ISSN 2071-5021

СИСТЕМА АВТОМАТИЧЕСКОЙ РАЗМЕТКИ НЕСТРУКТУРИРОВАННЫХ ПРОТОКОЛОВ РЕНТГЕНОЛОГИЧЕСКИХ ИССЛЕДОВАНИЙ ГРУДНОЙ КЛЕТКИ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ СЕМАНТИЧЕСКОГО АНАЛИЗА

07.03.2023 г.
Информатизация здравоохранения

DOI: 10.21045/2071-5021-2023-69-1-12

[1]Ронжин Л. В., [1]Астанин П. А., [2]Кокина Д. Ю., [2]Семенов С. С., [2]Арзамасов К. М., [1]Раузина С. Е.

[1]Федеральное государственное автономное образовательное учреждение высшего образования «Российский национальный исследовательский медицинский университет имени Н. И. Пирогова», г. Москва, Россия

[2]Государственное бюджетное учреждение здравоохранения города Москвы «Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы», г. Москва, Россия

Резюме

В настоящее время не существует единого структурированного стандарта описания рентгенологических исследований органов грудной клетки. Сложность создания такого стандарта заключается в многочисленности методик лучевой диагностики, разнообразии диагностических задач и особенностях работы отдельных медицинских организаций. Разработка инструментов разметки существующих неструктурированных протоколов рентгенологических исследований позволит усовершенствовать систему электронного документооборота в сфере медицины за счет автоматизации процессов формализации данных, а также подготовить наборы данных для машинного обучения.

Целью настоящего исследования является разработка системы автоматической разметки текстовых заключений в протоколах рентгенологических исследований органов грудной клетки на основе экспертных методов и методов машинного обучения.

Материалы и методы. В качестве материала исследования выступают диагностические данные о пациентах, проходивших рентгенологические исследования грудной клетки в подключенных к Единому радиологическому информационному сервису Единой медицинской информационно-аналитической системы амбулаторных и стационарных медицинских организаций Москвы и Московской области. Для обработки неструктурированных текстовых протоколов использованы методы семантического анализа, экспертные правила и алгоритмы машинного обучения.

Результаты. В ходе исследования выявлены языковые паттерны, свойственные классам наиболее важных патологических состояний и классу «норма», а также созданы соответствующие им регулярные выражения. Составлен словарь рентгенологических терминов и сокращений (397 слов), после чего разработан алгоритм коррекции грамматических ошибок в протоколах. Совместно с врачами-рентгенологами экспертной группы сформированы правила для многозначной классификации протоколов рентгенологического исследования и оценена их эффективность. При решении задачи многозначной классификации с использованием только экспертных правил процент точных совпадений составил 84%. В связи с недостаточной эффективностью решателей для таких состояний, как «инфильтрация/консолидация» и «очаг затемнения», проведена настройка моделей машинного обучения.

Заключение. Наилучшие результаты классификации показала рекуррентная нейронная сеть, позволившая достичь значений показателя чувствительности в 89 и 99%, соответственно, для «инфильтрации/консолидации» и «очага затемнения», что позволило статистически значимо (p=0,039) повысить общий процент точных совпадений до 87%.

Ключевые слова: семантический анализ; машинное обучение; неструктурированные данные; анализ текстов; NLP; языковые модели

Контактная информация: Астанин Павел Андреевич, med_cyber@mail.ru

Финансирование. Исследование выполнено в рамках государственного задания «Научные методологии устойчивого развития технологий искусственного интеллекта в медицинской диагностике» и федеральной программы «Приоритет 2030».

Конфликт интересов. Авторы декларируют отсутствие явных и потенциальных конфликтов интересов в связи с публикацией данной статьи.

Соблюдение этических стандартов. Данный вид исследования не требует прохождения экспертизы

локальным этическим комитетом.

Для цитирования: Ронжин Л.В., Астанин П.А., Кокина Д.Ю., Семенов С.С., Арзамасов К.М., Раузина С.Е. Система автоматической разметки неструктурированных протоколов рентгенологических исследований грудной клетки с использованием методов семантического анализа. *Социальные аспекты здоровья населения* [сетевое издание] 2023; 69(1):12. Режим доступа: http://vestnik.mednet.ru/content/view/1455/30/lang,ru/. DOI: 10.21045/2071-5021-2023-69-1-12

SEMANTIC ANALYSIS METHODS IN THE SYSTEM FOR AUTHOMATED MARKING OF THE UNSTRUCTURED RADIOLOGICAL CHEST EXAMINATION PROTOCOLS

[1]Ronzhin LV, [1]Astanin PA, [2]Kokina DYu, [2]Semenov SS, [2]Arzamasov KM, [1]Rauzina SE

[1]Pirogov Russian National Research Medical University, Moscow, Russian Federation

[2]Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department, Moscow, Russian Federation

Abstract

Currently, a unified structured standard for describing radiological chest examination does not exist. The complexity of developing such text report templates lies in the diversity of instrumental methods, variety of diagnostic objectives and specific work characteristics of individual medical organizations. Development of tools for marking the unstructured radiological chest examination protocols makes it possible to improve the system of electronic document management in healthcare due to automation of data formalization processes as well as develop data sets for machine learning.

The purpose of this study is to develop a system for automated marking of text reports of the unstructured radiological chest examination protocols using heuristic approach and machine learning algorithms.

