EXTRACTION OF MEDICAL DATA FROM ELECTRONIC MEDICAL RECORDS USING NLP ALGORITHMS

[a]ALEKSANDR V. GUSEV, [b]ROMAN E. NOVITSKIY,
[c]ALEKSANDR A. IVSHIN, [d]JULIIA S. BOLDINA,
[e]ALEKSEY S. SHTYKOV, [f]ALEKSEY S. VASILEV

[a,b]K-SkAI LLC, 20 Premises, 17 Naberezhnaya Varkausa, Petrozavodsk, Republic of Karelia, Russia, 185031
[c,d,e,f]Petrozavodsk State University, Lenina ave., 33 Petrozavodsk, Republic of Karelia, Russia, 185910
email: [a]agusev@webiomed.ai, [b]roman@webiomed.ai, [c]scipeople@mail.ru, [d]juliaisakova-20@mail.ru, [e]shtykoff@petrsu.ru, [f]alvas@petrsu.ru

Acknowledgements: This research was financially supported by the Ministry of Science and Higher Education of the Russian Federation Theme No. 075-15-2021-665. This study was performed using the Unique Scientific Unit (UNU) «Multicomponent software and hardware system for automated collection, storage, markup of research and clinical biomedical data, their unification and analysis based on Data Center with Artificial Intelligence technologies (reg. number: 2075518).

Abstract: Development of artificial intelligence methods in medicine requires large volumes of input data available. The source of this data is electronic health records. Extraction of data from health records is accompanied by a number of difficulties, mainly associated with their being filled out in any form and doctors using various abbreviations when putting down the information. The paper describes a method of processing electronic health records which allows extracting the necessary information from them – the one needed for building work algorithms of artificial intelligence software complexes and their learning. The method was developed and tested out on electronic health records filled out in Russian. For working with medical documents filled out in other languages, it needs no special adaptation. For this, it is sufficient to change teaching data and perform complete learning of all models.

Keywords: electronic health record, artificial intelligence, patient, prediction of disease development, machine learning.

1 Introduction

To diagnose and predict the development of diseases in medicine, machine learning (ML) methods are increasingly being used as a part of artificial intelligence (AI). Using ML requires a large amount of medical data, called Big Data. This data can be obtained from prospective studies, such as 70 years long Framingham heart study (FHS). The FHS was the first longitudinally-followed large cohort to study cardiovascular disease (CVD) epidemiology in the USA, now including a multigenerational community-based cohort of Framingham population.

Such study requires a lot of time, effort and money. More and more financing of medicine is required every year for medical research, therefore the new less expensive methods to get medical data are on demand. One of them – obtain the medical data from already existing medical documents. Electronic medical records (EMRs) and/or electronic health records (EHRs) can be considered as such medical documents. Currently, a large number of EMRs and EHRs are available, on the one hand, using their content by a physician is too time consuming and costly, and on the other hand, using algorithms to extract medical data and medical facts from them make this task very promising, makes the clinical information contained therein more accessible. The medical data, such as, EMR and EHR contains sufficient information to predict medical events successfully, can be considered sufficiently complete for this purpose (Malmasi et al., 2019).

Natural language processing (NLP) is a subsection of AI that deals with the intellectual processing of everything related to human language. Currently, there are many frameworks and libraries for working with natural languages, that are aimed at solving common problems, for example, for analyzing text, such models are trained on texts taken from newspaper articles or scientific papers. And this explains the many errors and inaccuracies in the operation of NLP algorithms with medical documents, this is due in addition to medical topics and the specifics of the type of documents, as well as using the abbreviations that the physician makes when filling out medical documents. It has often been necessary for a new NLP tool to be developed or adapted for each medical database, and even for each clinical event, when processing EMR free text. This is labor intensive, as it requires the tools to be tested on significant amounts of text already annotated by human experts.

Modern machine learning methods model the temporal sequence of structured events from a patient's clinical record using convolution and recurrent neural networks to predict future health events. There are other approaches extracting knowledge from data through data mining based on the domain ontology. Using the approach of extracting medical data in Russian is difficult because there is no necessary software and labeled medical corpora for this. To use AI methods in medicine we need «big medical data», to get it we need algorithms to extract this data from medical documents.

