Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 178 (2020) 284–290

Procedia Computer Science

www.elsevier.com/locate/procedia

9th International Young Scientist Conference on Computational Science (YSC 2020)

# Preprocessing of unstructured medical data: the impact of each preprocessing stage on classification

M. Kashina*, I.D. Lenivtceva, G.D. Kopanitsa

*ITMO University, Saint-Petersburg, Russia*

**Abstract**

Nowadays, it is still important to develop methods for processing data, in particular medical texts, in Russian. In this paper, we checked how each stage of text pre-processing affects the result of the classifier. The paper analyzed 269923 records of allergic anamnesis of patients, 11670 of which were placed for further processing. We consider the main stages of pre-processing: tokenization, deletion of stop words, error correction, document cropping, normalization, class harmonization, and vectorization. To vectorize the data, we have selected the Bag-of-Words. The method of logistic regression was chosen for classification, since it has easy reproducibility and interpretation. Precision, recall and F-measure were selected as evaluation metrics. The results (F = 88.12%) showed that the most effective was the stage of normalization and error correction.

© 2020 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the 9th International Young Scientist Conference on Computational Science

*Keywords:* preprocessing; tokenization; classifier; medical text; natural language processing; allergy.

## 1. Introduction

Allergies are one of the most common diseases in the world. As of 2018, every third inhabitant of the planet suffers from allergic rhinitis and every tenth from bronchial asthma [1]. The main causes of allergies are various environmental factors (allergens) such as dust, pollen, animal hair, citrus fruits, etc [2]. When diagnosing health in medical centers, all information received from the patient is registered in the electronic medical card (EMC).

Medical information is very complex. It is very difficult to establish semantic interoperability. Interaction of information systems is relevant, especially data exchange both within and between medical institutions.

* Corresponding author. Tel.: +79215520151;
*E-mail address:* k.mariya1997@gmail.com

1877-0509 © 2020 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the 9th International Young Scientist Conference on Computational Science
10.1016/j.procs.2020.11.030

Interoperability is the ability of two or more information systems or components to exchange information and to use the information obtained as a result of exchange. Patient health data is recorded in various formats and forms [3]. Information may include various laboratory test results, treatment and medication records, treatment progress records, referrals, notes, and so on. Most of the data is unstructured, which is why there is a problem with data interoperability [4].

Text mining is the process of searching for or extracting useful information from text data [5]. Text mining is a growing field of research in computer science that attempts to solve problems in machine learning, data mining, information retrieval, classification, information extraction and natural language processing (NLP) [6].

Natural language processing (NLP) is a field of research that studies the problems of computer analysis and synthesis of natural languages. In the future NLP will allow to create a convenient form for human-computer interaction [7]. NLP technologies can help extract valuable information from medical texts and electronic medical records [8].

For modern languages, primarily English, text mining and natural language processing technologies are best developed and are constantly being improved [9]. There are a large number of methods for Text Mining in English, such as Information Retrieval, Information Extraction, Opinion Mining, Sentiment Analysis, Question Answering, Machine Translation, etc [10]. However, there is a drawback of quality methods for other languages, in particular, for Russian. Especially for such a specific subject area as medicine. Creating applications for processing the Russian language is now important. For example, there is an NLPub resource where you can view existing projects on natural language processing [11]. The research [12] presents a comprehensive system of intelligent data processing in a multidisciplinary pediatric center. The methods used in the system for extracting information from clinical tests and analyzing medical data are presented. An experimental study of the developed system based on data from a multidisciplinary pediatric center was conducted. In the study, the novel TRE model performed best with the result F = 72.1%.

In the process of intellectual analysis of texts, there are three stages: collection of text data, pre-processing and analysis [13]. The stages of data collection and pre–processing are the most time-consuming and complex in text analysis [14]. Therefore, in this article, we aim to demonstrate the effectiveness of each stage of pre-processing.

Pre-processing and clearing text data are important tasks that must be completed before data can be used for model training. These tasks are part of the data processing and analysis process and usually involve an initial study of the data set used to determine and plan the necessary pre-processing [15].

Pre-processing helps to bring the text to a form that is convenient for further work, consists of various stages that may differ depending on the task at hand.

The pre-processing method plays an important role and it is the first step in text processing. The task of text pre-processing is to clear the text and bring it to a format suitable for computer processing [16]. Pre-processing, depending on the goals set, includes the following steps and procedures, applied in various combinations and sequences. The pre-processing stage is crucial for determining the quality of classification. In this article, we consider the main stages of pre-processing: tokenization, deletion of stop words, error correction, document cropping, normalization, class harmonization, and vectorization. Normalization is necessary, for example, for fast word search in dictionaries, syntactic and semantic parsing of the text [17]. The Russian language has a rich morphology (strong inflections due to grammatical variation) [18], so normalization is especially relevant for Russian-language text. Stop words are usually understood as articles, interjections, conjunctions, etc., which do not carry a semantic load, so you need to get rid of irrelevant words.

