FACTA UNIVERSITATIS (NIŠ)
Ser. Math. Inform. Vol. 35, No 3 (2020), 825-841
https://doi.org/10.22190/FUMI2003825A

NORMALIZATION OF HEALTH RECORDS IN THE SERBIAN LANGUAGE WITH THE AIM OF SMART HEALTH SERVICES REALIZATION *

Aldina R. Avdić, Ulfeta A. Marovac and Dragan S. Janković

Abstract. The development of information technology increases its use in various spheres of human activity, including healthcare. Bundles of data and reports are generated and stored in textual form, such as symptoms, medical history, and doctor's observations of patients' health. Electronic recording of patient data not only facilitates day-to-day work in hospitals, enables more efficient data management and reduces material costs, but can also be used for further processing and to gain knowledge to improve public health. Publicly available health data would contribute to the development of telemedicine, e-health, epidemic control, and smart healthcare within smart cities. This paper describes the importance of textual data normalization for smart healthcare services. An algorithm for normalizing medical data in Serbian is proposed in order to prepare them for further processing (F1-score=0,816), in this case within the smart health framework. By applying this algorithm, in addition to the normalized medical records, corpora of keywords and stop words, which are specific to the medical domain, are also obtained and can be used to improve the results in the normalization of medical textual data.

Keywords: telemedicine; e-health; epidemic control; smart healthcare; medical data mining.

1. Introduction

In medical information systems, a large amount of data is created and stored every day. They allow storing of common information (number of patients examined by a doctor, consumption of materials, prescriptions, etc.), but they also include medical reports that contain patient information such as anamnesis, diagnosis, symptoms, etc. These data collected daily should be used for analyzes and predictions to improve medical information systems. Therefore, it is necessary to

Received October 13, 2019; accepted March 10, 2020
2010 *Mathematics Subject Classification.* Primary 68T50; Secondary 92C50
*This paper is partially supported by the Ministry of Education, Science and Technological Development Republic of Serbia, Projects No. III44007 and ON174026.

prepare and appropriately process this data. Well prepared data can be processed for different purposes, for example in smart health services [1]. Smart health services are important for improving the quality of services provided, the efficiency of health services, etc. As such, they are considered the basis for the implementation of smart medical IS as an indispensable part of the concept of smart cities [2], which is becoming more and more relevant today. The fact is that the problem of organizing life and optimization, especially in big cities, is one of the current problems that is being intensively solved. The goal is to provide a range of services that will make life easier and cheaper. Healthcare occupies a significant place in this concept.

A smart city is a place where information and telecommunications technologies are used to enhance traditional services. It is a city that connects physical infrastructure, information technology infrastructure, social infrastructure, and business infrastructure to enhance the collective intelligence of the city. Within the smart city, there are branches such as smart transportation, smart healthcare, energy efficiency, smart technology and infrastructure, smart education, smart management, and smart people. Smart healthcare is e-Health aimed at promoting public health services in smart cities [1]. Smart healthcare uses technological innovations in the health care system. As we mentioned before, part of the medical records could be processed and used for the purposes of smart health and its services, such as epidemic control, the visualization of vaccination data, disease prevention, a self-diagnosis, etc.

The motivation for our paper is creating conditions for building smart health services using text mining techniques. For the purposes of this paper, a corpus that is containing 5,261 medical reports were created. In this paper, the aim was to create the algorithms for deleting the non-relevant data from the input set of medical data and preparing the relevant data for further processing. The relevant data is purified from excessive words, punctuation marks and other data that did not carry any informative value. Obtained redundant and keywords are remembered, in order of using them in the normalization of new data.

The paper is organized in the following way. The second section describes related researches. An overview of smart health services is given in the third section. The fourth section describes the proposed smart health framework for epidemic control. Next, the description of the data set and medical data are given to understand the need for their normalization. The following section presents our approach to normalization of medical data written in Serbian. Then the results of the application of the proposed method on the above-mentioned medical corpus are given. Finally, conclusions and directions for further research are given in the last section.

1.1. Related Work

Papers from our field of research can be divided into two groups. The first group includes papers describing the normalization and processing of medical textual data which is not written in the Serbian language. These papers present universal characteristics of medical data and their normalization, which are independent of the language.

