D. Hayn et al. (Eds.)

doi:10.3233/978-1-61499-524-1-35

Development of Text Mining Based Classification of Written Communication within a Telemedical Collaborative Network

Katharina GRUBER[a,1], Robert MODRE-OSPRIAN[a], Karl KREINER[a], Peter KASTNER[a] and Günter SCHREIER[a]
[a] *AIT Austrian Institute of Technology, Austria*

Abstract. Chronic diseases like Heart Failure are widespread in the ageing population. Affected patients can be treated with the aid of a disease management program, including a telemedical collaborative network. Evaluation of a currently used system has shown that the information of the textual communication is of pivotal importance for the collaboration in the network. Thus, the challenge is to make this unstructured information useable, potentially leading to a better understanding of the collaboration so as to optimize the processes. This paper presents the setup of an analysis pipeline for processing textual information automatically, and, how this pipeline can be utilized to train a model that is able to automatically classify the written messages into a set of meaningful task and status categories.

Keywords. Telemedicine, collaboration, text mining, classification, mHealth

1. Introduction

Chronic diseases like Heart failure (HF) are becoming more and more a serious public health problem and are responsible for a large share of healthcare costs and frequent hospitalizations and an increasing number of deaths [1]. Different disease management programs aim to decrease the impacts of heart failure. The healthcare provider Tiroler Landeskrankenanstalten GmbH (TILAK) started a collaborative HF network named HerzMobil Tirol in April 2012, which combines mHealth/telemedicine; nurse-led patient education and home visits embedded in a network of dedicated physicians in private practices [2]. One of the key aspects of HerzMobil Tirol is the interdisciplinary collaboration across institutions. This collaboration was supported by a web-based telemonitoring software in which the actors of the network documented their interactions with patients by editing patient-specific notes.

In the proof-of-concept phase of HerzMobil Tirol, which was finished in September 2014, 47 patients were included and 1,564 notes (55,737 words) containing patient-specific content have been entered by 10 physicians in private practices, 4 clinic physicians, 3 nurses and 1 coordinator. The participating physicians and nurses of the network were interviewed about strengths and weaknesses of the project. Most

[1] Corresponding Author: Katharina Gruber, AIT Austrian Institute of Technology, Reininghausstraße 13/1 8020 Graz, E-Mail: Katharina.Gruber.fl@ait.ac.at

interviewed persons emphasized that the textual information is useful to get a comprehensive view on the patients' medical and social situations.

Unstructured textual information is not only recorded in collaborative networks. The American project *The Health Story Project* [3] estimated that in the US healthcare system 1.2 billion documents are produced annually, in which 60% of the documents exclusively contains unstructured information. Further studies indicate [4] that the narrative display of functional information is of high importance for the physicians in clinical decision-making.

Therefore a method for automated analysis of the textual communication would be of interest. The present paper deals with the construction of a reusable pipeline to process the unstructured textual information and to generate a classifier for specific tasks or statuses. In a second step this pipeline was applied to real-world data from the HerzMobil Tirol network.

2. Methods

Since the textual information contained also different personal data (e.g. names, telephone numbers and addresses) about the patients, their relatives and other persons like involved physicians, the first required preprocessing step supported by the pipeline was de-identification (anonymisation respectively pseudonymisation). Pseudonymisation of patient names in the text was done by substituting the names of the patients with their patient ID. Anonymisation of all the other personal data was facilitated as follows: Names of physicians were substituted by their first letters by using an automatically generated dictionary of names. This dictionary contained all the names of the stakeholders in the network. Telephone numbers and mailing addresses were replaced by means of a regular expression. To save time, the actors used a lot of abbreviations. These abbreviations were normalized with the aid of a list which consist of 85 typical abbreviations and contracted words. The next problem of the unprocessed textual communication was that the notes were often long and consisted of many different topics. Thus to simplify the categorization of the notes, the notes were separated in so called note snippets using following separators: ".","?", "!" and ";".

After the preprocessing a classifier was trained. With the aid of word clouds certain tasks and statuses were identified that might be of interest for the actors in the network. Exactly six types of tasks and four types of statuses were specified. The tasks were: *medical adaptation*, *technical helpdesk*, *organizational task*, *home visit*, *patient deactivation* and *vacation replacement*. *General state of health*, *home environment*, *independence* and *compliance* were specified as the relevant statuses. The Support Vector Machine (SVM), a supervised learning method, was used because of its promising results for textual classification [5]. The corresponding annotation of the notes and feature selection are described in more detail in the following sections 2.1 and 2.2.

