engineering proceedings

MDPI

Proceeding Paper

Extracting and Processing of Russian Unstructured Clinical Texts for a Medical Decision Support System [†]

Irina Bolodurina [1,2], Alexander Shukhman [1], Leonid Legashev [1,*], Lyubov Grishina [1] and Arthur Zhigalov [1]

[1] Research Institute of Digital Intelligent Technologies, Orenburg State University, Prosp. Pobedy 13, 460018 Orenburg, Russia
[2] Department of Public Health and Healthcare, Orenburg State Medical University, Sovetskaya Street 6, 460000 Orenburg, Russia
[*] Correspondence: silentgir@gmail.com

[†] Presented at the 15th International Conference “Intelligent Systems” (INTELS’22), Moscow, Russia, 14–16 December 2022.

Abstract: The rapid growth in the volume of medical data is pushing the development and implementation of artificial intelligence (AI) tools. One of the directions of the application of AI in the field of healthcare is the use of natural language processing methods to build medical decision support systems based on electronic medical record (EMC) data. As a result of this study, a module for the extraction and pretreatment of patients’ EMC was developed. In addition, an approach was implemented to extract features from the unstructured textual information of patient admission protocols, with the formation of an appropriate vector representation of data. Predictive models for the diagnosis of groups of diseases based on the logistic regression model and BERT were developed. The highest efficiency in the experiments was shown by the logistic regression model, with a F1-score of 0.81 and Matthews correlation coefficient of 0.75. The obtained results have been posted for public access based on the django framework and can be used for preliminary assessment of patient health status, as well as integrated into existing medical decision support systems.

Keywords: electronic health records; medical decision support system; natural language processing; BERT; logistic regression

check for updates

Citation: Bolodurina, I.; Shukhman, A.; Legashev, L.; Grishina, L.; Zhigalov, A. Extracting and Processing of Russian Unstructured Clinical Texts for a Medical Decision Support System. *Eng. Proc.* **2023**, *33*, 41. https://doi.org/10.3390/engproc2023033041

Academic Editors: Askhat Diveev, Ivan Zelinka, Arutun Avetisyan and Alexander Ilin

Published: 26 June 2023

1. Introduction

Cardiovascular diseases (CVD) continue to be the most urgent health problem in most countries of the world, including in the Russian Federation. In 2020, cardiovascular diseases became the most common cause of death (47%) and claimed the lives of more than 900 thousand Russians. In this regard, it is necessary to develop new approaches to reducing this indicator.

From an economic point of view, the direct costs of public health for the treatment and diagnosis of CVD amount to about RUB 220 billion. This indicator is 8-times higher than the cost of screening and prevention, with which 40% of CVD cases can be prevented with a proper assessment of the risks of development [1,2].

The large growth of medical data is pushing the development of AI tools, for implementation, processing, and analysis. One of the directions of the application of AI in the field of healthcare is the use of NLP methods to build systems to support medical decision-making based on electronic medical records. One of the tasks of a medical decision-making system is the task of determining a diagnosis according to the ICD and based on patient complaints. Thus, the task of multiclass classification based on the text documents of the EMC arises.

2. Related Work

Currently, natural language processing (NLP) methods allow analyzing unstructured information and building highly efficient AI models [3]. In this regard, scientists around the world are engaged in the development and application of NPL methods in the field of digital medicine.

Thus, the study in [4] presented an approach to the processing and analysis of electronic medical records (EHR) of patients based on NLP and deep learning methods for prediction in healthcare. The presented methodology can be used to evaluate various health indicators and in subsequent decision-making. In [5], this approach was also highlighted as the main tool for developing end-to-end applications using multimodal data (images, quantitative analysis data, etc.).

The authors of the study in [6] proposed a deep learning model for predicting heart failure according to EHR data in the UK. However, the resulting model in testing demonstrated an AUROC equal to 0.6965, which generally does not correspond to predictive models of high accuracy.

An effective convolutional neural network (CNN) model for estimating the costs and duration of hospital stays was presented in [7]. The peculiarity of this model is its ability to extract potential knowledge from clinical data with low-frequency medical events.

In [8], an algorithm for predicting a diagnosis based on a deep neural network and by analyzing the data of the EMC of a department of pediatrics was proposed. In their study, the authors used an unstructured and unbalanced data set to build a model using bidirectional recurrent neural networks. The accuracy of the predictive model was 80.9 according to the precision metric.

