klinische Angaben“ (engl. “clinical data”), “Fragestellung“ (engl. “clinical situation”), “Befund“ (engl. “finding”), ”Beurteilung“ (engl. “assessment”). Within this text the headings are displayed by pseudo-html tags like /.br.//. For re-separating the fields, we developed a regular expression (regex) term addressing these tags. By doing so, this regex also may search for further headings which might be feasible for integrating in clinical routine. Thus, the final regex is:

(“\S\.br\S\w+[:]\S[.]\w+\S”)

The regex and the NLP pipeline was implemented by using using the natural language toolkit Python NLTK for general text processing [8, 9].

The NLP pipeline for splitting the text into their pre-defined sections includes several steps such as:

- Loading the data
- Hard-coded misspelling-cleansing
- Extracting the headings with the regex and export the top 10-headings
- Splitting the text according to their predefined headings into predefined fields

[1] ICD-GM: “International Classification of Diseases, German Modification“is the official classification for diagnoses in outpatient and inpatient health care in Germany.

2.2 Machine Learning Pipeline for Classification

The ML training data consists of 2,000 randomly selected radiology findings. The potential radiology findings have already been selected by ICD codes (range of lower leg injuries). To create a binary classification, the full text of the ICD diagnosis was searched for the keyword "Weber". In addition, the service type "ankle joint", which was documented in the metadata of the radiological findings, served as a further criterion to create the positive class.The negative class was created by using radiological findings not having these specific attributes.

The findings then were prepared in a machine learning pipeline, consisting of a text preprocessing part and a classification part. The text preprocessing steps for both the training data and the data to be classified were carried out with the Python NLTK packages as well. These included tokenizing, removing stopwords, transformation into bags of words and converting the bags of words into a Python Pandas DataFrame.

For dimension reduction within the bags of words, we inserted the following steps and compared the results:

- Filtering dates and times using the RegexpTokenizer (optional step)
- Only the twenty-most often tokens where represented in the bags of words, using the built-in Counter function.

Afterwards, we joined the bags-of-words dataframe with metadata features: service group, service type, analysis device, operating department. The prepared training data was used for classifier training. Various ML classifier were selected for comparision: Naive Bayes Classifier, Support Vector Machines, Decision Trees and Random Forest and Logistic Regression algorithms and accessed them via the Python scikit-learn package. We chose a 10-fold cross-validation with 70/30 training/test split.

Additionally, selected algorithms were judged according their confusion matrices while predicting test data. “Unknown” medical records for prediction where read directly from the ECRDW once they are in the same ICD range as selected above.

Prediction was conducted on the medical record as well as on the section “Befund” (“medical indication”) only. In Table 1, the process of prediction is outlined.

Table 1. Steps in Machine Learning pipeline from loading and preparing the data to prediction and analysis.

Step	Description
1	Selecting and annotating training data
2	Feature selection
3	Loading unknown data for prediction
4	NLP pipeline
5	Reduction of *bags-of-words* dimensions of the "unknown" data down to the dimensions of the training data set
6	Training of the algorithm, 10-fold cross-validation (70/30 split)
7	Prediction on test data with confusion matrix
8	Classification of unknown medical records
9	Embedding the prediction results to the data

3 Results

3.1 Finding Patterns and Separating the Text

As expected, the headings of the predefined fields were found most often, but not all of them actually were filled (Fig. 1). Additionally to the predefined headings, the regex found some potential new headings, for example "Methodik" ("methods"), which is named in about 5 % of all findings. The other headings found by the regex are mainly indicating anatomical issues. By these results, one can recommend to add the "methods" as an additional section within the radiological findings.