Material and methods. The study used patient data on radiological chest examinations of medical organizations connected to the Unified Radiological Information Service of the Unified Medical Information and Analysis System of inpatient and outpatient medical organizations of Moscow and the Moscow region. Semantic analysis methods, expert rules and machine learning algorithms were used for processing the unstructured text reports.

Results. The study has identified language patterns associated with important pathological conditions and “norm” class as well as developed regular expressions for these classes. A dictionary of radiological concepts and abbreviations (397 items) was compiled, followed by the development of an algorithm for correcting grammar mistakes in the protocols. In collaboration with the expert group, the rules of multilabel classification of the radiological examination protocols were created and their efficiency was tested. When solving the multilabel classification problem using only the expert rules, the percentage of exact matches equaled to 84%. Inasmuch as classifiers for conditions such as “infiltration/consolidation” and “blackout focus” were not effective, we have adjusted the models of machine learning.

Conclusion. The best classification results were demonstrated by the recurrent neural network with the long-short term memory architecture ensuring sensitivity of 89% and 99% for “infiltration/consolidation” and “blackout focus” classes, respectively. This made it possible to statistically significantly (p=0.039) increase the total percentage of the exact matches up to 87%.

Keywords: semantic analysis; machine learning; unstructured data; text analysis; NLP; language models

Corresponding author: Pavel A. Astanin, email: med_cyber@mail.ru

Information about authors:

Ronzhin LV, https://orcid.org/0000-0002-4653-1611

Astanin PA, https://orcid.org/0000-0002-1854-8686

Kokina DYu, https://orcid.org/0000-0002-1141-8395

Semenov SS, https://orcid.org/0000-0003-2585-0864

Arzamasov KM, https://orcid.org/0000-0001-7786-0349

Rauzina SE, https://orcid.org/0000-0002-9535-2847

Financial support. The study was implemented within the framework of the state assignment “Scientific methods for sustainable development of artificial intelligence technologies in medical diagnostics" and the federal program “Priority 2030”.

Competing interests. The authors declare the absence of any conflicts of interest regarding the publication of this paper.

For citation: Ronzhin LV, Astanin PA, Kokina DYu, Semenov SS, Arzamasov KM, Rauzina SE. Semantic analysis methods in the system for automated marking of the unstructured radiological chest examination protocols. *Social'nye aspekty zdorov'a naselenia* [serial online] 2023; 69(1):12. Available from: http://vestnik.mednet.ru/content/view/1455/30/lang,ru/. DOI: 10.21045/2071-5021-2023-69-1-12 (In Rus).

Введение

Внедрение электронного документооборота (ЭДО) является ключевым звеном формирования единого цифрового контура системы здравоохранения [1]. Структурированные электронные медицинские документы (СЭМД) активно используются в работе медицинских информационных систем (МИС), однако большинство из них характеризуются низким уровнем формализации данных [2]. По оценкам П. А. Тучковой [3], доля неструктурированных и неформализованных данных в системе ЭДО медицинских организаций (МО) может составлять более 80%.

Потребность в анализе неструктурированных медицинских данных присутствует во всех клинических областях [4–8], включая различные направления лучевой диагностики [9]. Несмотря на стремительное развитие средств визуализации медицинской информации, за последние годы протоколы лучевых исследований практически не подверглись изменениям не только со структурной, но и с содержательной стороны [10]. Большое разнообразие стилей протоколов, используемых специалистами лучевой диагностики, свидетельствует об отсутствии уникального и единого формата описания результатов исследований [11].

Важно отметить, что использование однозначной и согласованной терминологии служит основным правилом представления диагностических результатов исследований и должно изначально предусматриваться разработчиками шаблонов при их создании [12–14]. Основная сложность структурирования и формализации информации в данной области заключается в разнообразии анатомических структур, функциональных особенностей, методик исследования и особенностей работы отдельных медицинских организаций. Очевидно, что полноценное удовлетворение потребности системы здравоохранения в СЭМД потребует значительных трудозатрат и времени [15].

В настоящий момент для обработки неформализованных документов могут применяться алгоритмы семантического анализа, реализованные в виде комбинации экспертных решений и машинных методов [16]. Семантический анализ, в ходе которого производится обработка текстовых данных, считается одной из наиболее сложных и неизученных областей интеллектуального анализа данных [17]. С одной стороны, анализ медицинских текстов тесно связан с человеческим фактором, из чего следует наличие большого количества уникальных для предметной области терминов, опечаток, ошибок, аббревиатур и жаргонизмов [18]. С другой стороны, каждый язык имеет уникальную семантическую специфику [19], из-за которой опыт зарубежных коллег не может быть полноценно адаптирован и использован для решения задач анализа неструктурированной информации на иных языках [20].

Одной из задач семантического анализа медицинских текстов является создание инструментов для решения задач классификации с использованием математических алгоритмов и логических правил, обеспечивающих в дальнейшем поддержку принятия врачебных решений [21]. Классификация медицинских текстов предполагает их соотнесение с классом заболевания, с группой риска, с тяжестью состояния и иными категориями.