The study in [9] presented an approach to deep learning for identifying risk factors for cardiovascular diseases based on EHR analysis. The experimental results showed that the proposed models for the binary classification of the presence of CVD using several individual factors (smoking, diabetes, genetic predisposition, etc.) had a high accuracy of prediction (from 0.81 to 0.96).

Thus, at the moment, research in the field of the diagnosis of diseases and risk factors is often based on NLP and deep learning methods. In addition, the results of evaluating the quality of AI models for solving similar problems showed high predictive power. In this regard, within the framework of this work, a study was conducted on the effectiveness of various models for classifying groups of diagnoses of diseases based on the textual information of patient complaints from EHRs.

3. Problem Statement

At the moment, a medical information system (MIS) is a comprehensive software product, the main purpose of which is to automate the main processes related to the work of medical institutions of general and narrow specialization.

The problem of developing an intelligent decision support system (DSS) for the operational interaction of the patient and the doctor at the reception is as follows: the MIS databases with information about the protocols of visits, the results of additional examinations, etc., are stored in a distributed manner, and additional tools for extraction, transformation, and structuring are needed for the implementation of AI models.

In this regard, within the framework of this study, a methodology of interaction with the regional MIS is presented from the stage of information extraction to the implementation of forecasting results, which is schematically presented in Figure 1.

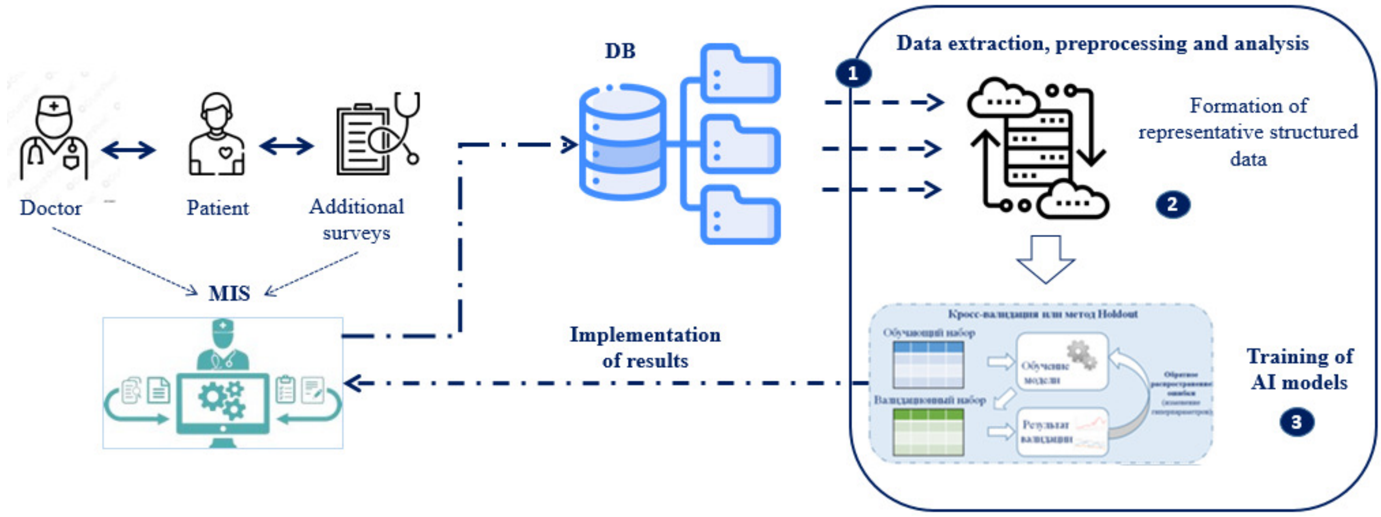

DB
Data extraction, preprocessing and analysis
1
Formation of representative structured data
2
Doctor
Patient
Additional surveys
MIS
Implementation of results
Кросс-валидация или метод Holdout
Обучающий набор
Обучение модели
Валидационный набор
Результат валидации
Обратное распространение ошибки (изменение гиперпараметров)
Training of AI models
3

Figure 1. Methodology of the development and implementation of models in MIS.

Based on the presented scheme, there are three main modules for the development of intelligent DSS:

1. A module for extracting and preprocessing large depersonalized data of the electronic medical records of patients;
2. A module for extracting signs from the unstructured textual information of patient visits to a medical organization and for forming a vector representation of data;
3. A module for diagnosing disease groups, to support medical decision-making.