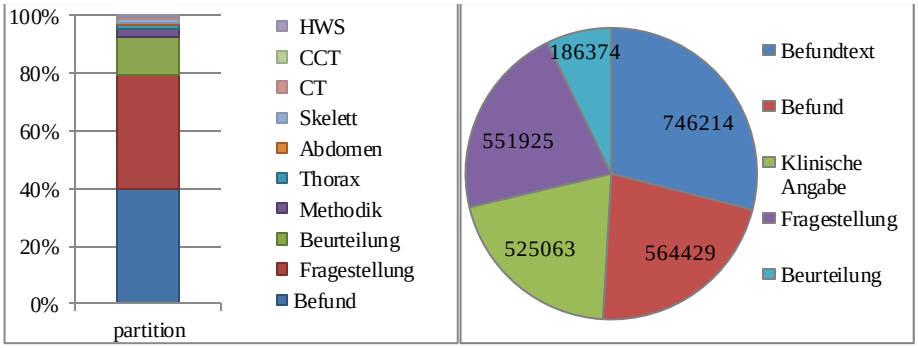

Fig. 1. Left side: The 10 most-often headings extracted from medical record texts, plotted according their occurrences. Right side: Filled section quantity by their pre-defined headings.

3.2 Finding the Appropriate Machine Learning Algorithm

After comparing the accuracies in 10-fold cross-validation of all algorithms and hyperparameter modes (Table 2), we selected two algorithms for classification of unknown findings: the standard Decision Tree and the standard Support Vector Machine (SVM) algorithm.

Additionally, we decided to add the RegExpTokenizer for dimension reduction. Although there was only minor effect in the accuracies whether the additional tokenizer was conducted or not, it helped to reduce feature dimensions drastically.

After selecting the algorithms, they were chosen to predict test data within a confusion matrix. As shown in Fig. 2, both prediction results vary seriously from each other. As expected, the Decision Tree algorithm performs slightly better than the SVM algorithm: While the correct-positive rate and the correct-negative rate of the Decision Tree is quite high (171:174), the SVM predicts a high rate of correct-positive records but a very high false-positive class (197:140).

Table 2. Results of comparing several classification algorithms accuracies in 10-fold cross-validation, 70/30 split mode. As there 1,000 positive and 1,000 negative training data, the baseline indicator is set to 50 %.

Modus	Without RegexpTokenizer		With RegexpTokenizer	
Standard SVM	54,00%	(+/- 0,08)	54,00%	(+/- 0,08)
SVM (kernel="linear", C=0.025)	54,00%	(+/- 0,08)	54,00%	(+/- 0,08)
SVM (gamma=2, C=1)	58,00%	(+/- 0,04)	58,00%	(+/- 0,04)
Standard Decision Tree	70,00%	(+/- 0,08)	70,00%	(+/- 0,08)
Decision Tree (max_depth=5)	58,00%	(+/- 0,08)	59,00%	(+/- 0,07)
Standard Random Forest	70,00%	(+/- 0,09)	70,00%	(+/- 0,09)
Random Forest (max_depth = 5, n_estimators =10, max_features=2)	50,00%	(+/- 0,04)	51,00%	(+/- 0,02)
Naive Bayes_Gaussian	54,00%	(+/- 0,07)	54,00%	(+/- 0,07)
Naive Bayes_BernoulliNB	50,00%	(+/- 0,03)	50,00%	(+/- 0,03)
Naive Bayes_MultinomialNB	55,00%	(+/- 0,08)	55,00%	(+/- 0,08)
Standard Logistic Regression	54,00%	(+/- 0,09)	54,00%	(+/- 0,09)

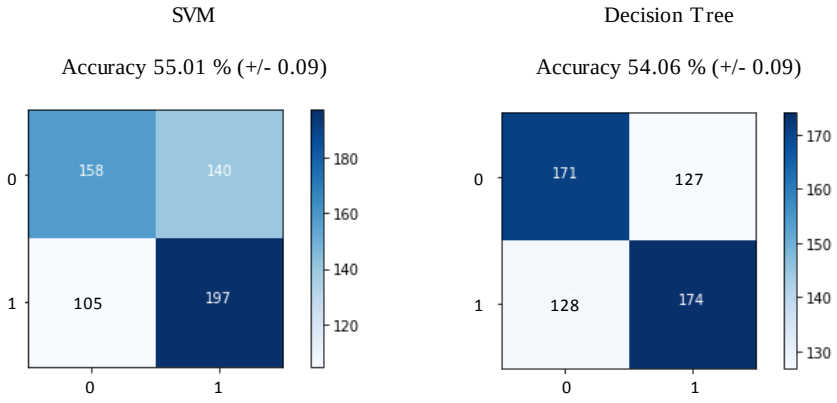

Fig. 2. Confusion matrices: Comparison of decision tree and support vector machine (SVM) algorithms for predicting test data. 0 – negative class, 1 – positive class, horizontal axis: Predicted class, vertical axis: Real class. Thus, the upper left box is the correct-negative, the lower right box the correct-positive rate.