Разметка текстовых протоколов рентгенологических заключений является необходимым этапом формирования набора данных для обучения интеллектуальных систем компьютерного зрения [22]. В свою очередь, актуальность разработки систем автоматической разметки медицинских текстов обусловлена потребностью в трудоемкой ручной разметке с привлечением клинических экспертов. Появление таких систем обеспечит ускорение процесса подготовки данных и обучения классификаторов изображений, а также позволит снизить количество ошибок, связанных с человеческим фактором, при диагностике наиболее значимых классов заболеваний [23].

Целью настоящего исследования является разработка системы автоматической разметки текстовых заключений в протоколах рентгенологических исследований органов грудной клетки на основе экспертных методов и методов машинного обучения.

Материал и методы

Исследование проводилось с декабря 2021 г. по июнь 2022 г. на базе Кафедры медицинской кибернетики и информатики ФГАОУ ВО «Российский национальный исследовательский медицинский университет имени Н. И. Пирогова» (РНИМУ им. Н. И. Пирогова) и Отдела медицинской информатики, радиомики и радиогеномики ГБУЗ «Научно-практический клинический центр диагностики и телемедицинских технологий» Департамента здравоохранения г. Москвы (ГБУЗ «НПКЦ ДиТ ДЗМ»).

Объектом настоящего исследования являются диагностические данные о пациентах, проходивших рентгенологические исследования грудной клетки в подключенных к Единому радиологическому информационному сервису (ЕРИС) Единой медицинской информационно-аналитической системы (ЕМИАС) амбулаторных и стационарных МО Москвы и Московской области. Предметом исследования являются неструктурированные текстовые протоколы (с описательной частью и заключением от врачей-рентгенологов) рентгенологических исследований органов грудной клетки, распределенных на наиболее значимые с клинической точки зрения нозологические группы.

Обучающая выборка получена из базы данных единой радиологической информационной системы (ЕРИС) и включает 4983 протокола рентгенологических исследований грудной клетки. В обучающей выборке 4384 (88,0%) образцам соответствует класс «норма», 42 (0,8%) – «плевральный выпот», 3 (0,1%) – «пневмоторакс», 108 (2,2%) – «очаг затемнения», 292 (5,8%) – «инфильтрация/консолидация», 4 (0,1%) – «диссеминация», 8 (0,2%) – «полость». Остальным 142 (2,8%) экземплярам соответствуют редко встречающиеся патологические изменения, объединенные в класс «другое». К таким изменениям относятся «кальцинаты», «ателектаз», «консолидированный перелом», «нарушение целостности кортикального слоя», «расширение тени средостения» и «кардиомегалия».

Тестовая выборка включает 507 протоколов со следующим распределением классов: «норма» – 191 (37,7%) образец, «плевральный выпот» – 64 (12,6%), «пневмоторакс» – 26 (5,1%), «очаг затемнения» – 29 (5,7%), «инфильтрация/консолидация» – 85 (16,8%), «диссеминация» – 5 (1,0%), «полость» – 15 (3,0%), «другое» – 92 экземпляра (18,1%).

Помимо этого, дополнительная выборка из 5000 протоколов без классовой разметки использовалась для вычисления векторных эмбеддингов (от англ. embedding – вложение) – тензорного отображения контекстной сочетаемости слов.

Настоящее исследование включало в себя два основных блока: препроцессинг (подготовка) данных и их последующий анализ. На первом этапе препроцессинга осуществлялись стандартные процедуры обработки данных: удаление некачественных образцов и дубликатов с помощью регулярных выражений, а также устранение грамматических ошибок [24]. Далее вводились ограничения на минимальное количество экземпляров в классе. В рамках балансировки классов производилось использование алгоритмов субдискретизации – искусственного уменьшения размера подвыборок мажоритарных (превосходящих по объему) классов, передискретизации – искусственного увеличения размера подвыборок миноритарных (малочисленных) классов, а также аугментации – генерации новых текстов на основе исходных [25]. Важно отметить, что, ввиду отсутствия в свободном доступе словаря синонимов медицинских терминов, было принято решение использовать векторные эмбеддинги для замены некоторых слов.

Последующие итерации препроцессинга данных включали работу непосредственно с текстом: перевод символов в нижний регистр, удаление знаков препинания, синонимов, «стоп-слов» – наиболее распространённых слов с низкой семантической ценностью (частиц, предлогов, союзов) [26]. В целях снижения размерности признакового пространства осуществлена лемматизация – приведение слов к единой нормальной форме с использованием библиотеки PyMorphy2. Данный морфологический анализатор позволяет производить быстрый поиск по словарю и получать список нормальных форм, если их может быть несколько. В случае, если слово отсутствует, алгоритмы библиотеки делают предположение о нормальной форме слова, основываясь на морфемном разборе. Итоговым результатом применения перечисленных этапов препроцессинга является преобразование текста в последовательность лаконичных синтаксических конструкций – токенов.

После предобработки текстов наступал этап извлечения признаков и представления текста в числовом виде. В данном исследовании был использован подход, основанный на извлечении ограниченного количества слов с наибольшим значением TF-IDF (term frequency–inverse document frequency) меры. TF-IDF мера позволяет ранжировать слова по их семантической ценности в пределах всего корпуса (набора) документов и отбирать наиболее значимые слова из всего словаря. Благодаря применению TF-IDF были сформированы векторные представления текстов, которые в дальнейшем подавались на вход алгоритмам классификации.