The differences between clinical and ordinary texts and problems related to obtaining information from medical texts are described in [3]. In [4, 5], the methods used in the normalization of electronic medical data are given. In the paper [6], one way of classifying medical data and the application of neural networks in solving this problem is described. Here are processed texts from the Internet that contain a description of the health status of patients, the symptoms and the like. Paper [7] proposes a method for normalizing symptoms written in the Chinese language. A complete system that normalizes and extracts information from medical records and its architecture is described in [8]. In [14] the characteristics of clinical reports are presented, from the corpus of medical reports in Sweden taken from 2014-2015.

On the other hand, the methods described in the previous papers are not sufficient to fully apply to the medical text in the Serbian language, and of course, do not include lexical resources specific for the medical domain. The second group includes papers dealing with the normalization of textual documents and their analysis in the Serbian language, but even here the normalization was not carried out in clinical texts. The papers [10, 11] give the description of the process of normalization of informal documents in the Serbian language with the aim of faster searching. In [9] is presented the normalization of text in Serbian language using n-gram analysis. The specific language resources are needed for different data processing [12, 13]. Some steps from these methods can be used in the normalization of medical data, but most of them need to be adapted. The structure of medical data (reports) and their contents are much different from other types of text documents, so they need to be adapted according to their specificities.

1.2. Smart Health Services and the Smart City

The smart city infrastructure includes physical infrastructure, information and communication technology (ICT) infrastructure and services. Physical infrastructure is the part of a smart city including roads, railways, a water supply system etc. ICT infrastructure consists of computer and information systems, networks, sensors etc. so it is a link between the other two infrastructures of the smart city. It is based on the Internet of Things (IoT) and Big Data [2, 15]. Service infrastructure is based on physical infrastructure and can have some ICT components [1].

The application of mobile and ubiquitous computing enabled the collection of data from the user's environment. These data are contextual, and applications using them are called context-aware applications [1]. The use of smartphones to improve the health status of the user led to the formation of a new sub-feature in electronic healthcare, which is mobile healthcare (m-Health). The synergy of mobile and electronic healthcare with the concept of smart cities has come up with a new term – smart healthcare (s-Health) [1].

M-Health and s-Health are subsets of e-Health, but they also have their intersection. S-Health and m-Health differ in the source information they use, and the flow of this information. For m-Health, the source information comes from the users/patients, while in s-Health, besides this information, it uses the collective data obtained from the infrastructure of the smart city. Regarding the difference in the

data, it refers to the fact that after processing, the m-Health returns the response to the user, and in the s-Health, the data are returned to the user, but also affect the collective data of the smart city.

To better understand the flow of data from patients to the ICT infrastructure of the smart city and back, some specific examples of services and their classification are given. Most smart healthcare services can be classified into one of these categories [16]:

- **Services based on the collective intelligence of the city.** They use data collected from sensors located on the territory of a smart city and analyze the obtained data for predictions for different purposes. Such service is suggesting a route with less air pollution for users with a respiratory problem.

- **Context-oriented services** analyze the close user environment, based on the image, the sound of motion detection, and so on. Examples are image processing and analysis for detecting abnormal phenomena in the relevant diseases. Motion detection is used for determining emotions, stress levels, breathing etc.

- **Services based on IoT devices** - They also analyze the close user environment, over the wearable IoT devices. The examples are the measurement of blood sugar, electrocardiogram of the heart, blood pressure and temperature using IoT bracelets.

- **Smart houses for patients** are smart environment (home) care for the elderly and the chronically ill, using technology infrastructure (sensors, cameras, wearable devices, and web services) to respect the wishes of patients to be at home.

- **Services based on crowdsourcing and medical data mining** - Similar to the first group of services, just with a large number of the original information from a wider area, which may be from the outside of the territory of the city. The examples are services for concluding in cases of measuring and obtaining abnormal results in some diseases and for detecting depression levels through crowdsourcing by comparing data from other users and using medical data mining.

The smart health framework which we are developing is based on medical data mining and it is described in the next section.

2. The Smart Health Framework

Smart health services that can be created by analyzing patient history data are smart health services based on medical data mining (Figure 2.1). They could be displayed on a public health portal and would include the following reports:

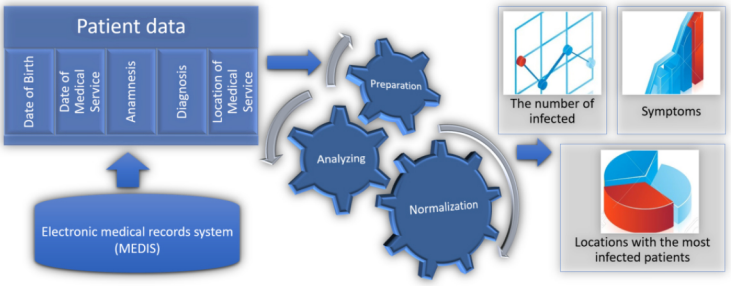

FIG. 2.1: Smart health framework for epidemic control

a) Report on how many people have a specific diagnosis daily in the city, as well as on a weekly, monthly, and annual basis. Based on this service, the user would be aware of the existence of the epidemic disease in his city and its status, whether it took a significant hit depending on whether the number of patients admitted day by day increased or decreased.