The data storage structure describes for each patient a “case diagnosis” and a “treatment step”, which are related to the actual “services rendered“. For patients with a history of CVD, within the framework of this study, all protocols of visits to a medical organization, including diagnoses of other groups of diseases, were uploaded. Thus, the purpose of this study is to extract and process unstructured clinical texts in Russian, to build a prognosis of a group of diseases and integrate the results into a medical decision support system.

4. Development of a Module for Extracting and Processing Electronic Health Records

The Medical Information and Analytical Center (MIAC) of Orenburg provided an opportunity to connect to the regional MIS to download depersonalized data from their server. After a preliminary analysis of the data, more than 1 million records of various protocols of medical visits were found, with institutions of patients with CVD (diaries of patient appointments, conclusions of additional examinations (Electrocardiography, blood test, etc.).

For the provided xlm documents, it is necessary to automatically recognize the most informative blocks suitable for building AI models. A characteristic problem of this stage is the availability of documents of various structures - due to the possibility of correction by the doctor of the template of visit protocols, individual concepts for filling in information about additional medical examinations. laboratories, etc. Thus, it is necessary to develop a unified approach to the processing of heterogeneous documents and their informative blocks of textual information.

Within the framework of this, study, methods of parallel reading and processing of the data stream, models and methods of deep learning, as well as NLP text information processing methods were used to build DSS in the diagnosis of groups of diseases.

4.1. Extracting MIS Data

When visiting a patient of a medical organization, the doctor fills out a diary, which consists of data from the objective study of the patient, anamnesis of life, and the complaints of the patient. The anamnesis of life contains information about heredity, bad habits, etc. The generated xml documents with patient examination data do not have a single structure and are modified directly by the doctor.

In this regard, for the development of predictive AI models, the following modules of interaction with the regional MIS of the city of Orenburg were implemented: the XMLParseModule module loads impersonal protocols via the MIAC API in xml format; and for processing heterogeneous templates, a DictParseModule module for automatic conversion of xml documents was developed.

The DictParseModule module for extracting information from heterogeneous xml protocols is based on an approach to recursive node search, with sequential analysis of the contents (Figure 2). A distinctive feature of the proposed approach is the creation of a service record tree in the MOD, which allows analyzing the relationship of certain factors within the document.

Thus, the presented modules convert the xml documents of patient visits into a json file that contains information about complaints at the reception, test results, lifestyle information, etc. into the "key" format:value". As a result of the work of this xml parser module for September–December 2021, 364,020 protocols were uploaded in xml format for patients diagnosed with CVD. The volume of xml files ranged from 3 KB to 1008 KB. The dataset was preprocessed; missing values and records in which the length of the patient's complaint line was less than 100 characters were removed. The final distribution of patient complaint protocols at the reception by disease group is shown in Figure 3.

It should be noted that in addition to cardiovascular diseases, the control group of patients also reported "Acute respiratory infections of the upper respiratory tract" (J0), as well as "New diagnoses of unclear etiology" (U0). These diagnoses were considered in the general order and included in the predictive model.

4.2. Text Information Preprocessing

For AI models, the EMC data obtained after processing xml documents in the form of textual information had to be represented at the input by a feature vector. Let us consider several approaches to feature extraction.

At the first stage, we perform numerical encoding of the target variable—the names of six groups of diseases according to the ICD—and also determine the dictionary of stop words from the Russian-language corpus of the nltk library and the minimum and maximum length of n-grams from 1 to 4.

The first approach to extracting features is as follows: First, operations are performed to convert tokens to lowercase, and remove punctuation, stop words, accents, etc. Next, a collection of unstructured text documents with patient complaints is converted into a matrix of the number of tokens using the CountVectorizer method (bag-of-words model). The resulting vector text embeddings are divided into training and test samples and are used subsequently to train a logistic regression model with cross-validation support.