В случае использования TF-IDF представления текста целесообразно использование моделей машинного обучения, игнорирующих порядок следования токенов. К таким методам относятся Logistic Regression (логистическая регрессия), Support Vector Machine (метод опорных векторов), Random Forest (ансамбль деревьев решений), Extreme Gradient Boosting (градиентный бустинг), k-Nearest Neighbors (метод k-ближайших соседей).

Разработка алгоритмов разметки неструктурированных текстов протоколов рентгенологических исследований осуществлялась на основе двух подходов: с использованием решающих правил и с использованием методов машинного обучения. Создание классификационных моделей на основе решающих правил осуществлялось экспертным путем с использованием регулярных выражений. Разработка моделей на основе машинного обучения включала использование следующих алгоритмов: ансамбль деревьев решений, логистическая регрессия, полносвязная нейронная сеть прямого распространения с тремя скрытыми слоями, нейронная сеть с архитектурой LSTM.

Оценка качества каждой модели основана на вычислении стандартных метрик бинарной классификации: точности, чувствительности, специфичности и F-меры для каждого класса. Для интегральной оценки качества работы алгоритма в задаче многозначной классификации, подразумевающей возможность наличия у пациента сразу нескольких патологических состояний, использованы метрики точности и отношения точных совпадений – доли экземпляров, для которых алгоритм не допустил ошибки ни в одном из классов. Оценка доверительных интервалов (ДИ) осуществлялась методом Уилсона с поправкой на непрерывность. В отличие от симметричного нормального интервала аппроксимации, интервал оценки Уилсона является асимметричным и не страдает от проблем преодоления границ и интервалов нулевой ширины, которые затрагивают нормальный интервал. Данный метод оценки ДИ можно безопасно использовать для малых выборок, несбалансированных классов и искаженных наблюдений.

Для технической реализации всех этапов настоящего исследования применялись средства языка программирования Python 3.9 и среда разработки Google Colaboratory. Для обучения и тестирования моделей машинного обучения использовались библиотеки Scikit-learn, Keras, TensorFlow, PyTorch, а также гибридный аппаратный ускоритель на виртуальной машине, выделенной Google Colaboratory.

Результаты

В ходе очистки и предобработки полученные данные были очищены от пропусков (их оказалось всего 2), после чего произведено удаление незначимых лексем (адресов, имён, цифр, знаков препинания, обозначений доз облучения, всех слов с корнем «рентген» и вариантами его сокращений). Все буквы приведены к нижнему регистру, все пробельные символы (множественные пробелы, переносы строки, переносы каретки) заменены на одиночный пробел. Буквы «ё» заменены на «е».

Изначально протоколы рентгенологических исследований органов грудной клетки содержали большое количество орфографических ошибок, что было следствием их рукописного ввода. Для правильной работы распознающих алгоритмов документы подвергнуты коррекции с использованием редакционного расстояния Левенштайна – наиболее распространённого метода оценки лексической схожести текстов [27]. В качестве эталонного словаря для данного алгоритма выбран словарь Open Office, встроенный во многие текстовые редакторы. Однако при его использовании большинство медицинских терминов, сокращений, профессиональных жаргонизмов и аббревиатур исправлялось неправильно. Поэтому было принято решение дополнить словарь вручную путем включения в него слов, нераспознанных алгоритмом. Из 4983 протоколов было извлечено 479 слов, характерных для рентгенологических протоколов и отсутствующих в обычном словаре русского языка. После пополнения словаря алгоритм позволил корректно исправить ошибки правописания и получить готовые к дальнейшему использованию тексты.

Важно отметить, что для обучающей выборки был характерен выраженный дисбаланс классов (даже на уровне «норма-патология»). Поскольку большинство заключений о норме являются однотипными, реализована субдискретизация класса «норма» посредством удаления дубликатов, что позволило снизить количество экземпляров до 862. Для обучения моделей, игнорирующих порядок слов в тексте, классы были уравнены алгоритмом SMOTE, который предполагает генерацию новых объектов с использованием данных о существующих образцах миноритарного класса [28]. Для обучения моделей, учитывающих порядок входных токенов, классы были аугментированы при помощи векторных эмбеддингов. Идея данного подхода состояла в вычислении косинусной меры схожести для некоторых слов в тексте и их последующей замене на ближайших соседей. Аугментация позволила увеличить объем подвыборок некоторых классов почти в 2 раза: класс «очаг затемнения» – с 108 до 206, а класс «инфильтрация/консолидация» – с 292 до 480 экземпляров (таблица 1).

Таблица 1

Пример аугментации токенизированного текста с использованием векторного эмбеддинга

Пример для аугментации	Аугментированный текст
«признак левосторонний инфильтрация утолщение междолевой плевра слева»	«признак правосторонний пневмония утолщение междолевой плевра справа»

На первом этапе разработки алгоритма классификации неструктурированных текстов рентгенологических протоколов были созданы логические решатели, основанные на использовании экспертных правил. Правила разрабатывались совместно с врачами-рентгенологами экспертной группы ГБУЗ «НПКЦ ДиТ ДЗМ» для каждого класса. Всего было создано 58 правил, описанных с использованием 86 регулярных выражений: для «нормы» – 22, для «плеврального выпота» – 4, для «пневмоторакса» – 2, для «очага затемнения» – 21, для «инфильтрации» – 19, для «диссеминации» – 3, для «полости» – 7, для класса «другое» – 8. На рисунке 1 представлен фрагмент решателя для выявления класса «Очаг затемнения».