b) Reporting how many people have been diagnosed with a disease at a particular health station. Here a citizen could see how many people are ill in his immediate area.

c) Report on the presence of the disease in different age groups, where the citizen could decide if he belongs to the riskier group.

d) Report on the most common symptoms in people diagnosed with the disease. Here, a citizen could see the number of people who showed up with certain symptoms and had a diagnosis of a disease whose epidemic was ongoing. In this way, he could more easily recognize the symptoms and contact the doctor himself if he had any or most of the symptoms.

All these services would help to keep the citizen up to date with the epidemic in his place and take measures to avoid or treat the disease, and in this way the consequences of the epidemic will be reduced.

The first three services require data analysis and visualization, which is not demanding, while the fourth requires specific textual processing for proper symptom extraction. Analyzing the anamnesis we used, we came across abbreviations, misspellings, different word forms, and synonyms for the same symptoms. The anamnesis should be cleansed of words that have no meaning, and the words of significance should be reduced to the same form. Abbreviations should also be processed and preserved. There are also negations of symptoms in anamnesis, so the service would not show the true number of patients who have a symptom if negation were not taken into account. Addressing these issues is a key motive and contribution of this paper.

3. The Data Set

The specificity of the language in which the report is written further complicates the process of normalization. Corpora are required to identify the specifics of medical reports. Medical records are sensitive to research because of the confidentiality of the information they carry, so appropriate medical identification of medical records must be made and all personal data of patients, as well as doctors, are removed. There are several English-language corpora available such as: Informatics for Integrating Biology the Bedside, BioScope Corpus and the Thyme corpus [17]. There is no Serbian corpus in electronic form publicly available. We have used about five thousand medical records written in the Serbian language from 32 outpatients belonging to the Health Center Niš (DZ Niš), collected by the MEDIS.NET information system. The medical reports were written by 169 different doctors. This corpus is made by all ethical standards, with the de-identification of patients and medical staff as well as maintaining links to the affiliation of multiple reports to the same patient.

4. The Description of Medical Data

Medical reports are mostly generated by the hospital's internal needs. Clinical reports are needed for different stakeholders such as: medical staff to keep up with day-to-day activities, patients to document their health status, clinical research (medical researchers, pharmacists, epidemiologists, etc.), hospital management to keep track of finances and inventory planning, budget etc. Medical reports may contain numerical and textual information. Medical data is of mixed type (structured, semi-structured and unstructured) and therefore requires more complex processing that involves the existence of appropriate specialized lexical resources. The structure part contains values of specific variables, so it is the easiest to process (name, surname, year). The semi-structured part gives descriptive values for some parameters, but the structure is still known (temperature, pressure and laboratory analysis). The unstructured part consists of free text that the doctor gives and consists of symptoms, history, observations, conclusions. Unstructured data contains linguistically incomplete, informal and non-standard abbreviations which makes it difficult for computer processing and analysis. For this, it is necessary to pre-process the data before analysis to bring it into a standardized form.

Table 4.1 gives an example of a medical report that we are processing. In it, we can identify the structural part containing the date of service, name of service, diagnosis, diagnosis code, organizational unit in which the service was provided and location of the service. Also, this report contains an unstructured section consisting of an anamnesis. This part is more complex to process because it needs to extract relevant data and transform it into a standardized format, suitable for further processing.

Medical reports are most often written by doctors and nurses. Because of the speed at which they are written, they often contain many errors. Very often, the sentences are incomplete, for example, the auxiliary verbs are omitted as well as

Table 4.1: An example of the used medical record

Date of the service	23-03-18
Name of the service	Re-examination of adults
Anamnesis	Pacijent dobio sinoc osip po koži. Makulopapulozna ospa po kozi iza ušiju, čela i spušta se na trup. Vezikularni disajni šum (en. The patient received a skin rash last night. Maculopapular rash on the skin behind the ears and forehead and going down to the hull. Vesicular breathing noise)
Diagnosis	Morbili – Measles
Diagnosis code	B05
The organizational unit of the service	General medicine
Location of the service	Central building

the subject when it is obvious that the subject is the patient himself. Even the attachments are rarely found in the medical data, only in the description of the symptoms (e.g., fever, sweating, shortness of breath). An example is given in Table 4.2.