Re-use of medical data based on their extraction from the EHR has appeared relatively recently and has become used, especially when data is needed to train artificial intelligence algorithms. One of the problems of using this method in Russia is that all medical texts, including EHR, are written in Russian. All available solutions are only developed for English and a small amount for German, French and Chinese. An attempt to translate EHR from Russian into English would lead nowhere since in Russia a different system of clinical guidelines is used. Clinical guidelines are a document based on proven clinical experience on the prevention, diagnosis, treatment, and rehabilitation, including models of patients, actions sequence of a health worker, diagnostics schemes and treatment, complications and comorbidities, and other factors influencing treatment outcomes. There are many NLP libraries for the English language, and there are several medical corpora to work with medical documents. Even there is a scispacy – Python package containing models for processing biomedical, scientific or clinical text.

The major challenge holding back the use of artificial intelligence technologies in medicine is the difficulty of extracting data from medical records. This is particularly pressing in working with EHRs filled out in Russian. With regard to this, the research topic covered in the paper is deemed relevant.

2 Literature Review

Electronic medical records tend to contain a variety of medical information, both structured and in the form of arbitrary text. It is rarely possible to use only one method of extracting information from an electronic medical record. The rules-based method, Classification Machine Learning method, and NLP Name entity recognition (NER) are used in the Webiomed system simultaneously (Gavrilov et al., 2020).

The most common machine learning algorithms for text classification: The Naive Bayes family of algorithms, support vector machines, and deep learning.

The Naive Bayes classifier is one of the basic classification algorithms. However, very often it works as well as, or even better than, more complex algorithms. The advantage of a naive Bayesian classifier is the small amount of data required for training, parameter estimation, and classification. All model parameters can be approximated by relative frequencies from the training data set. These are estimates of the maximum likelihood of probabilities.

Support Vector Machines. A text in the section «Complaints» frequently contains «noise» that might cause inaccuracy while applying the SVM as a text classification method.

Advantages of SVM:

- works well with a large feature space;
- deals well with small data volumes;
- the algorithm maximizes the dividing band reducing like «airbag» the number of classification errors;
- a problem always has a single solution (the dividing hyperplane with certain hyperparameters of the algorithm is always one), since the algorithm is reduced to solving the problem of quadratic programming in a convex domain.

Disadvantages of SVM:

- long learning time for large data sets;
- instability to noise: outliers in the training data become reference objects that violate it and directly affect the construction of the dividing hyperplane;
- general method for constructing kernels and straightening spaces that are most suitable for a specific problem in the case of the linear inseparability of classes are not described. Selecting useful data transformations is an art.

Hybrid Systems. This approach combines all or most of the principles discussed above and consists of applying classifiers based on them in a certain sequence.

To work with text data, unlike rule-based methods, words have to be converted into a vector representation. Unfortunately, there are no universal recommendations for choosing a particular method and a word vectorization method for particular information. Therefore, it has to be checked which vectorization method and which classification method gives the best result in each particular case.

It is difficult to perform a medical text, ignoring features of the language structure, which analyses the proposals, i.e. the essential or identifying relationships between objects of sentences or syntactic analysis of the sentence structure. For these tasks, numerous methods and ready-made software tools for the English language have been developed. There are fewer ready-made solutions for Russian, and many of them are commercial.

Obtaining EHR-derived datasets for COVID-19 used in others research, one of them this work (Pedrera-Jiménez et al., 2021). It showed effective reuse of EHRs in a tertiary Hospital during COVID-19 pandemic. Extracting household patient EHR data proved to be as effective at tracking transmission as COVID-19 contact tracing, according to research (Metlay et al., 2021). Additionally, analytic methods do not always give real-time results, it is easy to overlook or underuse EHR data. Overall, EHR data could support COVID-19 control efforts, so as long as infrastructure and methods are in place to put this to scale.

EHR data extraction errors can be explained by the following problems:

- numerous spelling errors in EHR,
- abbreviations, specific for this medical institution,
- inconsistent sentences, not in accordance with language rules.