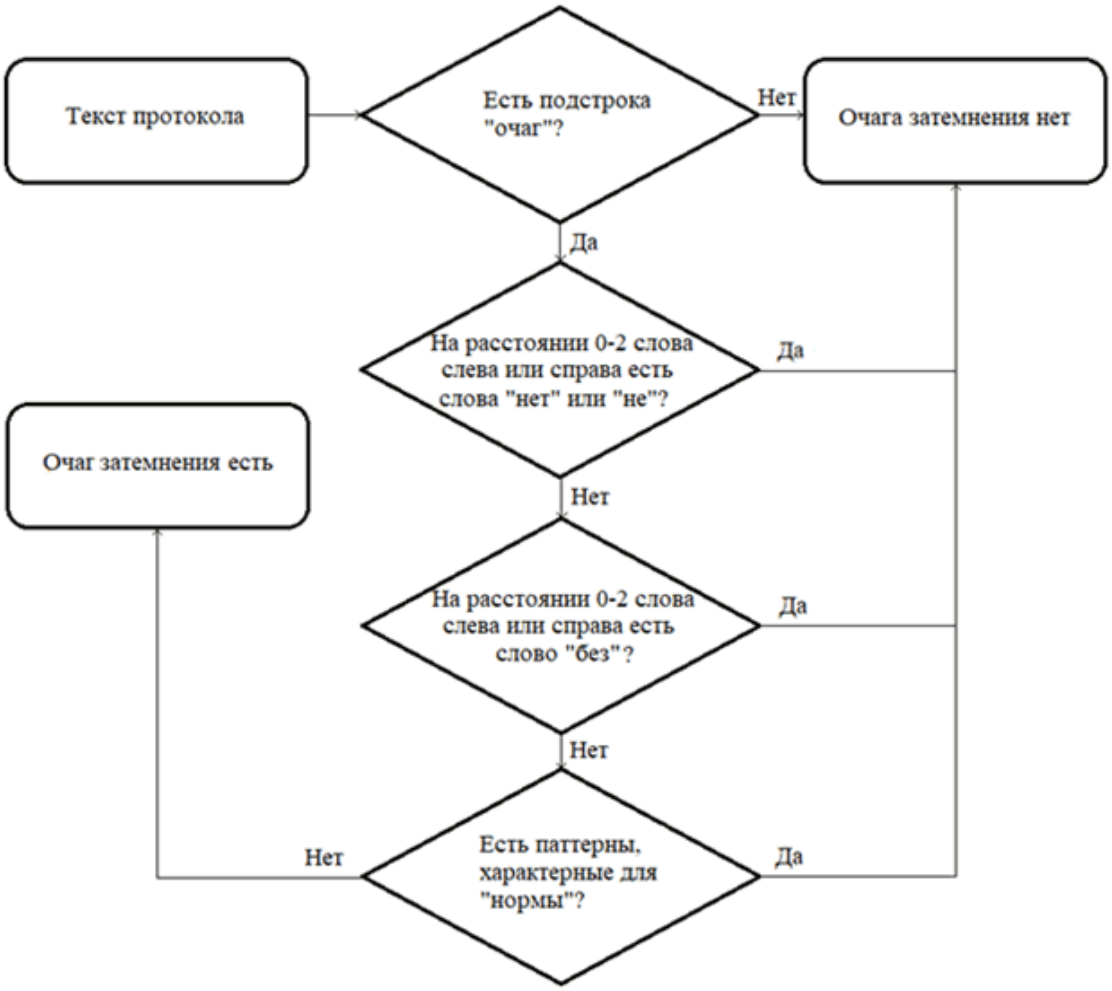

Рис. 1. Фрагмент решателя для класса «Очаг затемнения»

Для каждого состояния была произведена оценка качества бинарной классификации, результаты которой представлены в таблице 2.

Таблица 2

Оценка качества моделей, основанных на правилах

Описываемое состояние	Метрика качества классификации [ДИ$_{95\%}$], %				
	1	2	3	4	5
Плевральный выпот (n=64)	96 [94; 97]	80 [70; 88]	99 [98; 100]	95 [87; 98]	96 [94; 98]
Пневмоторакс (n=26)	100 [99; 100]	100 [87; 100]	100 [99; 100]	100 [87; 100]	100 [99; 100]
Очаг затемнения (n=29)	97 [95; 98]	71 [54; 83]	99 [97; 99]	83 [65; 92]	98 [96; 99]
Инфильтрация (n=85)	94 [92; 96]	84 [75; 90]	97 [94; 98]	86 [77; 92]	96 [94; 98]
Диссеминация (n=5)	100 [99; 100]	100 [56; 100]	100 [99; 100]	100 [56; 100]	100 [99; 100]
Полость (n=15)	100 [98; 100]	100 [77; 100]	100 [98; 100]	87 [62; 96]	100 [99; 100]
Другое (n=92)	97 [95; 98]	93 [86; 97]	98 [96; 99]	93 [86; 97]	98 [96; 99]
Норма (n=191)	91 [88; 93]	93 [88; 96]	90 [86; 93]	85 [79; 89]	95 [92; 97]
Примечание: 1 – точность, 2 – чувствительность, 3 – специфичность, 4 – прогностическая ценность положительного результата (ПЦПР), 5 – прогностическая ценность отрицательного результата (ПЦОР)					

Из данных таблицы 2 следует, что наилучшие результаты классификации показали модели, предназначенные для выявления пневмоторакса, диссеминации и полости. Относительно наихудшие результаты показали модели бинарной классификации для выявления инфильтрации и очага затемнения. При оценке качества

многозначной классификации (для всех моделей в целом) доля точных совпадений (MER) составила 84 [81; 87] %.