Such descriptions are concise and carry the most important information that, in a sense, facilitates the processing of such text. In medical reports, every word has more weight. In addition to deliberately omitting certain words, there are very common mistakes and spelling mistakes such as:

- two words are merged, and instead of space is a letter,

Table 4.2: An example of incomplete sentences in anamnesis

Original anamnesis text	
Serbian	English
“pečenje i svrab po celom telu, difuzna ospa koja svrbi”	”Burning and itching all over the body, diffuse wasp that itches”
Extended meaning	
Pacijent ima pečenje i svrab po celom telu u obliku difuzne ospa koja svrbi	The patient has burning and itching all over the body in the form of diffuse pox that itches

- misspelled word,
- incorrect writing of diacritical symbols, (eg. izvestaj, etc.), and often writes dj instead đ,
- a misspelled or omitted letter often, or a letter of excess e.g. "kašljke" instead of "kašlje",
- there are abbreviations of medical, punctuation marks, brackets, numbers, etc.
- x in anamnesis, (e.g. extremiteti).
- z and y mixed because of the keyboard.

When compared to other types of text according to Ehrentraunt and others [18], twice as many spelling mistakes (10%) are found in medical reports than in manuscripts, newspaper articles, web articles, etc. This is not surprising given the time limit they have for patient screening - average time due to scheduled appointments.

We often find abbreviations in medical reports. Abbreviations are ways of writing longer words without all the letters. It makes the text easier to write and read, but only on the condition that the reader knows the meaning of abbreviations. The problem is non-standardized abbreviations. Sometimes the same abbreviation can have an ambiguous meaning (for example, feb. for February and febrile). It is often the case that different abbreviations are used for the same term: shorten the temperature differently (e.g. T*, t, temp, etc.); In addition to abbreviations, acronyms are often used. Standardized acronyms are used in medical reports. However, there may also be problems because not often do acronyms have more meaning, for example: DIK is an acronym for pediatric infectious clinic (*dečija infektivna klinika*) and disseminated intravascular coagulation(*diseminovana intravaskularna koagulacija*), EEG for the electroencephalography and the electroencephalogram, etc.

The interpretation of acronyms, in this case, is likely to be related to the specialty of the physician who writes them or to the diagnosis. This association can be processed manually but also by machine learning methods with the appropriate corpus of data. Words that have a Latin or Greek root are common in medicine. However, in recent decades, more and more English words have been used for which there is no adequate translation into Serbian, so they are most commonly used in their original form. This is especially striking when it comes to medical techniques, medical devices, and certain surgical procedures. Depending on what the purpose of further processing of medical texts is in normalization, it may be desirable to simplify the litigants used in medical reports (e.g. pulmo – lung, hyperemia – the increase of blood flow to different tissues in the body).

The occurrence of negation in medical texts is very common because it excludes the existence of some symptoms that indicate disease. The presence of negation in the report significantly affects the meaning of the report itself, given the structure of the medical texts (short and concise). An example of the anamnesis with negation

Table 4.3: An example of the anamnesis with negation

Serbian	**English**
"Izbilo ga nešto po telu, pre dva dana, temper. nema, ne kašlje, ne boli ga grlo"	"There was something on the body, two days ago, temper. no, no cough, no pain"

is given in Table 4.3. The importance of the processing of negation in medical texts is demonstrated by the existence of the English corpus BioSope with a radial report in which the negation is manually marked as well as its scope of action [17]. For Serbian, there are rules of negation that can significantly improve data processing [19].

5. The Applied Normalization Method

The method for normalizing medical reports that we propose consists of several steps: tokenization, removal of stop words, processing of negation, removal of punctuation, and cutting off into n-grams. The algorithm we applied to the described data is shown in Figure 5.1.