The results of automatic text extraction can be improved by using automatic spelling correction methods, improving dictionaries (for dictionary-based methods), and training sampling (for machine learning methods) (Meystre et al., 2008).

Instead of relying on the manually created method based on rules, text classification with the help of machine learning solves the problem of classification based on the marked up training data set. Using pre-marked examples as training and validation data, the machine learning algorithm can identify associations between text excerpts and a particular opinion (diagnosis) expected for the particular input (e.g., the text of complaints). To apply this classifier, machine learning features have to be extracted: the method is used for transforming each text into digital representation in the form of a vector.

Classification results can be improved with the use of boosting technology. Boosting is a procedure for sequentially constructing a composition of machine learning algorithms when each next algorithm seeks to compensate for the shortcomings of the composition of all previous algorithms. Boosting is an optimization algorithm for constructing a composition of algorithms. Initially, the concept of boosting arose in works on probably correct training dealing with the question whether it is possible to obtain a good algorithm from the number of poor (slightly different from random) ones (Mayr et al., 2014).

For the past 10 years, boosting has remained one of the most popular machine learning methods, along with neural networks and reference vector machines. The main reasons are simplicity, versatility, flexibility (the ability to build various modifications), and most importantly, high generalizing ability.

Boosting over Decision Tree is considered one of the most effective methods in terms of classification. In many experiments, there was an almost unlimited decrease in frequency errors on an independent test sample as the composition grows. Moreover, quality on the test sample often continued to improve even after achieving error-free recognition of the entire training samples. This upended long-held perceptions of the fact that to increase generalizing ability, it is necessary to limit the complexity of algorithms.

Of interest is an open-source natural language processing (NLP) package SpaCy. It is written in Python that performs tokenization, Part-of-Speech (PoS) tagging, and dependency parsing. It is the fastest NLP parser available and offers state-of-the-art accuracy. It is presented on the site «Training SpaCy's Statistical models» at https://spacy.io/usage/training.

The most recent extensive evaluation of existing dependency parsers has been performed by. They evaluate 10 different off-the-shelf parsers for accuracy and speed; reporting labeled attachment scores (LAS) of 85% to 90%. While SpaCy does not perform the most accurate in their evaluation, it shows to be fastest maintaining comparable accuracy.

SpaCy models are statistical, and every decision they make, which part of speech tag to assign or whether a word is a named entity is a prediction. This prediction is based on model training examples. To train a model first training data, the special text, and placemarks, the model has to predict, are needed. This can be a part-of-speech tag, a named entity, or any other information. This package also used in Healthcare NER Models (Amogh et al., 2020).

SpaCy does not offer a pre-trained model for the Russian language, but provides an opportunity to conduct a Russian model training. SpaCy 2.0 offers new neural models for tagging, parsing, and entity recognition (Honnibal & Johnson, 2015).

3 Materials and Methods

The objective of this work is to develop a solution which allows extracting specific data from electronic health records using artificial intelligence technologies.

The tasks completed in the course of the research consisted in the following:

- to study the nature of data contained in EHRs;
- to identify data sets required for building prognostic models of the course of diseases;
- to select methods which are the most suitable for structuring and extracting data from EHRs, taking into account the objective of information search and convenience of its subsequent use;
- to develop and verify the prediction model finding out relationships between data extracted from health records and predicting the development of medical events on this basis.

When completing the tasks and achieving the set objective, the following methods were used:

- information search through sources of scientific technical information and medical information ones on the research topic;
- analysis of the collected material with its subsequent systemization;
- the comparison method which has allowed finding the most suitable ways of solving the problem under study in line with the set objective.

The research was conducted with anonymized medical data of patients from three hospitals in Russia. The authors used data on the condition of patients aged 18 to 70 having cardiovascular diseases; the data was collected by their physicians in charge during the patients' first in-office visit to the health facility. Medical records were made by different attending physicians in a free form and represented non-structured text.

As medical data to be analyzed, the following information about patients was extracted: gender, age, height, weight, arterial blood pressure (systolic and diastolic), respiratory rate, heart rate, Covid-19 symptoms, information on bad habits (smoking, alcohol abuse), the list of previous diagnoses (according to ICD-10), patient complaints, results of examination by the physician, and presence of cardiovascular diseases in parents. Recommendations outlined by doctors and indicated medical therapy were taken into account, too.

