Navigating Healthcare Through Challenging Times
D. Hayn et al. (Eds.)

doi:10.3233/SHTI210104

Natural Language Processing for Free-Text Classification in Telehealth Services: Differences Between Diabetes and Heart Failure Applications

Fabian WIESMÜLLER[a,b,1], Dieter HAYN[a,c], Karl KREINER[a], Bernhard PFEIFER[d,e], Gerhard PÖLZL[f], Peter KASTNER[a], and Günter SCHREIER[a]
[a] *AIT Austrian Institute of Technology GmbH, Graz, Austria*
[b] *FH Joanneum, Graz, Austria*
[c] *Ludwig Bolzmann Institute for Digital Health and Prevention, Salzburg, Austria*
[d] *AIT Austrian Institute of Technology GmbH, Hall in Tirol, Austria*
[e] *Landesinstitut für Integrierte Versorgung – LIV Tirol, Innsbruck, Austria*
[f] *Department of Internal Medicine III, Cardiology and Angiology, Medical University Innsbruck, Austria*

Abstract. Telehealth services for long-term monitoring of chronically ill patients are becoming more and more common, leading to huge amounts of data collected by patients and healthcare professionals each day. While most of these data are structured, some information, especially concerning the communication between the stakeholders, is typically stored as unstructured free-texts. This paper outlines the differences in analyzing free-texts from the heart failure telehealth network *HerzMobil* as compared to the diabetes telehealth network *DiabMemory*. A total of 3,739 free-text notes from *HerzMobil* and 228,109 notes from *DiabMemory*, both written in German, were analyzed. A pre-existing, regular expression based algorithm developed for heart failure free-texts was adapted to cover also the diabetes scenario. The resulting algorithm was validated with a subset of 200 notes that were annotated by three scientists, achieving an accuracy of 92.62%. When applying the algorithm to heart failure and diabetes texts, we found various similarities but also several differences concerning the content. As a consequence, specific requirements for the algorithm were identified.

Keywords. Heart failure, diabetes, natural language processing, telehealth

1. Introduction

1.1. Background

A rising prevalence of chronic diseases like heart failure and diabetes mellitus is a major concern for healthcare systems, since they are not only accompanied by a high burden for the patients but also by high costs [1,2]. To counteract these medical and financial

[1] Corresponding Author: Fabian WIESMÜLLER, AIT Austrian Institute of Technology GmbH, Graz, Austria, E-Mail: fabain.wiesmueller.fl@ait.ac.at

issues, different disease management programs were implemented in recent years. E.g., in 2012, a heart failure telehealth system called *HerzMobil* was launched in Tyrol, Austria, by the healthcare provider Tiroler Landeskrankenanstalten GmbH [3]. In 2010, a diabetes management program called *DiabMemory* was started by the Versicherungsanstalt für öffentlich Bedienstete, Eisenbahnen und Bergbau (BVAEB). Both systems are based on a telehealth platform developed by the AIT Austrian Institute of Technology [4]. In both programs, patients are equipped with a specific smartphone app for submitting data concerning their chronic disease to a backend system, where healthcare professionals review the data. While in *HerzMobil*, the focus is set on cardiac parameters such as blood pressure and heart rate [3], participants of the *DiabMemory* program primarily upload data like blood sugar values, insulin administration and their daily food intake [5]. Both systems store these data in a structured format. However, it is possible for the participants (patients and healthcare professionals) of the *HerzMobil* as well as the *DiabMemory* network, to upload additional information in free-text format. This function is used as an additional communication channel between participants. In the *HerzMobil* network, this interaction is mostly used between physicians, nurses and other healthcare professionals, whilst in the *DiabMemory* network, most free-text messages derive from the communication between patients and healthcare professionals. Since these free-texts often contain important information, analyzing them in a structured way, based on natural language processing techniques, would provide additional value.