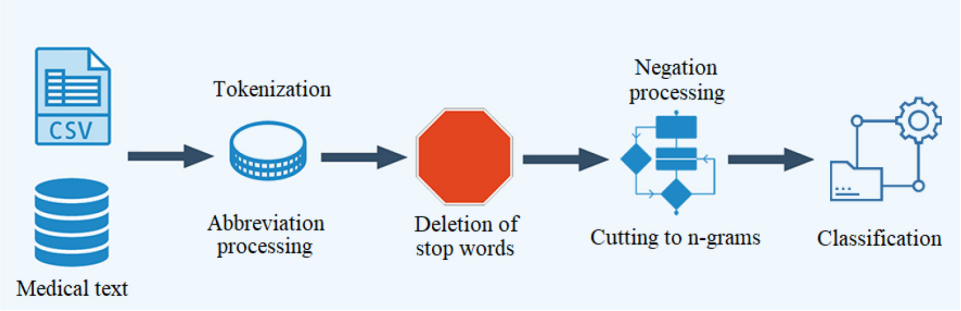

Fig. 5.1: Normalization of medical reports

Tokenization: In the first step of preparing a clinical text for further processing, it is necessary to delete unnecessary elements (multiple blanks, dates, special characters, etc.) from the anamnesis to identify the annotated words. The anamnesis were taken in the Serbian language using a Latin letter containing diacritical symbols ć, č, ž, š, đ, so in this step they are changed by the symbols cx, cw, zx, sx and dx respectively, since x and w are not used in Serbian letter, and proposed combinations are not found in corpus of diseases and related health problems (Serbian and Latin version) [20]. This makes it easier to process diacritical symbols because otherwise they will be transformed into special characters. In this step, the abbreviation processing is also done. Abbreviations belonging to standard medical

Table 5.1: Some of stop words from dictonary

Stop words	**Stop words**
kakva (en. which)	igde (en. where)
iako (en. although)	će (en. will)

abbreviations are generally acronyms and are capitalized. Non-standardized abbreviations specific to the Serbian language are separated and reduced to a single form without a point at the end. Punctuation marks remain in this step, because of their meaning in the processing of negation. When the reports are cleared in this way, then they can be divided into words, that is, tokens, at the level of further processing.

Deletion of stop words: The reports also contain words that are not meaningful, and these words are called stop words. These words should be deleted from medical records to reduce their scope and to retain only relevant information. By reducing the volume of data, the speed of their further processing increases. Stop words are usually adverbs, prepositions, pronouns, conjunctions, words and other words that are not relevant for determining the meaning of the text. The process of removing the stop word begins by creating a dictionary of stop words. The dictionary of stop words in Serbian, which is the result of our previous research [9], contains 3117 stop words. In Table 5.1, some words from this dictionary are given. Some words need to be removed from the set of stop words because they represent medical abbreviations or abbreviations for chemical elements. Their processing will be performed afterwards, considering whether the appropriate elements represent an abbreviation or stop word (whether they are capitalized, do they appear with other keywords, for example, "se" can be the auxiliary verb, and Se is the value of selenium from blood tests). Negation signals are also excluded from the set of stop words in order not to lose information about the negation of the existence of some symptoms.

Negation processing: The specificity of medical reports is that they are written in the third person and are concise, so that not all the negation signals appear. From the negation signals in the processed corpus, next words were found: ne (en. not), nije (en. it is not), nema (en. haven't), bez (en. without) and ni (en. nor). The extent of the action of this negation will be determined by the first non-stop word following the negation signal to the punctuation mark, or if it does not exist then it will be taken first before the negation signal to the punctuation mark. A word in the negation range receives a prefix ne_ and the negation signal is deleted. An example is given in Table 5.2. Punctuation marks can be removed after the negation has been processed.

Cutting off to n-grams: Bearing in mind the richness of the Serbian language, the presence of cases, phonetic changes, and changing verbs by person, gender and number, it happens that words that carry the same meaning can be found in many forms. Word-based grammatical rules would make this problem complex. Also, a morphological dictionary of Serbian, which is not publicly available in electronic

Table 5.2: An anamnesis after processing abbreviations and negations

Before processing	**After processing**
Izbilo ga nešto po telu, pre dva dana, temper. nema , ne kašlje, ne boli ga grlo	izbilo telu, dana, ne_temperaturu , ne_kasxlje, ne_boli grlo

form, would be necessary. So, we decided on a language-independent variant, which is a cutting to n-grams. An n-gram is a subset consisting of n elements of a given string. For example, the word "učiti" consists of the following n-grams: u-č-i-t-i (length 1), uč-či-it-ti (length 2), uči-čit-iti (length 3), učit-čiti (length 4) and učiti (length 5). N-grams are obtained by moving the frame of length n, whose origin can be in positions 1 to $m - n + 1$, where m is the length of the string [9]. By analyzing the content of the n-gram, the correlation between the appearance of the n-gram and the characteristics of the text can be noticed. N-grams are suitable for use in the analysis of textual documents in natural languages, due to language independence. N-gram analysis is a procedure applied to text, and its result is to obtain a set of n-grams of a given length. In proposed method, we used n-grams which length is 4 (4-grams). When cutting to 4-grams, the negation prefix is ignored (ne_temperaturu is cut to ne_temp). This length was chosen as optimal because it gives the best results for Serbian compared to the analysis of 3-grams and 5-grams (it is compared in next section).