The most common way to classify data is logistic regression (LR) [19] because of its versatility and ease of interpretation. Logistic regression is very easy to train and its results can be interpreted, since you can easily extract all the most important coefficients from the model [20]. For our work, the method of logistic regression was chosen to classify patients' allergic anamneses.

## 2. Methods

### *2.1. Dataset*

Dataset were provided by the Federal State Budgetary Institution “V.A. Almazov National Medical Research Center” of the Ministry of Health of the Russian Federation. 269923 records of allergic anamnesis of patients. Of these, 11670 data were marked up and tagged, such as:

A - record contains information about allergies;
AL - record contains an allergen;
R - record refers to an allergic reaction;
N - record does not apply to allergies;
NN - record indicates a lack of allergies and intolerance.

### *2.2. Pre-processing*

The algorithm includes:

- search for entries by keywords: «allergy» and «intolerance»
- deleting duplicate entries
- removing characters and extra spaces
- regular expression error correction;
- normalization;
- removal of stop words;
- vectorization;
- harmonization of the classes.

We were interested in records containing information about allergies and intolerance. But since medical records contain a large amount of information about the patient's health, there was a need to highlight sentences containing keywords such as "аллергия" (allergy) and "непереносимость" (intolerance).

After that, we identified unique records, thereby eliminating possible repetitive records, to avoid processing the same sentences.

Since the text enters the program in an unprocessed form that contains unnecessary information for our task, it is necessary to remove all unnecessary information from it before processing begins. The first task of this stage is to clear the text. Technical marks, time, date, conjunctions and prepositions, punctuation marks, and other data that are not of value in the research are removed from the text.

Regular expressions were used to:

- put spaces between numbers and letters (18mg - 18 mg);
- words written together were separated, when each word is written with a capital letter (YellowRed - Yellow Red);
- removed spaces when writing dates in the format dd. mm. yyyy;
- dots are placed between two words when the second word is capitalized (Red apple Yellow - Red apple. Yellow);
- put spaces between words written together if one of the words inside unseparated word is written with a capital letter (yellowREDblue - yellow RED blue).

When normalizing the text, all words in the text were converted to the dictionary form: to the nominative case, singular (if any), or infinitive for verbs and all words were reduced to a same letter case.

We used the NLTK library, which has a preset list of stop words.

After the text has been converted to a cleared normalized sequence of words, started the process of their vectorization – conversion to numeric vectors. As a vectorization method has been chosen Bag-of-Words (BoW) [21]. BoW is a detailed representative model for simplifying the processing of text content. It does not take into account grammar or word order and is mainly used to determine the number of occurrences of individual words in the analyzed text.

This article uses the SMOTE algorithm [22] for the harmonization of the classes to create synthetic examples in the data sample.

### *2.3. Classification algorithm*

Logistic regression (LR) is a probabilistic method for analyzing data. LR returns the probability that a vector belongs to a particular class. Logistic regression was used to classify the main types of medical communications. LR works well with large-size data (Bag-of-Words).

### *2.4. Metrics for evaluating of classification*

Precision, recall and F-measures are compared for processed and unprocessed data, to evaluate the effectiveness of pre-processing in classifying medical texts.

$$Precision = \frac{TP}{TP+FP} \tag{1}$$

$$Recall = \frac{TP}{TP+FN} \tag{2}$$

$$F-measure = \frac{2*Precison*Recall}{Precision+Recall} \tag{3}$$

TP and TN correspond to the number of true positive and true negative results, FP and FN, on the other hand, correspond to the number of false positive and false negative results.

### *2.5. Experiments*

During the experiments, the next stage of preprocessing was added to each stage sequentially, then the metric values were compared before and after the function was applied. The results are listed in table 1 in hierarchical order.

## 3. Results

In figure 1 shows how regular expressions work using clippings from allergological anamnesis.


Fig. 1. Regular expressions error correction.