1.2. State-of-the-art

The amount of unstructured clinical notes in modern healthcare systems is already very high and increases rapidly [6]. Therefore, previous studies have already made efforts towards an analysis of free-text clinical notes. A study by Hebal et al. [7] compared methods for automated and manual data extraction from clinical notes. However, the scope of this study was rather small with 149 individual notes and, additionally, these notes had to follow a specific template for the automated extraction. Another approach for analyzing clinical notes is to detect named entities by using a domain ontology [8]. Amongst other techniques it is also possible, to use systems based on Support Vector Machines or Statistic Language Modeling [9,10].

1.3. Objectives

In this paper, automated classification of clinical free-text notes from a heart failure telehealth network and a diabetes telehealth network are compared. The paper focuses on a) differences in the frequency of free-text categories in the two corpuses and b) on implications on the natural language processing algorithm when applying tools developed for a heart failure application to free-texts from a diabetes telehealth service.

2. Methods

2.1. Corpus

The datasets for the analysis stem from the two telehealth systems: *HerzMobil* and *DiabMemory*. The *HerzMobil* dataset contained 1,564 notes with 55,737 tokens. These

1,564 notes were tokenized so that every resulting text snippet represented a separate sentence, resulting in 3,739 notes on a sentence layer.

In the *DiabMemory* network, the clinical notes can be further divided into *comments*, which were written by the patients themselves, and *feedbacks* which represent a response from a physician. The physicians' respond must not necessarily concern a patient's free-text comment, but it can also refer to an uploaded measurement or any other issue. For the evaluation of the *DiabMemory* system, 71,241 comments written by patients and 156,868 free-text feedbacks from physicians were available.

The clinical notes from both systems were exclusively written in German language, which had to be considered during keyword and regular expression creation in chapter 2.3.

Ethics approval for our analyses was granted by the Ethikkommission der Medizinischen Universität Innsbruck (vote nr. AN2015-0131 35/4.2 374/5.10(4092a)) and the Niederösterreichischen Ethikkommission (vote nr. GS1-EK-4/534-2018).

2.2. Natural language processing

Our algorithm was based on an approach developed by Gruber et al. [11, 12] in 2015, based on *HerzMobil* notes. Adaptions of these algorithms and of the categories used for classification were applied. The final algorithm consisted of pre-processing, stemming, and filtering based on regular expressions, which are described in detail in the following. Details concerning adaptions as compared to Gruber et al. are described in chapter 3.

2.2.1. Pre-processing

Since the notes contained personal information about patients and healthcare professionals, all personal data had to be de-identified using a rule-based algorithm during preprocessing, including data like names, age, address, etc., Additionally, in order to prevent linkage of the de-identified data with telehealth data especially in case of abnormal values mentioned in the free-texts, all numeric values were masked, resulting in e.g. *<Digit> km of running* instead of an actual distance, or *hba1c <Digit>*.

A previously developed algorithm by Wiesmueller et al. was applied to tag all notes containing time-related information like e.g. *gestern (yesterday)* [13]. Even though this algorithm was developed using the notes from the *HerzMobil* network, it was not yet available during the work by Gruber et al.

2.3. Regular expressions and keywords

A set of keywords which were analyzed by regular expressions was applied to all free-text notes. Each expression corresponded to one category. E.g., free-texts containing the word *krank (sick)* were classified as category *state of health.* The keywords for each category were stored in separate plain text files for further use in a Python algorithm.
In general, most regular expressions had the following structure

(?:^|\s)expression(?:\s|$).

The enclosing expressions were start and stop parameters which escaped the expression in the case of e.g. a blank character or a new line. As compared to Gruber et al., the number of keywords was reduced by simply using the stem of a word and appending a

leading and trailing *\w**. An example, where this technique was useful, is the expression *(?:^|\s)\w*schlecht\w*(?:\s|$)*. This single expression covered all cases of *schlecht (bad), schlechter (worse), verschlechtern (worsen), Verschlechterung (worsening),* etc.

2.4. Filter process

A Python[2] script was developed to analyze the notes. Therefore, all the prebuilt keyword files were transformed into a Python list object, in which each element represented an expression. Python's own library for regular expressions was used to filter the input notes. Afterwards, the results for each individual note were aggregated and grouped by patients and physicians, based on each note's sender and recipient ID, respectively.