3.3 Predicting the Weber Fracture in Medical Records

As described above, two different classification approaches were taken to predict unknown radiological findings. Fig. 3 shows the prediction results exemplarily, plotted for the radiological service types.

The most often positively-predicted records were in service group "NERMRTCRC" (neurological MRT) and "NERANGIO" (neurological angiography), which implies quite bad prediction rates for the classifiers.

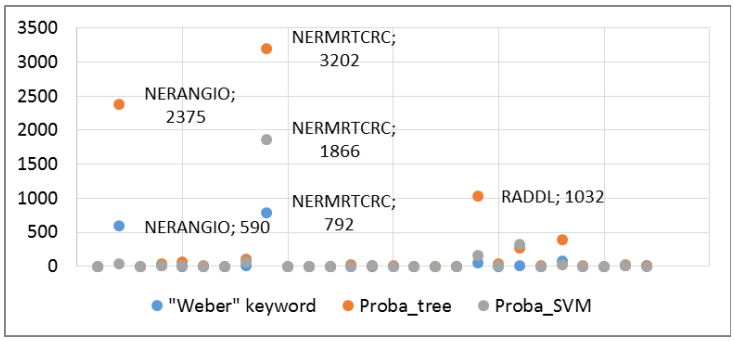

Fig. 3. Classification results: Absolute occurrences of medical records predicted as "Weber"-positive according to the algorithm (Proba_tree: decision tree; Proba_SVM: support vector machine) in radiological service groups at the MHH.

3.4 Implementation into the ETL Process

We implemented the developed pipeline as a script component within the data integration task that loads the HL7 messages from the RIS. The result of the script component enriches the ECRDW core repository by splitting the findings in defined columns and by creating further information from the classification task that can be used as additional features when it comes to querying the data for Weber fracture.

4 Discussion and Conclusion

In this work the Weber fracture was to be identified in radiological medical findings. We used a dataset containing pseudonymised master data of a patient, ICD10-GM diagnosis code, as well as the free text, localisation, and certainty of every diagnosis captured during the hospital stay. The dataset also contains the radiological finding, that was split into the four fields, and additional metadata from the RIS (e.g. service group, service type text, analysis device, operating department, observation time). Various ML techniques were applied and compared with each other with regard to their suitability, accuracy and specificity. The results described in this paper show the basically feasibility of classifying texts using ML techniques. However, the results differ considerably depending on the chosen method.

The pre-selection of possible algorithms was based on algorithms that were used in the literature for similar questions: Naive Bayesian classifiers are used in many works when it comes to text classifications and sentiment analyses. The advantage of decision trees is their simple application and comprehensible interpretation [9]. SVM have been used in a number of studies, as soon as there were high dimensional vector spaces [10]. SVMs are used in a variety of problems, such as clustering, regression or classifications. The invaluable advantage of SVMs is that they work even if the features differ in test and training data sets. SVMs generate a higher dimensional vector space based on similar words by embedding the unknown tokens. Although the present paper does not use other features in the unknown data set for classification than in training, this approach could be tested in a further study.

The use of a regular expression for additional filtering of date and time information was not sensitive enough. As a result, dimensions were created in both the training and prediction datasets that might have been superfluous. Nevertheless, the use of a regular expression tokenizer showed a significant reduction of the dimensions. In a direct comparison of the classifiers, in which the dimensions were created with or without RegexpTokenizers, a clear change up to doubling of the accuracy was shown in some individual experiments.

In addition, an overfitting of the models must be considered: the training data set was created from findings texts of the radiology department of the MHH. Accordingly, it can be assumed that the models cannot easily be applied to applications at other universities or even other departments.

Classifications of medical texts, such as findings or doctor's letters, have become increasingly important in recent years, especially since they are increasingly digital and therefore available in machine-readable form [11].