Figure 2 shows an example of anamnesis processing: text after normalization and after removing stop words.

| | ANAMNESES | NORMALIZATION | STOPWORDSREMOVING |
|---|---|---|---|
| 0 | Аллергологический анамнез без особенностей (Allergological anamesis without features) | аллергологический анамнез без особенность (allergological anamesis without a feature) | аллергологический анамнез без особенность (allergological anamesis without a feature) |
| 1 | Аллергологический анамнез без особенностей (Allergological anamesis without features) | аллергологический анамнез без особенность (allergological anamesis without a feature) | аллергологический анамнез без особенность (allergological anamesis without a feature) |
| 2 | Аллергологический анамнез без особенностей (Allergological anamesis without features) | аллергологический анамнез без особенность (allergological anamesis without a feature) | аллергологический анамнез без особенность (allergological anamesis without a feature) |
| 3 | Непереносимость лекарств отрицает (Drugs intolerance denies) | непереносимость лекарство отрицать (drug intolerance to deny) | непереносимость лекарство отрицать (drug intolerance deny) |
| 4 | Непереносимость лекарств отрицает (Drugs intolerance denies) | непереносимость лекарство отрицать (drug intolerance to deny) | непереносимость лекарство отрицать (drug intolerance deny) |
| ... | ... | ... | ... |
| 11666 | Учитывая неоднократные рецидивы тиреотоксикоз... (Considering repeated relapses of thyrotoxicosis...) | учитывать неоднократный рецидив тиреотоксикоз... (consider repeated relapse of thyrotoxicosis) | учитывать неоднократный рецидив тиреотоксикоз... (consider repeated relapse of thyrotoxicosis) |
| 11667 | У пациентки с кордароиндуцированным тиреоток... (In a patient with cordaron-induced thyrotox...) | у пациентка с кордароиндуцировать тиреотокси... (in a patient with cordarone-induce thyrotox...) | пациентка с кордароиндуцировать тиреотоксикоз... (patient with cordarone-induce thyrotox...) |
| 11668 | Плохо переносит жаркую погоду, одышка без нар... (Does not tolerate hot weather, shortness of breath without nar...) | плохо переносить жаркий погода одышка без нар... (bad to tolerate hot weather shortness of breath without nar...) | плохо переносить жаркий погода одышка без нара... (bad to tolerate hot weather shortness of breath without nar...) |
| 11669 | У пациентки рецидив тиреотоксикоза на фоне отмены (The patient has a relapse of thyrotoxicosis on the background of canceling) | у пациентка рецидив тиретоксикоз на фоне отмены (the patient have a relapse thyrotoxicosis on the background note...) | пациентка рецидив тиретоксикоз на фоне отмена тер... (patient have relapse of thyrotoxicosis background of cancellation...) |
| 11670 | Аллергологический анамнез аллергия на поливал... (Allergological anamesis allergy polyva...) | аллергологический анамнез аллергия на поливал... (allergological anamesis allergy polyva...) | аллергологический анамнез аллергия на поливалент... (allergological anamesis allergy polyva...) |

Fig. 2. anamnesis processing.

The results of experimental analysis of the classification dataset are illustrated in table 1.

Table 1. Results of experiments

| Stages of Preprocessing | Precision (%) | Recall (%) | F – measure (%) |
|---|---|---|---|
| None | 80,00 | 88,50 | 87,69 |
| Error correction | 80,50 | 88,79 | 88,01 |
| Normalization | 80,68 | 88,92 | **88,12** |
| Stop-words | 80,63 | 88,88 | 88,10 |
| SMOTE | 80,57 | 88,83 | 88,06 |

Error Correction in the text using regular expressions (figure 1) improved the quality of classification. After applying normalization, the metric values also increased (F = 88.12%). However, after deleting the stop words, the

values decreased slightly. The processed text after applying normalization and deleting stop words is shown in figure 2. Harmonization of classes using the SMOTE algorithm also reduced the quality of classification. To improve the quality of the SMOTE algorithm, you need to carefully pre-process the data, including correcting typos in the text. Correcting typos in the Russian-language text is a separate big task for language processing.

## 4. Discussion

Normalization included translating all words to the same case, bringing the words to their initial form. The "Bag-of-words" method was used for vectorization. We have harmonized SMOTE classes using the algorithm. To evaluate the effectiveness of preprocessing in classifying medical texts, precision, recall, and F-measures are compared for processed and unprocessed data. As a result, the most effective steps were normalizing and correcting errors; deleting stop words and SMOTE, on the contrary, reduced the values of metrics.

The results obtained will be used in the future to extract allergens and reactions from medical records.

## Acknowledgements

We thank V.A. Almazov National Medical Research Center for providing the dataset used in paper. This work financially supported by the government of the Russian Federation through the ITMO fellowship and professorship program and Russian Fund for basic research: 18-07-00870

## References

[1] WHO | World Health Organization, (n.d.). https://www.who.int/ (accessed May 19, 2020).