To normalize the results for each participant, the overall number of free-text elements for each category were subsequently divided by the overall number of free-text elements of the respective participant.

2.4.1. Classification

We slightly adapted the 28 classes used by Gruber et al. and merged similar classes, leading to a smaller set of ten different categories as described in Table 1.

The developed regular expressions categorized each of the free-text elements into none, one or more of these categories.

Table 1 – Categories used for classifying the free-text elements, including short name, description, some examples and the number of keywords that were used in the regular expressions.

Short name	Description	Example	Number of keywords
State of health	Physical and mental well-being, pain, illness, etc.	Had a headache today.	91
Measurements	Measurements of physical activities and vital parameters	*DIGIT* km *DIGIT* kcal	10
Nutrition	Meals, food/drink intake	Had bread and ham for breakfast.	42
Activities	Physical activities, chores, errands	*DIGIT* h cycling	34
Medication	Medication and medication forms (pills, capsules, etc.)	Insulin injected at 10 pm.	245
Technical issues	Technical malfunctions/outages, test messages and phone-related notes	I could not transmit my measurements.	26
Medical appointment	Stationary and ambulatory visits at hospitals, rehab-institutes, physicians, etc.	I am going to be at the hospital next week.	11
Absence	Vacations and travelling	I am going abroad for two weeks.	6
Relatives	Relatives, their activities and visits	Yesterday was my daughter's birthday.	22
Timestamps	Time-related information	Today, tomorrow, at 10 pm, on Monday…	-

[2] http://www.python.org

2.5. Statistical analyses and evaluation

2.5.1. Reference annotation and validation

To validate our algorithm, a sample of 200 free-text notes were randomly selected from the *DiabMemory* dataset. These 200 notes consisted of an equal amount of 100 comments from the patients to the healthcare professionals and 100 feedbacks from the healthcare professionals to the patients. An annotation guideline was developed and provided, based on which three scientists independently annotated the whole set of 200 notes, resulting in three annotations per note. These annotators already had experience with classification of free-text clinical notes. The inter-annotator reliability was calculated via the Fleiss' kappa measure which is part of Pythons *numpy* library [14]. After the first annotation phase, divergent notes were analyzed jointly by the three annotators and an update of the guidelines was made to be more specific even in case of unclear free-text classifications. Afterwards, one of the three annotators applied the finale guideline on the in-sample set of 200 notes to create a *gold standard,* which was used for validating our algorithm.

2.5.2. Comparison of diabetes and heart failure free-text elements

The number and ratio of free-text elements per category as achieved for the *HerzMobil* and the *DiabMemory* corpus were compared with one another. The results of this comparison are depicted in chapter 3.3.

3. Results

3.1. Adaptions of the algorithm

During the development of the *DiabMemory* algorithm, two major differences between the *HerzMobil* and the *DiabMemory* system were identified. The first important difference was the origin of the notes. Whilst both datasets originated from a telehealth system, the *HerzMobil* notes were used for communication between various healthcare professionals like physicians and nurses. In *DiabMemory,* however, a healthcare professional directly communicated with the patient, which might have resulted in more inexpertly and personal expressions and conversations.

The second aspect was the different nature of the diseases heart failure and diabetes. Even though the cores of the algorithms were similar, some keyword categories had to be extended or changed, to fit the profile of both diseases. Nutrition is, for example, a very important aspect in the management of diabetes and, therefore, the nutrition-related keywords were vastly extended. Instead of the initially eight words from Gruber et al., the final dictionary for the *DiabMemory* system had 42 expressions for the nutrition category.

The list of medications had to be extended as well. Diabetes patients need different medication than heat failure patients, which had to be considered during the creation of the keywords. Insulin was especially important, since it is one of the most important medications of patients with diabetes. Activity-related notes represented an entirely new category. Since physical activities are a key factor in the management of diabetes, it was necessary to filter end extract this information. The importance of this category is also reflected by its relative frequency of 16.53% as shown in Table 3.