Medical records made in EHRs in the Russian language were used. Table 1 shows an example of such entries.

Table 1 Example Medical Records in Russian EHR (Originally in Russian Language)

General information: woman, 48 y.o. **Diagnosis List (ICD-10)**: F20.0, K02.1, I10, M42.1, N95.1, D10.4, D27 **Objective:** Overall condition satisfactory, emotionally labile. The skin is clean, normal color. Peripherals don't work. The throat is pink and clean. Cor tones are rhythmic. Blood pressure 130/80 mm Hg A pulse of 65 beats per minute. In the lungs, respiration is vesicular. Percussion pulmonary sound. The abdomen is soft b/b on palpation. Liver, spleen not taken away. T-36,5., Physiological functions are normal. The pasty stop. Smoking for 15 years.

Source: compiled by the authors

This data was downloaded as a text file from the clinical decision support system (CDSS) Webiomed that containing more than 50 million records (Gavrilov et al., 2020).

4 Results and Discussion

For an artificial intelligence model to start making accurate forecasts, one has to train it promptly. Training requires a large number of data samples mined from EHRs. This circumstance demanded solving a problem associated with studying the nature of data contained in EHRs.

It was found that the principal problem for artificial intelligence to overcome when extracting data from EHRs in the automated mode is the following: medical workers' using their own word abbreviations and brief note forms; spelling errors in the text; inconsistent summary of the results of examination and patient's complaints; building descriptive sentences with violation of the conventional grammar rules; and careless attitude to filling out all fields of the electronic record.

At the next stage of work, the authors selected methods which fitted best for structuring and extracting data from EHRs, taking into account the objective of information search and convenience of their subsequent use.

It has been found that at this stage of working with information, several methods have to be used simultaneously, because none of the existing methods can completely satisfy current needs in handling the data of EHRs which contain both structured information and free text. Detailed information and methods for extracting it are shown in Table 2. The rules-based method, Classification Machine Learning method, and NLP Name entity recognition (NER) are used in the Webiomed system simultaneously.

Table 2 NLP Methods in HER

Categories	Text data	Sections	NLP	Examples of Data
Presenting Problem	Structure or semi-structure (pre-defined sections)	Detailed description of problem(s)	NER Rules-based Hybrid	Up-to-day of important health problems, including diagnoses, symptoms, physical findings and physical test findings
		List of symptoms	Text classifi-cation	«chest pain» identifies portions of clinical note text where 1 of the terms describing PAIN (eg, pressure) either precedes or follows 1 of the terms describing the LOCATION (eg, chest).
		Mental status	NER	
Clinical Notes	Unstructu-red	Progress notes		
		Consultation letters	NER, Rules-based	
		Hospitals additional notes		
		Previous treatment history	NER	
Personal History	Structure or semi-structure (pre-defined sections	Develop-mental milestone	NLP: Diag-noses extrac-tion	
		Medical history		Past medical, surgery, developmental & social history
		Physical, emotional, sexual abuse		
		Diet, exercise		
Substance Abuse History	Unstructu-red	Pattern of use: onset, frequency, quantities		
		Drugs/habits of choice: alcohol, smoking	NLP: Extract informa-tion	ML methods
Family History	Unstructu-red	Age and health of parents, siblings	NLP: data extrac-tion	Medical history of the family members
		Description of relation		Rules-based methods
		Cultural and ethnic influence		
		History of illness, mental illness		
Employme nt and Education	Unstructu-red	Educational history		
		Employment history		
		Achieve-ments, patterns and problem		
Labs Tests	Structure or semi-structure (pre-defined	Interpreta-tions of laboratory, radiology, pathology	NLP NER	Labs test data

	sections)	and other		
Other	Imaging data	Echocardio-logy & electrocar-diology	NLP Annota-tion	
		CT scans		
	Scanned documents	MRI scans		
		Medical documents from external sources		

Source: compiled by the authors

The use of machine learning system based on rules requires special markup of the text. The excerpts singled out during the markup are used by artificial intelligence for finding relationships between them, in particular, between snippets identified in the text of patient complaints, results of tests, examination, and the subsequent confirmed diagnosis. The use of this method requires transformation of each text record symbol into numeric representation in the vector form which is accessible for the computer software complex to process.