В целях улучшения качества работы алгоритма классификации неструктурированных текстов протоколов рентгенологических исследований грудной клетки на втором этапе было принято решение применить методы машинного обучения для классов «очаг затемнения» и «инфильтрация/консолидация». Результаты оценки качества классификации для моделей бинарной классификации данных состояний представлены в таблицах 3 и 4, соответственно.

Таблица 3

Сравнительная оценка качества бинарной классификации для класса «очаг затемнения» с использованием различных алгоритмов машинного обучения

Алгоритм классификации	Метрика качества классификации [$ДИ_{95\%}$], %				
	1	2	3	4	5
Логистическая регрессия	94 [92; 96]	74 [56; 85]	96 [93; 97]	57 [43; 72]	98 [96; 99]
Случайный лес	96 [94; 98]	74 [56; 85]	98 [96; 99]	76 [59; 87]	98 [96; 99]
Полносвязная нейросеть	94 [92; 96]	76 [60; 88]	95 [93; 97]	57 [43; 71]	98 [96; 99]
Нейронная сеть LSTM	98 [96; 99]	85 [70; 94]	99 [97; 99]	85 [70; 94]	99 [97; 99]
Примечание: 1 – точность, 2 – чувствительность, 3 – специфичность, 4 – прогностическая ценность положительного результата (ПЦПР), 5 – прогностическая ценность отрицательного результата (ПЦОР)					

Данные, продемонстрированные в таблице 3, позволяют сделать вывод о том, что наиболее высокое качество классификации для состояния «очаг затемнения» продемонстрировала нейронная сеть LSTM. Чувствительность, являющаяся наиболее информативной метрикой при большом числе классов, для данной модели составила 85 [70; 94] % на тестовой выборке против 71 [54; 83] % для модели, основанной на правилах.

Таблица 4

Сравнительная оценка качества бинарной классификации для класса «инфильтрация/консолидация» с использованием различных алгоритмов машинного обучения

Алгоритм классификации	Метрика качества классификации [$ДИ_{95\%}$], %				
	1	2	3	4	5
Логистическая регрессия	84 [80; 87]	83 [73; 89]	84 [80; 87]	55 [46; 63]	95 [93; 97]
Случайный лес	88 [84; 90]	78 [68; 86]	90 [86; 93]	65 [55; 73]	95 [92; 96]
Полносвязная нейросеть	86 [82; 89]	77 [67; 85]	88 [84; 91]	60 [51; 68]	94 [91; 96]
Нейронная сеть LSTM	97 [95; 98]	89 [80; 94]	99 [97; 99]	94 [87; 97]	97 [95; 99]
Примечание: 1 – точность, 2 – чувствительность, 3 – специфичность, 4 – прогностическая ценность положительного результата (ПЦПР), 5 – прогностическая ценность отрицательного результата (ПЦОР)					

При построении моделей для обнаружения класса «инфильтрация/консолидация» наилучшие результаты вновь показала LSTM. Чувствительность модели на тестовой выборке составила 89 [80; 94] % против 84 [75; 90] % для модели, основанной на правилах.

На основании полученных оценок качества классификаторов принято решение использовать модели LSTM для выявления классов «инфильтрация/консолидация» и «очаг затемнения», а остальные классы выделять моделями, основанными на правилах. Схема работы итогового алгоритма представлена на рисунке 2 и включает следующие шаги: очистка текстов, коррекция грамматических ошибок, лемматизация, вычисление векторных эмбеддингов, выделение классов «плевральный выпот», «пневмоторакс», «полость», «диссеминация», «другое» моделью, основанной на экспертных правилах, выделение классов «очаг затемнения» и «инфильтрация» нейросетью LSTM. Если патологий нет, протоколу присваивается значение метки нормы, равное «1».