No synonyms were considered in selecting the training data. Thus, we expected to have a positive prediction by only having the token "Weber" within the data. Surprisingly, this was not the case: records without "Weber" were also classified positively and vice versa. Based on the occurrences in the full-text search, we expect the SVM algorithm to be more accurate in its prediction, which is quite the opposite of what we expected from the 10-fold cross-validation and confusion matrix. To have this hypothesis confirmed, a domain expert has to validate the results, which we will do so in our next steps. Additionally, further work will include a self-generating dictionary of synonyms by implementing word embeddings to increase the recall.

Another promising approach would be to use a semi-supervised learning method: first training would be performed by using a little fraction of the whole data and would then be post-trained continuously by various prediction and validation rounds on real data. It would be promissing to combine the semi-supervised techniques with a word embedding.

In summary, we have shown that implementing a NLP approach into a data warehouse ETL pipeline with Python is feasible. The developed pipeline provides more flexibility for data pre-processing and data cleansing of unstructured or semi-structured information than we would have had by using the standard data integration services of MS SQL Server. Additionally, adding a data mining pipeline for a specific research question upon this data is applicable, but its power definitively relies on validated gold standard training data and the validation of the predictions provided by a clinical expert, which will be our next step. Further limitations are due to the chosen dataset: As a medical record may have more than one medical diagnosis (e.g. differences in entry diagnosis and in release diagnosis), it may be rated as well as positive or as negative.

Nevertheless, this study proves the possibility of combining ETL processes with machine learning techniques. For one's own attempt of implementing, the pipeline has to be adapted to one's own IT infrastructure, since every hospital has its own, heterogeneous infrastructure and conditions.

5 References

1. Köppen, V., Saake, G., Sattler, K.-U.: Data Warehouse Technologien. mitp. 2014, ISBN 9783826694851.
2. Tolxdorff, T., Puppe, F.: Klinisches Data Warehouse. Informatik-Spektrum 2016, 39, 233-237, doi:10.1007/s00287-016-0968-3.
3. Zapletal, E., Bibault, J.-E., Giraud, P., Burgun, A.: Integrating Multimodal Radiation Therapy Data into i2b2. Appl. Clin. Inform. 2018, 09, 377-390, doi:10.1055/s-0038-1651497.
4. Dietrich, G., Krebs, J., Fette, G., Ertl, M., Kaspar, M., Störk, S., Puppe, F.: Ad Hoc Information Extraction for Clinical Data Warehouses. Methods Inf. Med. 2018, 57, e22-e29, doi:10.3414/ME17-02-0010.
5. Kharat, A., Singh, A., Kulkarni, V., Shah, D.: Data mining in radiology. Indian J. Radiol. Imaging 2014, 24, 97, doi:10.4103/0971-3026.134367.

6. Do, B. H., Wu, A. S., Maley, J., Biswal, S.: Automatic retrieval of bone fracture knowledge using natural language processing. J. Digit. Imaging 2013, 26, 709-13, doi:10.1007/s10278-012-9531-1.
7. Bird S, Klein E, Loper E. Natural Language Processing with Python. Sebastopol: O'Reilly Media Inc, ISBN 978-0-596-51649-9.
8. Perkins J. Python Text Processing with NLTK 2.0 Cookbook. Birmingham: Packt Publishing. 2010, ISBN 9781849513609.
9. Daumke, P., Simon, K., Paetzold, J., Marwede, D., Kotter, E.: Data-Mining in radiologischen Befundtexten. RöFo - Fortschritte auf dem Gebiet der Röntgenstrahlen und der Bildgeb. Verfahren 2010, 182, WS117_3, doi:10.1055/s-0030-1252462.
10. Kavuluru, R., Rios, A., Lu, Y.: An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records. Artif. Intell. Med. 2015, 65, 155-166, doi:10.1016/j.artmed.2015.04.007.
11. McNutt, T. R., Moore, K. L., Quon, H.: Needs and Challenges for Big Data in Radiation Oncology. Int. J. Radiat. Oncol. Biol. Phys. 2016, 95, 909-915, doi:10.1016/j.ijrobp.2015.11.032.