[2] A.M. Pope, Indoor allergens - Assessing and controlling adverse health effects, 1993. https://doi.org/10.1001/jama.1993.03500210021010.

[3] M. Lehne, J. Sass, A. Essenwanger, J. Schepers, S. Thun, Why digital medicine depends on interoperability, Npj Digit. Med. 2 (2019) 1–5. https://doi.org/10.1038/s41746-019-0158-1.

[4] P.B. Jensen, L.J. Jensen, S. Brunak, Mining electronic health records: Towards better research applications and clinical care, Nat. Rev. Genet. 13 (2012) 395–405. https://doi.org/10.1038/nrg3208.

[5] R. Talib, M. Kashif, S. Ayesha, F. Fatima, Text Mining: Techniques, Applications and Issues, Int. J. Adv. Comput. Sci. Appl. 7 (2016) 414–418. https://doi.org/10.14569/ijacsa.2016.071153.

[6] P. Shinde, S. Govilkar, A Systematic study of Text Mining Techniques, Int. J. Nat. Lang. Comput. 4 (2015) 54–62. https://doi.org/10.5121/ijnlc.2015.4405.

[7] B. Manaris, Natural Language Processing: A Human-Computer Interaction Perspective, Adv. Comput. 47 (1998) 1–66. https://doi.org/10.1016/S0065-2458(08)60665-8.

[8] W. Sun, Z. Cai, Y. Li, F. Liu, S. Fang, G. Wang, Data processing and text mining technologies on electronic medical records: A review, J. Healthc. Eng. 2018 (2018). https://doi.org/10.1155/2018/4302425.

[9] A. Névéol, H. Dalianis, S. Velupillai, G. Savova, P. Zweigenbaum, Clinical Natural Language Processing in languages other than English: Opportunities and challenges, J. Biomed. Semantics. 9 (2018) 1–13. https://doi.org/10.1186/s13326-018-0179-8.

[10] S. Singh, Natural Language Processing for Information Extraction, (2018) 1–24. http://arxiv.org/abs/1807.02383.

[11] NLPub, (n.d.). https://nlpub.ru/ (accessed May 19, 2020).

[12] R. Sergienko, T. Gasanova, E. Semenkin, W. Minker, Text categorization methods application for natural language call routing, ICINCO 2014 - Proc. 11th Int. Conf. Informatics Control. Autom. Robot. 2 (2014) 827–831. https://doi.org/10.5220/0005139708270831.

[13] S. Shah, N. Rathod, P.K. Saini, V. Patel, H. Rajput, P. Sheth, Soft Computing: Theories and Applications, 2019. https://doi.org/10.1007/978-981-13-0589-4.

[14] M.E. Megel, J.A. Heermann, Methods of data collection, Plast. Surg. Nurs. 14 (1994) 109–110. https://doi.org/10.5005/jp/books/13075_10.

[15] A. Famili, W.-M. Shen, R. Weber, E. Simoudis, Data Preprocessing and Intelligent Data Analysis, Intell. Data Anal. 1 (1997) 3–23. https://doi.org/10.3233/IDA-1997-1102.

[16] Z. Jianqiang, G. Xiaolin, Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis, IEEE Access. 5 (2017) 2870–2879. https://doi.org/10.1109/ACCESS.2017.2672677.

[17] B. Han, P. Cook, T. Baldwin, Lexical Normalization for Social Media Text, ACM Trans. Intell. Syst. Technol. 4 (2013). https://doi.org/10.1145/2414425.2414430.

[18] A. Mikheev, L. Liubushkina, Russian morphology: An engineering approach, Nat. Lang. Eng. 1 (1995) 235–260. https://doi.org/DOI:

10.1017/S135132490000019X.

[19] D.R. Cox, The Regression Analysis of Binary Sequences, J. R. Stat. Soc. Ser. B. 20 (1958) 215–232. https://doi.org/10.1111/j.2517-6161.1958.tb00292.x.

[20] R.E. -, A.R. -, B.M. -, Comparison of Classification Methods Based on the Type of Attributes and Sample Size, J. Converg. Inf. Technol. 4 (2009) 94–102. https://doi.org/10.4156/jcit.vol4.issue3.14.

[21] R. Zhao, K. Mao, Fuzzy Bag-of-Words Model for Document Representation, IEEE Trans. Fuzzy Syst. 26 (2018) 794–804. https://doi.org/10.1109/TFUZZ.2017.2690222.

[22] W.P.K.N.V.C.K.W.B. Lawrence O. Hall, snopes.com: Two-Striped Telamonia Spider, J. Artif. Intell. Res. 2009 (2006) 321–357. https://doi.org/10.1613/jair.953.