We used the method «bag of words», where the vector represents the frequency of a word in a predefined dictionary of words. The machine-learning algorithm then receives training data consisting of pairs of feature sets (vectors for each text example) and tags (diagnosis) to create a classification model. The common question for training data: how much data will be enough? There's still no «Golden Rule». It depends on the type of machine learning problem we are going to solve. In our case, we came to the following rule: 100 sentences of using one feature for training and 30 sentences for validation.

When developing the software for classifying the information listed in the Complaints section of the EHR, we used the following network architecture:

- input layer;
- convolution layer;
- max-pooling layer;
- fully connected hidden layer;
- output layer.

Additional layers will be added between the main layers to protect against network retraining (dropout). The principle of their operation is that the network «forgets» a certain percentage of weights (passed as a parameter in dropout). As activation in the convolution layer and the hidden layer will use the rectified linear activation function «Relu». For the output layer selected activation function «Softmax» is one of the special cases of the sigmoid function applied for multi-class classification. To implement this task the following software was selected:

- Python programming language containing a variety of libraries for dealing with data and neural networks;
- Keras part of the Tensorwlow now enabling CNN, LSTM & biLSTM implementation.

Until now, a technology based on rules and regular expressions was mainly used to extract words from texts. But even now, this method is the most effective for extracting features that can take a small number of values and practically do not change over time.

This method is used in the Webiomed system to extract COVID-19 symptoms from patient's medical records. The examples are given in Table 3.

Table 3 Examples of COVID-19 Symptoms

«The condition is extremely serious. Consciousness is a deep stupor. The situation is forced. The physique is correct. The normothermic Constitution»
«Complaints (according to the mother): weakness, low mood, lack of mobility, decreased appetite: «very thin», stupors state: «sits and looks at one point, does not sleep at night» The patient herself confirms these symptoms and indicates that she has suicidal thoughts»
«The numbness of 2-5 fingers of the left hand is constant, burning in them. Vertigo is not permanent. Anosmia for more than 20 years»
«Pain in the eyes when turning to the side. Noise in the ears. When asking questions, the sense of smell decreases»

Source: compiled by the authors

Classification results can be improved with the use of boosting technology, the essence of which was discussed above.

When extracting data from electronic medical records, we used an open-source natural language processing (NLP) package SpaCy. The advantages of using package SpaCy were discussed earlier.

The spacy pre-train command enables using transfer learning to initialize your models with information from raw data, using a language model objective similar to the one used in Google's BERT system, State-of-the-art techniques (SOTA) in NLP. While training, a model has to memorize examples as well as learn to make assumptions that can be generalized to other new examples on the site «Training SpaCy's Statistical models» at https://spacy.io/usage/training. To test the ability of models to generalize, new test data is needed. To train a model from scratch, hundreds of examples for both training and validation are required. To update an existing model, the results with very few examples can be achieved if they're representative.

Example of data description for training from the program in Python:

TRAIN_DATA = [

("vesicular breathing, no wheezing, RR 18 in min, heart tones clear, rhythmic, BP 130/80 mm Hg heart rate 72 beats per min.",

{'entities': [(34, 46, 'RR'), (79, 97, 'BP'), (98, 113, 'HR')]}),

….]

Example of data description for training from the command line interface (CLI):

"id": 9, "paragraphs": [

{"raw": "BP 120/70 мм Hg",

"sentences": [{

"tokens": [{"id": 0,

"orth": "BP",

"ner": "B-BP"},

{"id": 1,

"orth": "120/70",

"ner": "I-BP"},

{"id": 2,

"orth": "мм",

"ner": "I-BP"},

{"id": 3,

"orth": "рт.ст",

"ner": "L-BP"},

{"id": 4,

"orth": ".",

"ner": "O"}],

"brackets": []}],

"cats": []}]

Text classification is one of the main tasks of NLP and allows solving problems related to the text belonging to a particular class. We applied this method to determine the main diagnosis of an EHR based on labeled medical data. Features in section «Complaints» will be used as the dependent variable, and «Diagnosis» as the target.