Classifying n-grams is required after normalization, where the n-gram is sorted into keywords and stop words that are specific to the medical domain. Keywords include symptoms that can be remembered for the appropriate type of illness, while stop words are those that occur in all anamnesis, regardless of diagnosis. These are the words that most commonly appear in medical reports ("uput", "doznaka"). Since there are synonyms in keywords, they must be grouped. Classification is a work that requires special attention and will be the subject of extensive research in subsequent papers. Here, the classification is made in a simple way to show the impact of normalization on the results. The classification was made by looking for [20] n-grams among the symptoms and if they were found they are declared as keywords for the appropriate type of illness. Those n-grams which are not found in symptoms were declared as stop words in the medical domain.

6. Results and Discussion

After normalization over the described medical reports, the anamnesis was obtained in the normalized form. Of the 5261 medical reports, 2112 contained text in the non-structural section, while the rest contained only the structured section. After ejecting numbers and stop words, there were 16606 words left to normalize. The normalization with n-grams of lengths 3, 4, and 5 was performed, to determine which n-gram size yields the best results. To compare results precision, recall, and

Table 6.1: The results of different normalization methods

Normalization method	**WN**	**NW**	**CNW**	**Precision**	**Recall**	**F1-score**
3-gram	16606	13011	10213	0.7850	0.6150	0.6897
4-gram	16606	10330	9488	0.9185	0.5714	0.7045
5-gram	16606	8472	7891	0.9314	0.4752	0.6293
Proposed method	16606	13375	12232	0.9145	0.7366	0.8160

F1-score measures were used.

$$Precision = \frac{CNW}{NW} \tag{6.1}$$

$$Recall = \frac{CNW}{WN} \tag{6.2}$$

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{6.3}$$

Where CNW is the number of corectly normalized words, NW is the number of normalized terms, and WN is the number of words to be normalized.

The normalization results are shown in Table 6.1 and the best F1-score was for 4-grams. From the table it can be seen that increasing the length of n-grams increases the precision and the recall weakens. But for 3-grams the precision is much lower, while for 5-grams the recall is smaller, and therefore their F1-score is smaller compared to 4-grams. The precision for 3-grams of was expected to be lower because if the word had a prefix, the root (as the carrier of its meaning) is removed by normalization. Here, corectly normalized word is word which resulting n-gram was a part of the semantic root, and this was compared to a manualy labeled corpus.

Then we performed the proposed method as it is described in 5th section and the obtained results are also given in the Table 6.1. It can be obtained that the proposed method have best results for F1-score (0.8160). This results are comparable with methods for normalizing medical data in other languages (in [7] F1-score=0.6562). The words were reduced to 4-grams, and then counted in order to find their frequency in the anamnesis. The anamneses are now remembered in purified form, contain only relevant data and their volume has been reduced. Table 6.2 gives an example of anamnesis before and after normalization.

Then the n-grams with the largest number of occurrences in medical histories are divided into two groups, by meaning. The first group consists of keywords, which are those words whose meaning is closely related to the disease and they are shown in Table 6.3. Table 6.3 shows that the number of occurrences of symptoms in the corpus varies significantly depending on the application of the normalization method. So, if n-grams are searched in the corps, and abbreviations, synonyms and

Table 6.2: An example of one normalized anamnesis

Before normalization	**After normalization**
Kontrola: još uvek ima malaksa-lost, slabost, Savet , kontrola 30. 3. 2018	kont mala slab save kont
Pacijent dobio sinoc osip po koži. Makulopapulozna ospa po koži iza ušiju, čela i spušta se na trup. Vezikularni disajni šum	paci dobi osip kozxi maku ospa kozi usxij cwela spusx trup vezi disa sxum

negation is not considered, we will not get the exact number of symptoms. Table 6.3 shows that the percentage of occurrence for some symptoms is significantly increased (rash, because of synonyms and temperature, as it is often abbreviated). There are also those symptoms whose occurrence is reduced (change, for example) after the application of the proposed method, which is because they were found in negation.