The generated annotated data set and the selected feature set were then used for the training of a classifier by applying data from HerzMobil Tirol. Finally, the created model was evaluated by ten-fold cross validation and the scorer values precision, recall and F-measure values were calculated.

The pipeline was implemented using the open source analytics software KNIME [6], because of the large amount of existing machine learning tools and the available plugin for text processing.

2.1. Annotation of Data Set

The application of the word clouds resulted in ten interesting categories (four types of statuses and six types of tasks). To support annotating the notes, a description and some examples for every category were defined in so called annotation guides for both specified tasks and specified statuses. The design of the annotation guides was such that the procedure was easily comprehendible for people without prior knowledge of the telemedical disease management program. For the generation of an annotated data set with the aid of the annotation guides, the note snippets were *annotated* by students without prior knowledge about the telemedical disease management program. Hence, the human annotators had the same initial situation as the machine learning algorithm. Each student annotated a randomly selected set of note snippets. A gold standard was created by selecting the majority of votes for each note snippet.

2.2. Feature Selection

The features were selected based on term frequency, a process that was shown to be an easy but efficient approach in previous studies [7]. Before a feature set was selected, the following standardization steps were applied: First all the letters in the notes were *converted* to lower cases. Second, all words in the notes were tagged by a *Part-Of-Speech-Tagger* called Stanford Tagger [8] and certain tagged words were filtered (pronouns and particles). Afterwards, numbers, words with less than three characters and stop words of the German language were omitted. Certain frequently used expressions were *abstracted* to uniform terms (Table 1). All the text examples given in this paper were translated from German to English.

Table 1: The abstracted terms with their regular expression and an example

Term	Regular Expression	Example
telephone number	\d?\d{4}-?/?telephone number	0699/telephone number
medication	\d/?\d?-\d/?\d?-\d/?\d?-\d/?\d?	1-0-1/2
blood pressure	\d?\d\d/\d?\d\d	135/80
dosage	\d?x?,?\d\d?\d?\d?s?mg \d\d?\d?\d?s?mg \d\d?x\d\d?g?	25mg
time interval	\d\d?d\s	14d
date	\d\d?january\d?\d?\d?\d? one for every month	3februar2013
weight	\d\d?\d?\s?kg	76 kg
time	\d\d?\s?h	3 h

Then a so called *Bag-of-Words* (BoW) was created, i.e. a list, where all the used words (unigrams) in the notes are named and one column was added with the absolute frequency of the words. For the used combinations of two words (bigrams) the same BoW was created. The two lists were combined and sorted with a descending absolute frequency of the words. Finally a feature set was selected by using the first entries of this list with 17,448 features. The cut-off was different for each classifier and was selected empirical by comparing the calculated F-measures for different numbers of features.

3. Results

3.1. Pipeline

The pipeline contained modules for preprocessing, annotation, feature selection and evaluation (Figure 1). Nearly all of those modules except annotation of data set were implemented using the KNIME tool.