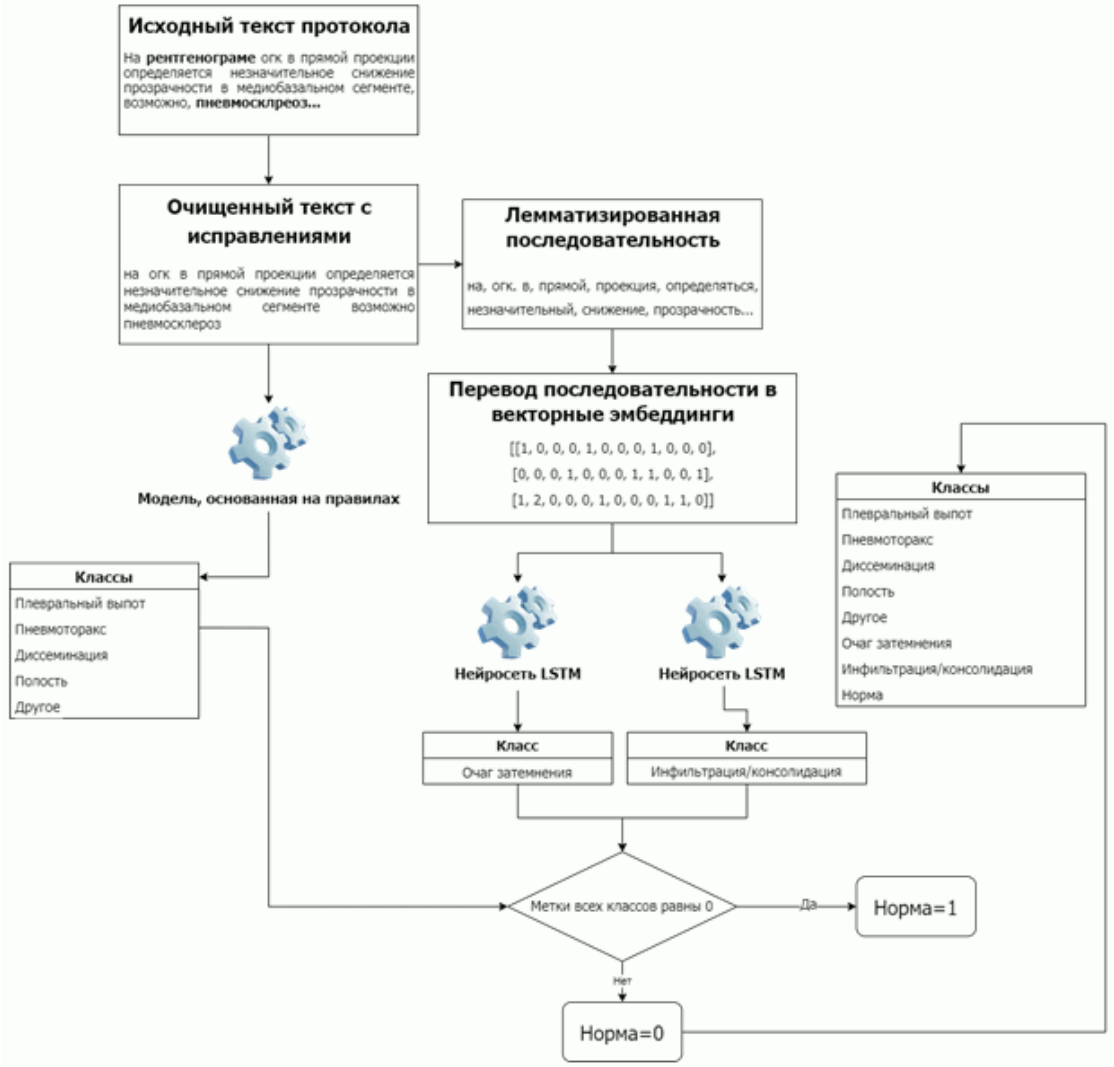

Рис. 2. Схема итогового алгоритма автоматической разметки протоколов рентгенологических исследований органов грудной клетки

В ходе оценки работы разработанной системы автоматической разметки текстов в задаче многозначной классификации процент точных совпадений составил 87 [84; 90] % против 84 [81; 87] % для аналогичного алгоритма, основанного только на правилах. Прирост точности алгоритма оказался статистически значимым (p=0,039).

Обсуждение

Проведённое исследование продемонстрировало основные этапы работ по созданию алгоритма классификации неструктурированных текстов протоколов рентгенологических исследований грудной клетки. В процессе исследования применялись не только экспертные решения, но и модели машинного обучения. Были выявлены языковые паттерны, свойственные представителям наиболее важных состояний, и созданы регулярные выражения, соответствующие им. В процессе предобработки текстов составлен словарь рентгенологических терминов и сокращений (397 слов), а также разработан алгоритм коррекции грамматических ошибок в протоколах. Совместно с экспертами сформированы правила для многозначной классификации протоколов рентгенологического исследования и оценена их эффективность. Для большинства состояний качество бинарной классификации оказалось высоким, а процент точных совпадений по всем моделям в совокупности составил 84 [81; 87] %

В связи с недостаточной эффективностью решателей для таких состояний, как «инфильтрация/ консолидация» и «очаг затемнения», проведена настройка следующих моделей машинного обучения: логистическая регрессия, ансамбль деревьев решений (Random Forest), полносвязная нейронная сеть прямого распространения и рекуррентная нейронная сеть (LSTM). Наилучшие результаты классификации показала LSTM, позволившая достичь значений показателя чувствительности в 89 [80; 94] % и 99 [97; 99] %, соответственно, для «инфильтрации/консолидации» и «очага затемнения», что позволило статистически значимо повысить общий процент точных совпадений до 87 [84; 90] %.

Важно отметить, что первоначальный выбор экспертных правил был связан с недостаточным объемом обучающей выборки и выраженным дисбалансом классов. Использование моделей машинного обучения при перечисленных ранее несовершенствах обучающего набора данных позволило улучшить работу итогового алгоритма за счет повышения качества классификации двух патологических состояний. Учитывая, что использование данных алгоритмов позволило значительно упростить и автоматизировать процесс разработки классификаторов, их преимущество не вызывает сомнений. Тем не менее поиск эффективных способов улучшения качества анализа неструктурированных медицинских текстов остается актуальным и требует дальнейшего изучения.