The second group consists of stop words whose meanings do not determine the patient's illness, and they can appear in any medical history or clinical document. As the anamnesis is already in purified form without the stop words that are characteristic of general documents, the stop words thus extracted refer to the medical domain and can be stored to improve the results of normalization of new anamnesis or documents. These are shown in Table 6.4. The most common occurring keywords are related to symptoms of the disease (rash, temperature, cough, etc.) and the most frequent stop words are medical terms that don't indicate the patient's condition (appointment, report, etc.).

Therefore, this normalization can be used in the execution of statistics in the control of epidemics. The simple example is given in Figure 6.1.

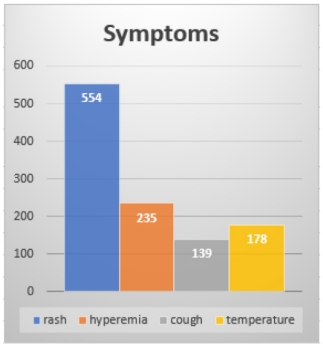

Fig. 6.1: Statistics in the control of epidemics

If we compare the Table 6.3 with the chart (Figure 6.1), it can be concluded

Table 6.3: The percentage of occurrence for the most occured keywords

N-gram	Simple 4-gram	Proposed method	Associated words	Associated n-grams of synonyms
hipe	10.62%	12.07%	hiperemija (en. hyperemia)	
kasxl	5.97%	6.63%	kašalj (en. cough)	kaša
kozxi	5.40%	6.01%	koža (en. skin)	kože, kožn, koži
licu	7.34%	10.89%	lice (en. face)	lica, lice
morb	17.90%	17.85%	morbili (en. morbili)	
ospa	13.92%	25.76%	ospa (en. rash)	osip
prom	5.16%	5.02%	promena (en. change)	
pulm	14.16%	13.92%	pulmo	
telu	10.89%	12.74%	telo (en. body)	tela, telu, telo
temp	8.43%	18.32%	temperatura (en. temperature)	
zxdre	12.36%	12.41%	ždrelo (en. pharynx)	

Table 6.4: The most occured stop words

n-gram	The percentage of occurrence in the anamnesis	Associated words
Kont	21.77%	kontrola (en. appointment)
Izve	13.01%	izveštaj (en. report)
Bolo	11.36%	bolovanje (en. sick leave)
Dana	10.46%	dan (en. day)
Dozn	9.28%	doznaka (en. remittance)
Uput	7.00%	uput (en. refer)
Nala	5.73%	nalaz (en. finding)
Preg	5.30%	pregled (en. examination)

that more information can be extracted to make the symptoms more accurate. For example, skin, face, body are mentioned in the table, which are the rash provisions. This means that the joint occurrence of n-grams in the anamnesis could be determined in order to obtain phrases, which will be the subject of the further research.

7. Conclusion

This paper outlines how medical reports in an electronic form stored daily can be used, as well as problems that can be encountered in their analysis and processing. The importance of text normalization for smart health services was demonstrated and a way to normalize clinical textual information written in Serbian was presented. In the experimental part, the results of the application of the proposed algorithm for normalization over the anamnesis collected using the MEDIS medical information system are presented. The described method extracts relevant data from medical reports so that it can be used for a variety of purposes, including public health and other smart healthcare services. The data stored in a purified form can be used for further processing. Applying the proposed method over the collected data, we obtained the F1-score (0.816). Since there is no adequate method in the Serbian language for comparison, compared to the methods for normalizing medical data in other languages [7], the proposed method produces good results. In this way, a corpus containing the stop words for a medical domain in the Serbian language can be formed, which can be used for processing medical text for various purposes. The subject of our further research will be the classification of medical terms and the labeling of entities in medical records and their application in smart healthcare services.

REFERENCES

1. A. Solanas, C. Patsakis, M. Conti, I. S. Vlachos, V. Ramos, F. Falcone and A. Martinez-Balleste: *Smart health: a context-aware health paradigm within smart cities*, IEEE Communications Magazine, vol. 52, no. 8, pp. 74-81, 2014.
2. S. P. Mohanty, U. Choppali and E. Kougianos: *Everything you wanted to know about smart cities: The internet of things is the backbone*, IEEE Consumer Electronics Magazine, vol.5, no. 3, pp. 60-70, 2016.
3. S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler and J. F. Hurdle: *Extracting information from textual documents in the electronic health record: a review of recent research*, Yearbook of medical informatics, vol. 17, no. 1, pp. 128-144, 2008.
4. W. Sun, Z. Cai, Y. Li, F. Liu, S. Fang and G. Wang: "*Data processing and text mining technologies on electronic medical records: a review,* Journal of healthcare engineering, vol. 2018, pp. 1-10, 2018.
5. I. Yoo, P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J. F., Chang and L. Hua, *Data mining in healthcare and biomedicine: a survey of the literature*, Journal of medical systems, vol. 36, no. 4, pp. 2431-2448, 2012.