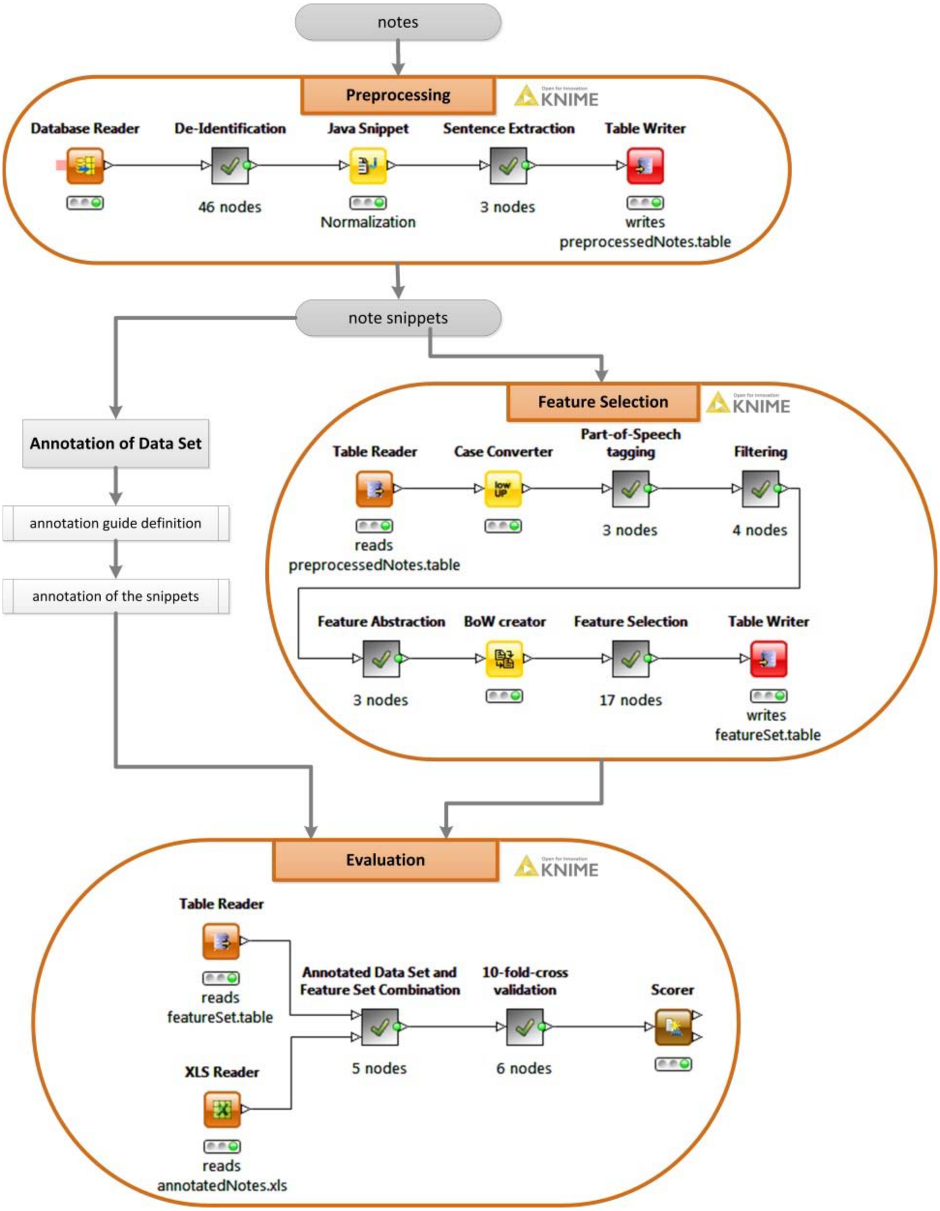

Figure 1: Pipeline for the preprocessing, annotation, feature selection and evaluation of the notes

Table 2: Examples for results of the preprocessing steps

Preprocessing step	Result
Original Note	Telephone c. with Ms. Smith and daughter Mary: Pat. has more appetite, eating much more, feels strong – WT increase with it explainable, has no leg edema; pointed out that the med. (40 mg to noon) should be confirmed.
De-Identification	Telephone c. with Ms. *patient<1458>* and daughter *M*: Pat. has more appetite, eating much more, feels strong – WT increase with it explainable, has no leg edema; pointed out that the med. (40 mg to noon) should be confirmed.
Normalization	Telephone call with *Mrs* patient<1458> and daughter M: Patient has more appetite, eating much more, feels strong – *weight* increase with it explainable, has no leg edema; pointed out that the *medication* (40 mg to noon) should be confirmed.
Sentence Extraction	*snippet1:* Telephone call with Mrs patient<1458> and daughter M: Patient has more appetite, eating much more, feels strong – weight increase with it explainable, has no leg edema *snippet 2:* pointed out that the medication (40 mg to noon) should be confirmed

We analyzed 1,564 notes with 55,737 words. The results of each single preprocessing step of an example note are shown in Table 2. The original note contained the patient name “Smith”. The de-identified note instead included only the word “patient” and the patient ID “<1458>”. At this preprocessing step the name “Mary” was also substituted with “M”. After the normalization, abbreviations like “WT” and “med.” were replaced with “weight” respectively “medication”. Finally, the result of the sentence extraction was two single note snippets (snippet 1 and snippet 2).