In addition, within the framework of this study, an approach using Russian-language models of BERT transformers on unstructured medical texts was considered, which consisted in forming a vector representation of tokens of medical texts. In this case, the maximum size of the dictionary num_words = 15,000 and the maximum message length max_len = 200 in tokens were set, and then the sentences of the original dataset were aligned to the same length (padding='post'). Tokenization of the training sample was performed using the EnRuDR-BERT model [10], pretrained on a collection of consumer reviews about taking medications. To solve the problem of classifying a group of diseases, the output layer is represented by six outputs in accordance with the ICD codes. Thus, as a result of training, an attention mask is created for each sample: those tokens that need to be taken into account when training and calculating gradients are filled with units, those tokens that should be skipped are filled with zeros.

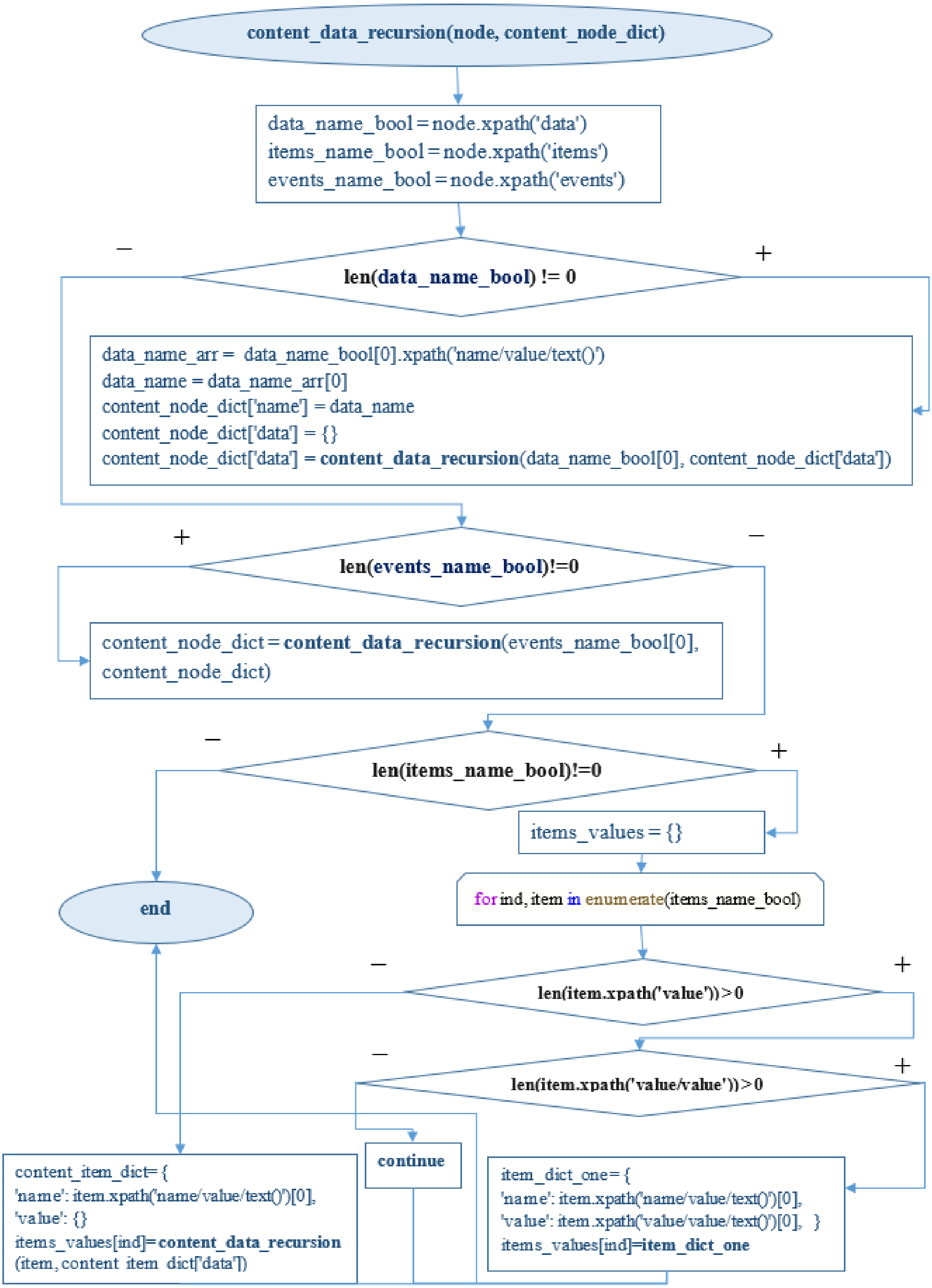

content_data_recursion(node, content_node_dict)
data_name_bool = node.xpath('data')
items_name_bool = node.xpath('items')
events_name_bool = node.xpath('events')
–
+
len(data_name_bool) != 0
data_name_arr = data_name_bool[0].xpath('name/value/text()')
data_name = data_name_arr[0]