Major problems of the traditional method of the text classification:

1. It is necessary to choose a method for converting text to a vector representation since the best classification quality is frequently shown for various tasks various methods.
2. The resulting feature space will have a high dimension, and so it will be highly discharged.
3. Frequently stop words have to be removed from the text to improve the classification quality.
4. To apply this method, it is necessary to use stamping or lemmatization, since the words with different declensions have the same meaning.

While doing pre-processing with the text in feature «complaints» the lemmatization Python library PyMorphy2 is used. There are many algorithms for automatic text classification, which can be grouped into three different types: rule-based, Machine Learning based, and hybrid.

Metrics and Evaluation

To choose a proper method for text classification and other NLP tasks as metrics we applied:

- accuracy: the percentage of texts that were predicted with the correct tag.
- precision: the percentage of examples the classifier obtained from the total number of examples predicted for a given tag.
- recall: the percentage of examples the classifier predicted for a given tag out of the total number of examples it should have predicted for that given tag.
- F1 Score: the harmonic means of precision and recall (Chip, 2019).

An important problem with text processing for feature extraction is a large number of errors and misprints. The electronic medical records are an official medical document signed by a physician; therefore, any corrections and adjustments can't be made. In this case, preprocessing of the input text, correcting errors, misprints and abbreviations have to be done.

SpaCy provides way to increase the accuracy of the new, created from scratch model. The models take a long time to train, so not enough experiments can be run to figure out the best hyper parameters for the model training. The following algorithms for optimization we came up:

- initialize with batch size 1, and compound to a maximum determined by the data size and problem type;
- use Adam solver with a fixed learning rate;
- use averaged parameters;
- use L2 regularization;
- clip gradients by L2 norm to 1;
- on small data sizes, start at a high dropout rate, with linear decay. This was developed experimentally.

To extract various features from the EHRs, an ensemble of 4 models was developed. The results of all models are shown in Table 4, where an indication of the purpose, extracted features, used methods for extraction, data sets for training, the corresponding metrics are specified for each model.

Table 4 Developed NLP Models for Extracting Features from Medical Records

Model name	Purpose of the model	Features	Method ML	Datasets	Metrics
Objec-tive Data	Model for extracting features of objective patient data from medical records	height, weight, respiratory rate, respiratory rate, BP	NER	More than 500 anno-tated texts	Precision=97. 6% Recall= 99.5% F1= 98.6%
Smo-king Extract	Model for extracting the sign of «Smoking» from medical records texts	smoking, non-smoking	CAR, TF-IDF, LSA	More than 500 anno-tated texts	Precision=88. 5% Recall = 96.1% F1 = 92.3%
Labo-ratory data	Model for extracting features from laboratory analysis texts	glucose, cholesterol, creatinine, urea, glycemic profile, hematocrit, triglycerides, HDL, LDL	NER	More than 700 anno-tated texts	Precision=91. 4% Recall=87.2% F1 = 89.2%
Covid-19-Symptoms	COVID-19 symptoms extraction model	body temperature, cough, shortness of breath, chest congestion, myalgia, confusion, headaches, hemoptysis, diarrhea, anosmia, conjunctivi-tis, stupor	NER	More than 1500 anno-tated texts	Precision=83, 6% Recall = 80.6% F1= 82.1 %

Source: compiled by the authors

For example, the model «Objective Data» was trained in more than 500 annotated texts, using the NLP Name entity recognition (NER) method, for 120 epochs, using Stochastic gradient descent (SGD) optimizer and binary cross entropy as the loss function. The loss function results were also weighted according to the proportion of samples with positive and negative samples.

There are two methods available to obtain the models: the program on Python and the command line methods. The program method significantly loses in training time, since it uses an interpreter, while the command line method uses the program on C using the GPU.