3.2. Annotation guidelines and annotation results

The 3,739 separated notes were then annotated by 18 students. To gain 3 votes per note snippet the notes were randomly divided into packages of 624 notes. The guidelines used by the students for annotating the statuses are shown in Table 3.

Table 3: Annotation guidelines

Status	Description	Example
General state of health	Mental and physical general state of the patient	The mother is still very tired and weak. Today mental and physical better. …feels a little bit better. The patient seems to be dying.
Home environment	Description of the personal environment of the patient; support of the family, 24-hour care	Son will support by the data transmission at the present. The new 24-hour care doesn’t speak German. Housekeeping is done mostly by the wife.
Independence	Independence of a patient by the transmission of date or by the handling of the devices	Handling of the devices without difficulty. ...transmit the data meanwhile independent Patient doesn’t cope with the health scale.
Compliance	Degree, with which the behavior of a patient corresponds with the medical or health advice (e.g. medication intake, data transmission)	The data transmission is very incomplete. … very motivated patient Patient intakes medication regular. Patient wants no home visits.

With the help of this annotation guides 3,739 note snippets have been annotated by the students. The amount of annotation varied for tasks (Table 4) from 36 of *vacation replacement* to 362 of *organizational tasks* and for statuses (Table 5) from 285 of *independence* to 1,388 of *general state of health.*

Table 4: Number of annotated note snippets for each task

Tasks	Annotated note snippets
Medication adaptation	242
Technical helpdesk	136
Organizational task	362
Home visit	153
Patient deactivation	252
Vacation replacement	36

Table 5: Number of annotated note snippets for each status

Status	Annotated note snippets
General state of health	1,388
Home environment	298
Independence	285
Compliance	295

The number of notes for each student to annotate was chosen such that every snippet was annotated at least by three students. The annotations of all students were joined, resulting in a gold standard category if at least two out of three students identified the same category in the note snippets. The complete list of categorized note snippets was then used to train and test the SVM classifier.

3.3. Evaluation results

Using a 10-fold cross validation, the annotated data set was divided into 10 sets of the same size and within 10 iterations the classifier was trained with 9 datasets and tested with the left out. For each iteration precision, recall and F-measure were calculated. The mean values for each category are shown in Table 6 and Table 7.

The calculated values were compared with the results of a Keyword Classifier as a benchmark. This kind of classification depends only on three keywords for each category. The 3-keyword classifier of the organizational task for example always classified a note as organizational task, when the note contained one of the words “medical report”, “appointment” or “reachable”. As can be seen in Table 6, in eight out of 10 cases the simple keyword classifier delivered lower F-measure values than the SVM classifier. The bold marked numbers in the Table 6 and 7 indicate the best F-Measure value for a given category.

Table 6: Scorer results of the binary classifiers for task derivations

Tasks	3-Keyword Classifier			SVM Classifier		
	Precision	Recall	F-Measure	Precision	Recall	F-Measure
Home Visit	0.39	0.86	0.54	0.73	0.73	**0.73**
Patient Deactivation	0.30	0.23	0.26	0.60	0.55	**0.58**
Vacation Replacement	0.34	0.64	**0.46**	0.47	0.28	0.35
Medication Adaptation	0.37	0.49	**0.42**	0.44	0.40	**0.42**
Organizational Task	0.41	0.29	0.34	0.43	0.37	**0.40**
Technical Helpdesk	0.18	0.20	0.19	0.39	0.37	**0.38**

Table 7: Scorer results of the binary classifiers for status derivations

Status	3-Keyword Classifier			SVM-Classifier		
	Precision	Recall	F-Measure	Precision	Recall	F-Measure
General Health Status	0.91	0.12	0.21	0.80	0.71	**0.76**
Independence	0.92	0.99	**0.95**	0.56	0.40	0.47
Home Environment	0.18	0.30	0.22	0.51	0.37	**0.43**
Compliance	0.23	0,04	0.06	0.28	0.23	**0.25**

4. Discussion

The increasing amount of unstructured textual information in medical documentation calls for the provision of additional information in an automatic way to involved healthcare providers like physicians and nurses. In this work a pipeline was designed which supports this process of automatic processing, including de-identification, normalization and sentence extraction of the textual information. Furthermore, the pipeline generates a model for classification of notes into specified tasks or statuses.