3.2. Statistical analyses and evaluation

3.2.1. Inter-observer-variability

The inter-observer-variability of the three annotators for the subset of *DiabMemory* notes was calculated based on the annotations prior the annotation guideline update. Based on Feiss' kappa [15], a slight agreement was identified (Fleiss' kappa = 0.017). The comparison of the results of each individual annotator with the gold standard resulted in an accuracy of 91.95%, 90.65% and 87.58%, respectively.

3.2.2. Accuracy of the analyzing algorithm for DiabMemory

To calculate the accuracy of the newly developed algorithm for *DiabMemory*, the manually annotated gold standard was compared to the results of the Python script. Using the previously annotated set of 200 notes resulted in an accuracy of 92.62% for the automated extraction algorithm.

Table 2. Contingency table as achieved for 200 notes, each of which could be positive or negative according to the ten categories, leading to 2,000 decisions as a whole.

	Positive (Annotation)	Negative (Annotation)	Total
Positive (Algorithm)	150	43	193
Negative (Algorithm)	112	1,695	1,807
Total	262	1,738	2,000

3.3. Comparing the relative frequency of categories

Table 3 shows the relative frequency of specific categories normalized by the overall number of notes.

Table 3. The relative frequency of notes per category for the *HerzMobil* and *DiabMemory* system in percent. Note: Relative frequencies do not sum up to 100 %, since 0, 1 or >1 categories can be applied to each note.

	HerzMobil	DiabMemory
State of health	**37.26**	25.33
Measurements	15.97	**27.43**
Nutrition	5.64	**19.8**
Activities	5.03	**16.53**
Medication	**38.05**	4.25
Technical Comments	**31.15**	3.04
Medical appointment	**22.90**	0.42
Absence	**4.09**	2.65
Relatives	**5.63**	3.57
Time references	**63.14**	23.55

4. Discussion

When trying to apply natural language processing methods developed within the *HerzMobil* telehealth program for heart failure patients on data from the *DiabMemory* diabetes monitoring program, we identified that several adaptions were required. These adaptions primarily related to the different focusses addressed in the two programs, i.e. hospital stays, medication adaptions, etc. for the (rather severely ill) heart failure patients, as compared to nutrition, physical activities, and vital parameters for the diabetes patients.

Although some of the differences identified related to the different type of communication (heart failure: between healthcare professionals, diabetes: patient to healthcare professionals), this effect had less influence than the different focus of the diseases.

All categories except three show higher frequencies in the *HerzMobil* notes than in the *DiabMemory* notes. This is because *HerzMobil* notes tend to be longer and, therefore, it is more likely for one note to have multiple categories. The three categories *measurements, nutrition* and *activity* seem to be of great importance for diabetes patients, since they are more common in the *DiabMemory* notes than in *HerzMobil*.

The inter-observer-variability with a Fleiss' Kappa result of 0.017 in between the three annotators shows, that annotating clinical notes is, in some cases, not quite clear and the annotator has room for interpretations. Comparing the accuracy of the three annotators with the newly developed algorithm shows, that the algorithm works with a slightly higher accuracy, even though the reference annotations significantly influenced the golden standard. However, since the annotation guidelines were adapted after the results used to calculate the inter-observer-variability and prior finalizing the golden standard, inter-observer-variability must be interpreted with care.

State of health is a very common category in *HerzMobil* as well as in *DiabMemory*. However, due to the nature of the keywords from Gruber et al., this category is predestined to contain false-positive results. Due to the very general expressions like *gut (good/well),* phrases which have no relation to the state of health might be considered as relevant. For example, the common greeting phrase *Guten Morgen (Good morning)* would be considered as a state of health expression. This is probably the reason for the high relative frequencies of 37.26% and 25.33% for the respective systems.

The difference in the number of medication-related notes between *HerzMobil* and *DiabMemory* might be explained by the origin of the notes. In an expert-to-expert communication like in *HerzMobil*, it is apparently more common to talk about the patient's medication, whilst in a patient to physician communication, the medication is a less frequent topic. Similarly, the rate of activity-related notes might be higher in *DiabMemory* because activities are not so much in the focus of the healthcare professionals (as compared to the patients), although activities are known to have a huge impact on the outcome in both, heart failure and diabetes management.