content_node_dict['name'] = data_name
content_node_dict['data'] = {}
content_node_dict['data'] = content_data_recursion(data_name_bool[0], content_node_dict['data'])
+
–
len(events_name_bool)!=0
content_node_dict = content_data_recursion(events_name_bool[0], content_node_dict)
–
+
len(items_name_bool)!=0
items_values = {}
end
for ind, item in enumerate(items_name_bool)
–
+
len(item.xpath('value'))>0
–
+
len(item.xpath('value/value'))>0
content_item_dict= {
'name': item.xpath('name/value/text()')[0],
'value': {}
items_values[ind]=content_data_recursion
(item, content item dict['data'])
continue
item_dict_one= {
'name': item.xpath('name/value/text()')[0],
'value': item.xpath('value/value/text()')[0], }
items_values[ind]=item_dict_one

Figure 2. DictParseModule for automatic xml document conversion.

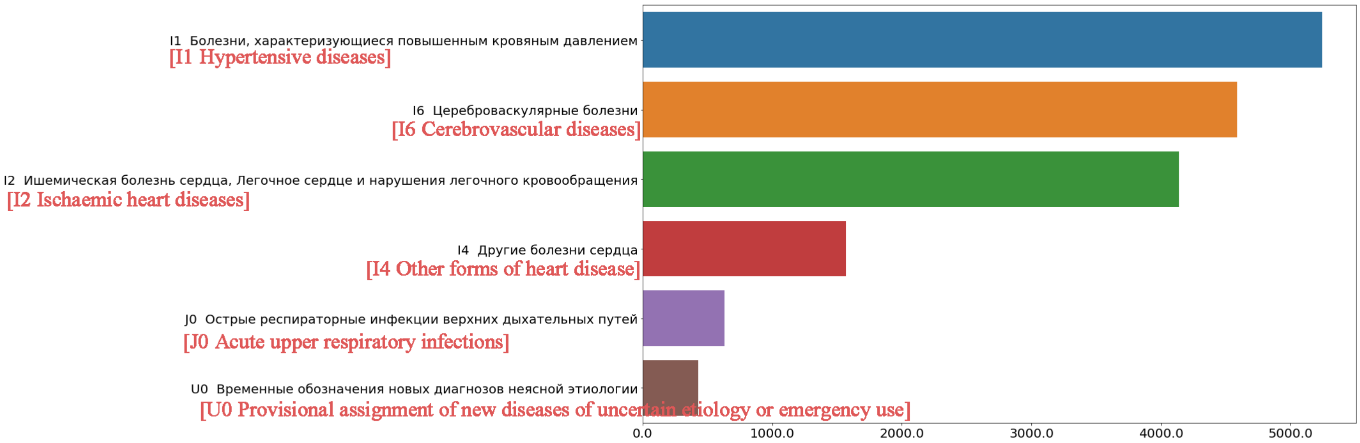


Figure 3. Distribution of patient complaint protocols by disease group.

5. The Training of AI Models

The next stage of constructing a prognosis of a patient's disease group is the training of AI models. Schematically, the process of learning a logistic regression model for classifying groups of diseases is shown in Figure 4. Note that for each sample object, when using this approach, the probability of belonging to one of the six groups of diseases according to the ICD is calculated.