The results show that some categories (e.g. general health status and home visit) can be extracted very well. Other categories (e.g. technical helpdesk and compliance) could not be determined reliably. Reasons for this could be: differences in the number of available annotated notes. As an example, for the status "General Health Status" we achieved the best results and this category has by far the most annotated note snippets (1,388 out of 3,739). For comparison: all other tasks or statuses consist of only 229 annotated note snippets on average. The classifier for "Compliance" delivers the lowest scorer result and a manual assessment of a subset of the textual communication has shown that this category is ambiguous in the sense that - for example - the two different notes "The patient did not want a home visit" and "The patient takes his medication regularly" both contain some information about the compliance of this patient.

The manual check was also indicating that the extraction of tasks is in general quite difficult, because of the similarity between open tasks and closed tasks. A source of errors here was that the annotation guide was primarily defined for annotating open tasks.

The 3-word classifier of the category "Independence" reached an F-Measure value of 0.95, while the SVM-classifier achieved only a score of 0.47. The three keywords for this category were "self-dependent", "tried" and "realized". For this category the keywords were so characteristic that the SVM-classifier with the low amount of annotated data set notes was not able to agree with the gold standard in a similar way.

But in most of the categories the SVM-classifier reached higher F-measure values than the simple 3-keyword classifier. Task "Patient Deactivation" with an F-measure value of 0.58 for the SVM-classifier and an F-measure value of 0.26 for the keyword-classifier is only one representative example for the better performance of the SVM-classifier.

5. Conclusion and outlook

We implemented a pipeline to process unstructured textual information and created a model that is able to classify textual notes into actionable information with relevance for a collaborative network. Most preprocessing steps were done automatically, but some manual interactions (like the creation of dictionaries or the annotation of the snippets) were needed.

This work shows that it is possible to find classifiers in the unstructured information of a telemedical collaborative network. The results are limited due to the small amount of patients and annotated notes. But applied to a running network with the increase of textual information the model can be regularly refined based on a new annotated data set and a new selection of features. Future work will extend the pipeline with additional linguistic analysis as for example was done in [9] using tools like T-LAB [10].

In the future the classifiers may be used in a productive collaboration network to determine specific characteristics of a patient and derivate certain treatments. For example if for a patient a superior amount of compliance statements were documented it is an indicator that a patient needs a consultation or an additional disease education. On the other hand the results of the classifiers might help to improve the quality of the network as it could provide overall statistics about tasks of the stakeholders and status information about the patients. This information can then be used in further education and training of the stakeholders of the network. The application of the pipeline in other telemonitoring networks will show, whether it can be adapted to other chronic diseases, like diabetes.

References

[1] K. Guha and T.McDonagh, Heart Failure Epidemiology: European Perspective, *Current Cardiology Reviews* **9** (2013), 23-127.

[2] A. Von der Heidt, E. Ammenwerth et al., HerzMobil Tirol network: rationale for and design of a collaborative heart failure disease management program in Austria, *Wien klinische Wochenschrift 2014 Nov* **126** (2014), 734-41.

[3] http://www.healthstory.com/, last access: 22.01.2015

[4] C. Weir, R. Dunlea et al., Comparing Narrative versus Numerical Display of Functional Information: Impact of Sense-Making, *European Federation for Medical Informatics and IOS Press* (2014), 609-613.

[5] T. Joachims, Text categorization with Support Vector Machines: Learning with many relevant features, *Lecture Notes in Computer Science Volume* **1398** (2005), 137-142.

[6] M.R Berthold, N. Cebron et al, KNIME - The Konstanz Information Miner, *SIGKDD Explorations* **11** (2009), 26–31.

[7] A.A. Argaw, A. Hulth, General-Purpose Text Categorization Applied to the Medical Domain, *Department of Computer an System Sciences – Research Report Stockholm University* **16** (2007).

[8] K. Toutanova, D. Klein et al, Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network, *Proceedings of HLT-NAACL 2003*, 252-259.

[9] N. Y. Osman, C. Schonhardt-Bailey, J. L. Walling, J. T. Katz and E. K. Alexander, Textual analysis of internal medicine residency personal statements: themes and gender differences, *Medical Education* **49** (2014), 93-102.

[10] M. Cortina, S. Tria, Triangulating Qualitative and Quantitative Approaches for the Analysis of Textual Materials: An Introduction to T-Lab, *Social Science Computer Review* **32** (2014), 561-568.