The models were applied by the Spacy train command-line method. The «Scorer» and «nlp.evaluate» now report the text classification scores, calculated as the F-score on a positive label for binary exclusive tasks, the macro-averaged F-score for all exclusive features. Most tokens in real-world medical documents are not a part of entity names as normally defined, so the baseline precision, recall, and F1 is extremely high, typically >90%; so, the entity precision, recall, and F1 values are reasonably good.

5 Conclusion

The result of this work, extracting features from EHR, was the creation of an industrial system, which is built into the Clinical Decision Support System (CDSS) Webiomed (https://webiom ed.ai). Features were extracted from over 600,000 electronic health records and more than 45 million features were extracted just for one month. Thanks to this work, it became possible to create datasets to create predictive models on medical data not from available datasets from other countries, but train on data from the population where the model will be used – in Russia. This made it possible to significantly improve the prediction accuracy. On Dataset – 3521 patients with 24 features, a model for predicting hospitalization in 12 months of patients with cardiovascular diseases from 18 to 70 years old was trained using this method.

The results obtained on the test data show good accuracy = 83 % and AUC = 0.89. The accuracy is captured in two metrics: the «False Alarm Rate» (the fraction of false positives out of the negatives) and the «Detection Rate» (the fraction of true positives out of the positives). For a binary classification system,

the evaluation of the performance using these two metrics is typically illustrated with the Receiver Operating Characteristic (ROC) curve, which plots the Detection Rate versus the False Alarm Rate at various threshold settings. Using the proposed method for extracting structured and unstructured data from medical documents in Russian, they will allow creating datasets for machine learning models that will be responsible for medical tasks and, most importantly, allow creating medical data of the local population with more accuracy.

Literature:

1. Amogh, K. T., Tiwari, A. Dhaimodker, V. N., Rebelo, P., Desai, R., Rao, D.: *Healthcare NER models using language model pretraining*. First health search and data mining workshop (HSDM 2020). 2020. doi: doi.org/10.48550/arXiv.1910.11241.
2. Chip, H.: *Evaluation metrics for language modeling.* The Gradient. 2019. Available from: https://thegradient.pub/understanding-evaluation-metrics-for-language-models/.
3. Gavrilov, D., Gusev, A., Korsakov, I., Novitsky, R., Serova, L.: *Feature extraction method from electronic health records in Russia*. Proceedings of the FRUCT'26. 2020. 497-500 pp.
4. Honnibal, M., Johnson, M.: *An improved non-monotonic transition system for dependency parsing*. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015. 1373-1378 pp.
5. Malmasi, S., Ge, W., Hosomura, N., Turchin, A.: *Comparison of natural language processing techniques in analysis of sparse clinical data: insulin decline by patients.* AMIA Joint Summits on Translational Science, 6, 2019. 610-619 pp.
6. Mayr, A, Binder, H, Gefeller, O, Schmid, M.: *The evolution of boosting algorithms. From machine learning to statistical modelling.* Methods of information in medicine, 53(4), 2014. 419-427 pp. doi: doi.org/10.3414/ME13-01-0122.
7. Metlay, J. P., Haas, J. S., Soltoff, A. E., Armstrong, K. A.: *Household transmission of SARS-CoV-2*. JAMA Network open, 4(2), 2021. doi: 10.1001/jamanetworkopen.2021.0304.
8. Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C., Hurdle, J. F.: *Extracting information from textual documents in the electronic health record: a review of recent research.* Yearbook of medical informatics. 2008. 128-144 pp. Available from: https://pubmed.ncbi.nlm.nih.gov/18660887/
9. Pedrera-Jiménez, M., García-Barrio, N., Cruz-Rojo, J., Terriza-Torres, A. I., López-Jiménez, E. A., Calvo-Boyero, F., Jiménez-Cerezo, M. J., Blanco-Martínez, A. J., Roig-Domínguez, G., Cruz-Bermúdez, J. L., Bernal-Sobrino, J. L., Serrano-Balazote, P., Muñoz-Carrero, A.: *Obtaining EHR-derived datasets for COVID-19 research within a short time: a flexible methodology based on detailed clinical models*. Journal of biomedical informatics, 115, 2021. Art No.103697. doi:10.1016/j.jbi.2021.103697

Primary Paper Section: I

Secondary Paper Section: IN, FS