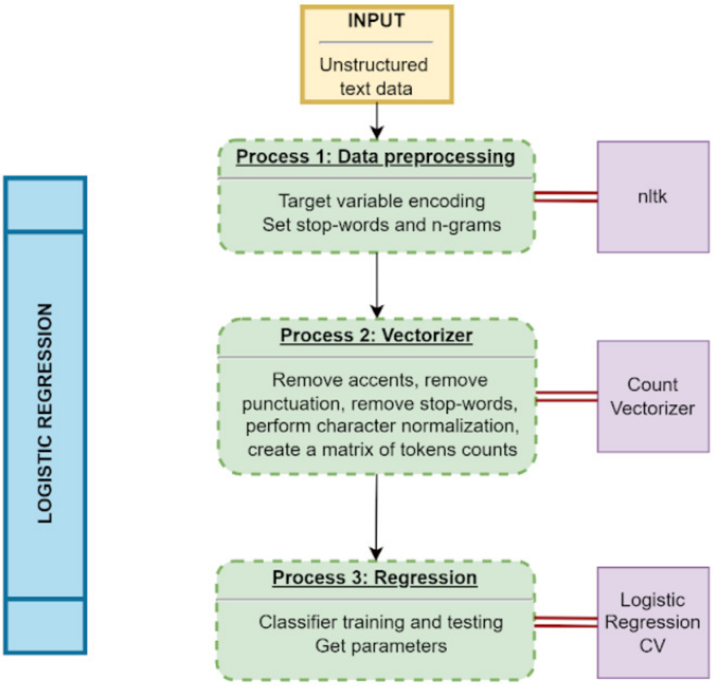

INPUT
Unstructured text data
Process 1: Data preprocessing
Target variable encoding
Set stop-words and n-grams
nltk
Process 2: Vectorizer
Remove accents, remove punctuation, remove stop-words, perform character normalization, create a matrix of tokens counts
Count Vectorizer
Process 3: Regression
Classifier training and testing
Get parameters
Logistic Regression CV
LOGISTIC REGRESSION

Figure 4. General scheme of the logistic regression approach.

The process of training a BAT-based model is shown in Figure 5. Embeddings are formed using the input layer of the neural network based on a list of dictionary numbers of text tokens. The model was trained and tested. The number of epochs was selected experimentally (epoch = 2). As a result, the error on the training and test dataset had the following values: train_loss: 0.5425, val_loss: 0.5644. The softmax function of the torch library was used to obtain the predicted probability of a sample belonging to one of the six groups of diseases according to the ICD.

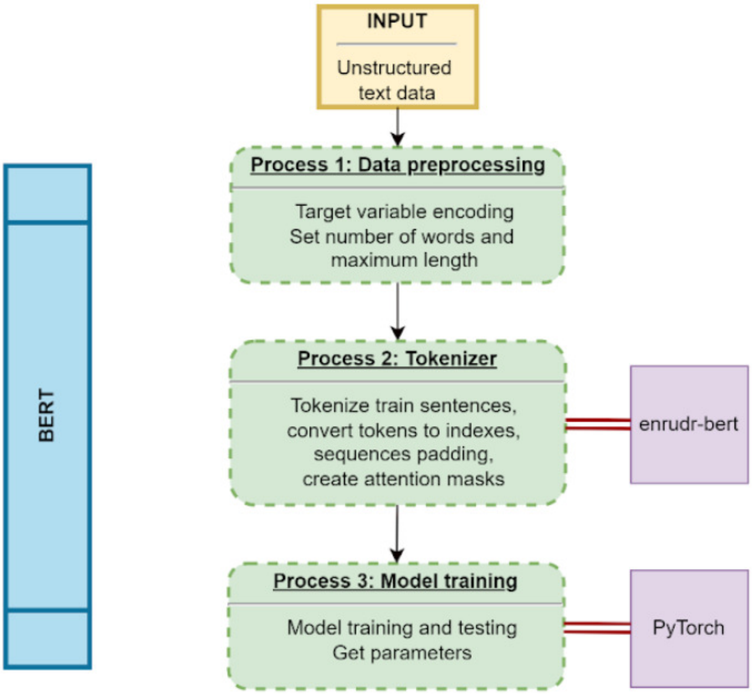

INPUT
Unstructured text data
Process 1: Data preprocessing
Target variable encoding
Set number of words and maximum length
Process 2: Tokenizer
Tokenize train sentences, convert tokens to indexes, sequences padding, create attention masks
enrudr-bert
Process 3: Model training
Model training and testing
Get parameters
PyTorch
BERT

Figure 5. General scheme of the BERT-based approach.

The results of comparing the proposed approaches to predicting groups of diseases using precision, F1-score, and Matthews correlation coefficient (MCC) metrics are presented in Table 1.

The logistic regression approach showed the best results for all metrics. At the same time, the BERT-based approach functioned a little worse, which may indicate the need to retrain the model on specialized medical texts.