6. K. Lee, S. A. Hasan, O. Farri, A. Choudhary, and A. Agrawal, *Medical Concept Normalization for Online User-Generated Texts*, In 2017 IEEE International Conference on Healthcare Informatics (ICHI), pp. 462-469, 2017.

7. Y. Wang, Z. Yu, Y. Jiang, et al.: *Automatic symptom name normalization in clinical records of traditional Chinese medicine*, BMC Bioinformatics 11, 40 (2010) doi:10.1186/1471-2105-11-40

8. G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler and C. G. Chute, *Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications*, Journal of the American Medical Informatics Association, vol. 17, no. 5, pp. 507-513, 2010.

9. U. Marovac, A. Pljaskovic, A. Crnisanin and E Kajan, *N-gram analysis of text documents in Serbian language*, In Telecommunications Forum (TELFOR), pp. 1385-1388 , 2012.

10. A. Ljajić, U. Marovac and M. Stanković: *Comparasion of the influnce of diferent normalization methods on tweet sentiment analysis in Serbian language*, Facta Universitatis (NI Ser. Math. Inform. Vol. 33, No 5 (2018), 683–696 https://doi.org/10.22190/FUMI1805683L , M51

11. A. Pljasković, D. Avdić, U. Marovac, A. Crnisanin and D. Rančić , *Pretraživanje dokumenata na srpskom jeziku za potrebe m-Uprave*, ETRAN, pp. RT4.6, 2013.

12. P. Rajković, D. Janković and D. Vučković: *Adaptation and Application of Daitch – Mokotoff SoundEx Algorithm on Serbian Names*, Conf. PRIM (book of abstracts), pp. 21, Kragujevac 2006.

13. P. Rajković, D. Janković and D. Vučković, *Using String Comparison Algorithms for Serbian Names*, Proceedings XLI International scientific conference on Information, communication and energy systems and technologies – ICEST, pp. 221-224, Sofia, June 29th – July 1st, 2006.

14. H. Dalianis, *Characteristics of Patient Records and Clinical Corpora. In: Clinical Text Mining*, Springer, Cham, 2018.

15. M. Batty, *Big data, smart cities and city planning*, Dialogues in Human Geography, vol. 3, no.3, pp. 274-279, 2013.

16. A. Avdić and D. Janković, *Healthcare in smart cities- privacy and security issues*, CPMMI 2018, 5th International Conference of Contemporary problems of mathematics, mechanics and informatics (CPMMI) Novi Pazar, Serbia, June 17-19, 2018

17. V1. Vincze, G. Szarvas, R. Farkas, G. Móra and Csirik J: *The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes*, BMC Bioinformatics. 2008 Nov 19;9 Suppl 11:S9. doi: 10.1186/1471-2105-9-S11-S9.

18. C. Ehrentraut, H. Tanushi, J. Tiedemann,and H. Dalianis (2012). *Detection of hospital acquired infections in sparse and noisy Swedish patient records.*, In Proceedings of the Sixth Workshop on Analytics for Noisy Unstructured Text Data (AND 2012) Held in Conjunction with Coling 2012, Bombay. ACM Digital Library

19. A. Ljajić and U. Marovac: *Improving Sentiment Analysis for Twitter Data by Handling Negation Rules in the Serbian Language*. Computer Science and Information Systems, Vol. 16, No. 1, 289-311. (2019), https://doi.org/10.2298/CSIS180122013L

20. *Međunarodna Statistička Klasifikacija Bolesti i Srodnih Zdravstvenih Problema, Deseta revizija*,EDITOR Dr Miljan Ljubičić, Institut za javno zdravlje Srbije "Dr Milan Jovanović Batut", edition 2010, vol. 1,(2013)

Aldina R. Avdić
State University of Novi Pazar
Department of Technical Sciences
36300 Novi Pazar, Serbia
aplaskovic@np.ac.rs

Ulfeta A. Marovac
State University of Novi Pazar
Department of Technical Sciences
36300 Novi Pazar, Serbia
umarovac@np.ac.rs

Dragan S. Janković
Faculty of Electronic Engineering
Department of Computer Science
18000 Niš, Serbia
dragan.jankovic@elfak.ni.ac.rs