Due to the large number of 228,109 available notes, using an in-sample set of notes for the evaluation should have a neglectable impact on the results of this paper.

4.1. Outlook

Up to now, we have applied our classification algorithm to the telehealth data retrospectively, to structure the information included in the free-text clinical notes. These analyses have already helped us and will further help us to identify, which important information need to be recorded in a structured way.

Currently, we apply various machine learning and artificial intelligence algorithms to the data derived within our telehealth services to predict events like hospitalizations or dropouts. As a next step, we will include not only structured data, but also the classes derived from the free-text data in these models, which might further improve the model accuracy.

5. Conclusion

Structuring free-text notes from telehealth services via natural language processing methods provides valuable information, which can complement structured data. Even though methods derived for a specific telehealth program can be a good starting point for analyzing data from other programs, various adaptions might be necessary, especially regarding disease specific aspects.

Acknowledgement

The authors would like to thank the Versicherungsanstalt für öffentlich Bedienstete, Eisenbahnen und Bergbau (BVAEB) and the Landesinstitute für Integrierte Versorgung Tirol for supporting this scientific work.

References

[1] K. Barnett, S.W. Mercer, M. Norbury, G. Watt, S. Wyke, B. Guthrie, Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study, *The Lancet* **380**(9836) (2012 Jul), pp. 37-43.

[2] K. Guha and T. McDonagh, Heart Failure Epidemiology: European Perspective, *Current Cardiology Reviews* **9**(2) (2013 May), pp. 23-127.

[3] 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* **126** (2014) 2014 Nov, pp. 734-41.

[4] AIT, DiabMemory, https://www.ait.ac.at/themen/telemedical-solutions/projects/diabmemory/, last access: 04.01.2018, URL, last access: 05.01.2021.

[5] V. Schusterbauer, D. Feitek, P. Kastner, H. Toplak, Two-Stage Evaluation of a Telehealth Nutrition Management Service in Support of Diabesity Therapy. *Studies in health technology and informatics* **248** (2018), pp. 314–321.

[6] H. J. Kong, Managing Unstructured Big Data in Healthcare System. *Healthcare informatics research*, ***25***(1) (2019), pp. 1–2.

[7] F. Hebal et al, Automated data extraction: merging clinical care with real-time cohort-specific research and quality improvement data., *Journal of Pediatric Surgery* **52**(1) (2017 Jan), pp. 149-152.

[8] E. Yehia et al, Ontology-based clinical information extraction from physician's free-text notes, *Journal of Biomedical Informatics* **98** (2019), pp. 103276.

[9] H. J. Lee et al, Identifying direct temporal relations between time and events from clinical notes, *BMC Med Inform Decis Mak* **18** (2018 Jul), pp. 49.

[10] R. Zhang et al, Detecting clinically relevant new information in clinical notes across specialties and settings, *BMC Med Inform Decis Mak* **17**(2) (2017 Jul), pp. 86.

[11] K. Gruber, R. Modre-Osprian, K. Kreiner, P. Kastner, G. Schreier, Development of text mining based classification of written communication within a telemedical collaborative network, *eHealth* (2015 Jun), pp. 35-42.

[12] Gruber K., Schreier G. Dashboard und Klassifikatoren in einer telemedizinischen Software zur Unterstützung der kollaborativen Herzinsuffizienz-Versorgung, *Technischen Universität Graz* (2015 Apr).

[13] Wiesmüller et al, Automated Extraction of Time References From Clinical Notes in a Heart Failure Telehealth Network (in press)

[14] Compute Fleiss' kappa using numpy, 2021 GitHub, Inc. https://gist.github.com/skylander86/65c442356377367e27e79ef1fed4adee, last access: 25.01.2021.

[15] Kappa statistics and Kendall's coefficients, 2019 Minitab, LLC, https://support.minitab.com/en-us/minitab/18/help-and-how-to/quality-and-process-improvement/measurement-system-analysis/supporting-topics/attribute-agreement-analysis/kappa-statistics-and-kendall-s-coefficients/, last accessed: 08.02.2021