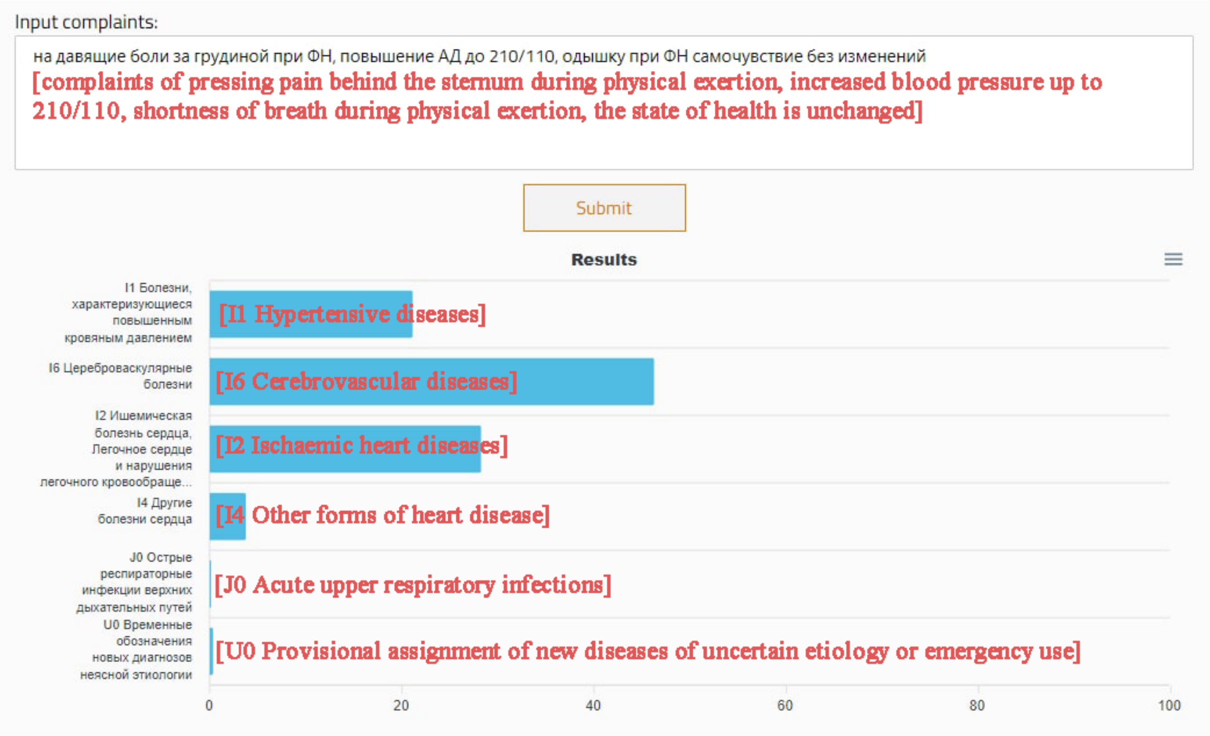
Table 1. An example of a table.

Algorithm	Precision	F1-Score	MCC
Logistic Regression	0.8187	0.8161	0.7551
BERT	0.8095	0.8088	0.7450

6. Implementation of a Medical Decision Support System Prototype

A demo version of the logistic regression model is available for general use at http://osudeepai.com/services/disease-ml (accessed on 4 June 2022), implemented using the django framework. The characteristics of the software provided by the provider are as follows: Intel(R) Xeon(R) Gold 6240R processor, 2.40 GHz CPU and 128 GB RAM.

Examples of the probability distribution of AI model classes for the complaints of patients with coronary heart disease are shown in Figure 6. According to the therapist's comments, the spread of probability classes occurred due to the fact that the listed complaints in some cases may relate to several groups of diseases, and in practice the doctor makes the final decision based on personal experience and, possibly, the results of additional examinations.



Input complaints:

на давящие боли за грудиной при ФН, повышение АД до 210/110, одышку при ФН самочувствие без изменений
[complaints of pressing pain behind the sternum during physical exertion, increased blood pressure up to 210/110, shortness of breath during physical exertion, the state of health is unchanged]

Submit
Results
I1 Болезни, характеризующиеся повышенным кровяным давлением
[I1 Hypertensive diseases]
I6 Цереброваскулярные болезни
[I6 Cerebrovascular diseases]
I2 Ишемическая болезнь сердца, Легочное сердце и нарушения легочного кровообраще...
[I2 Ischaemic heart diseases]
I4 Другие болезни сердца
[I4 Other forms of heart disease]
J0 Острые респираторные инфекции верхних дыхательных путей
[J0 Acute upper respiratory infections]
U0 Временные обозначения новых диагнозов неясной этиологии
[U0 Provisional assignment of new diseases of uncertain etiology or emergency use]
0
20
40
60
80
100

Figure 6. An example of the demo of the DSS based on logistic regression.

7. Conclusions

As a result of this study, a predictive model for the diagnosis of disease groups based on a logistic regression model was developed, which in the experiments showed a value of 0.81 for F1-score and 0.75 for MCC. To train the model, depersonalized regional MIS data obtained by extracting and preprocessing the patients' EMC were used. In addition, an approach to extracting features from the unstructured textual information of patient admission protocols and the formation of an appropriate vector representation of data was additionally implemented. The presented model of disease group prediction can be used for preliminary assessment of a patient's health status and also integrated into existing medical decision support systems. In the future, it is planned to implement a separate AI model for the DSS, which will check the data entered by the user for its relevance in relation to the service used. In addition, one of the areas of further research includes expanding the data set with examples of protocols with other groups of diseases, to scale the results obtained.

Author Contributions: Supervision, I.B.; conceptualization, I.B.; methodology, A.S.; software, A.Z.; investigation, L.L.; validation, A.Z.; writing—original draft preparation, L.G., writing—review and editing, L.G. All authors have read and agreed to the published version of the manuscript.

Funding: The research was carried out within the framework of the Priority 2030 program (Agreement No. 075-15-2021-1171/2 dated 11 May 2022), as well as scholarships of the President of the Russian Federation to young scientists and graduate students (No. SP-919.2022.5).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data sharing not applicable due to privacy restrictions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Federal State Statistics Service. Health Care. 2021. Available online: https://rosstat.gov.ru/folder/13721 (accessed on 3 June 2022).
2. British Heart Foundation. Heart and Circulatory Disease Statistics. 2021. Available online: https://www.bhf.org.uk/what-we-do/our-research/heart-statistics/heart-statistics-publications/cardiovascular-disease-statistics-2020 (accessed on 3 June 2022).
3. Basques, R. What Is Natural Language Processing? 2020. Available online: https://towardsdatascience.com/what-is-natural-language-processing-86a7123a076b (accessed on 3 June 2022).
4. Jain, K.; Prajapati, V. NLP/Deep Learning Techniques in Healthcare for Decision Making. *Prim Health Care* **2021**, *11*, 1–4.
5. Zhou, B.; Yang, G.; Shi, Z.; Ma, S. Natural Language Processing for Smart Healthcare. 2021. Available online: http://arxiv.org/abs/2110.15803 (accessed on 4 June 2022).
6. Denaxas, S.; Stenetorp, P.; Riedel, S.; Pikoula, M.; Dobson, R.; Hemingway, H. Application of Clinical Concept Embeddings for Heart Failure Prediction in UK EHR Data. 2018. Available online: http://arxiv.org/abs/1811.11005 (accessed on 4 June 2022).
7. Feng, Y.; Min, X.; Chen, N.; Chen, H.; Xie, X.; Wang, H.; Chen, T. Patient outcome prediction via convolutional neural networks based on multi-granularity medical concept embedding. In Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Kansas City, MO, USA, 13–16 November 2017; pp. 770–777.
8. Shi, J.; Fan, X.; Wu, J.; Chen, J.; Chen, W. DeepDiagnosis: DNN-Based Diagnosis Prediction from Pediatric Big Healthcare Data. In Proceedings of the Sixth International Conference on Advanced Cloud and Big Data (CBD), Lanzhou, China, 12–15 August 2018; pp. 287–292.
9. Chokwijitkul, T.; Nguyen, A.N.; Hassanzadeh, H.; Perez, S. Identifying Risk Factors for Heart Disease in Electronic Medical Records: A Deep Learning Approach. In Proceedings of the BioNLP 2018 Workshop 2018, Melbourne, Australia, 19 July 2018; pp. 1–10.
10. Tutubalina, E.; Alimova, I.; Miftahutdinov, Z.; Sakhovskiy, A.; Malykh, V.; Nikolenko, S. The Russian Drug Reaction Corpus and Neural Models for Drug Reactions and Effectiveness Detection in User Reviews. *Bioinformatics* **2020**, *37*, 243–249. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.