Заключение

Разработанный аналитический алгоритм реализован в виде программного кода и преобразован в сервис с использованием специальной платформы Postman API. В ходе его работы реализуются процессы

предобработки текстов и воспроизводятся алгоритмы классификации, основанные на построенных ранее экспертных правилах и машинных методах. На различных этапах своей работы указанный сервис обращается к словарю рентгенологических терминов, векторным эмбеддингам и сохраненным весам нейронных сетей LSTM. Все вышеперечисленные ресурсы образуют систему автоматической разметки текстов рентгенологических исследований грудной клетки. В настоящее время данный программный продукт находится на этапе внедрения в работу Отдела медицинской информатики, радиомики и радиогеномики ГБУЗ «НПКЦ ДиТ ДЗМ».

К перспективам дальнейших исследований в данной области следует отнести создание инструментов анализа неструктурированных данных на основе языковых моделей. Существующие языковые модели (BERT) используются в англоязычных странах для работы с медицинскими документами и показывают результаты, превосходящие по эффективности экспертные правила и современные архитектуры нейронных сетей [29]. Создание языковых моделей на русском языке станет большим шагом в развитии медицинской информатики в России [30] и позволит повысить качество анализа неструктурированных медицинских данных.

Библиография

1. Калугина Е.А. Система электронного документооборота, ее преимущества и переход на электронный документооборот. *Вестник Национального Института Бизнеса* 2019; (37): 110–113.
2. Чолоян С. Б., Екимов А. К., Байгазина Е. Н., Молодцов Н. С., Калинина Е. А., Поснов А. А. Современные подходы к решению задач управления медицинских организаций. *Менеджер здравоохранения* 2021; (10): 4–13. DOI: 10.21045/1811-0185-2021-10-4-13
3. Тучкова П. А. Применение методов обработки естественного языка для анализа текстовых и речевых данных в медицине. *Наукосфера* 2021; (5-1): 174–179. DOI: 10.5281/zenodo.4771893
4. Шулаев А. В., Галяутдинов Г. С., Бирюков Д. М., Марапов Д. И., Гарипов Р. З., Горнаева Л. И., и др. Формализация медицинских данных пациентов с артериальной гипертензией. *Уральский медицинский журнал* 2020; (8): 21–26. DOI: 10.25694/URMJ.2020.08.07
5. Первышин Н. А., Лебедева И. В., Лебедева Е. А. Формализация и информатизация амбулаторного приема пациентов с сахарным диабетом. *Профилактическая медицина* 2021; 24 (3): 14–21. DOI: 10.17116/profmed20212403114
6. Москалев И. В., Кротова О. С., Хворова Л. А. Автоматизация процесса извлечения структурированных данных из неструктурированных медицинских выписок с применением технологий интеллектуального анализа. *Высокопроизводительные вычислительные системы и технологии* 2020; 4 (1): 163–167.
7. Кротова О. С., Москалев И. В., Хворова Л. А., Назаркина О. М. Реализация эффективных моделей классификации медицинских данных методами интеллектуального анализа текстовой информации. *Известия Алтайского государственного университета* 2020; (1): 99–104. DOI: 10.14258/izvasu(2020)1-16
8. Зулкарнеев Р. Х., Юсупова Н. И., Сметанина О. Н., Гаянова М. М., Вульфин А. М. Методы и модели извлечения знаний из медицинских документов. *Информатика и автоматизация* 2022; 21 (6): 1169–1210. DOI: 10.15622/ia.21.6.4.
9. Андрианова М. Г., Кудрявцев Н. Д., Петряйкин А. В. Разработка тезауруса рентгенологических терминов для голосового заполнения протоколов диагностических исследований. *Digital Diagnostics* 2022; 3 (S1): 21–22. DOI: 10.17816/DD105703
10. Морозов С. П., Владзимирский А. В., Шулькин И. М., Ледихова Н. В., Арзамасов К. М., Андрейченко А. Е., и др. Целесообразность применения технологий искусственного интеллекта в лучевой диагностике (результаты первого года Московского эксперимента по компьютерному зрению). *Врач и информационные технологии* 2022; (1): 12–29. DOI: 10.25881/18110193_2022_1_12
11. Гусев А. В., Владзимирский А. В., Голубев Н. А., Зарубина Т. В. Информатизация здравоохранения Российской Федерации: история и результаты развития. *Национальное здравоохранение* 2021; 2 (3): 5–17. DOI: 10.47093/2713-069X.2021.2.3.5-17
12. Масловская Л. Ю. Особенности медицинской терминологии и пути её пополнения. *The Scientific Heritage* 2021; (63): 41–43. DOI: 10.24412/9215-0365-2021-63-5-41-43
13. Гаппарова Д. А., Юсупова С. Х., Искандаров Д. Ф. Проблемы лексикографического описания медицинских теминов. *Open innovation* 2019: 142–144.
14. Абаев Ю. К. Хороший доктор. Часть 9. Термины и «терминотворчество» в медицине. *Здравоохранение (Минск)* 2020; (878): 28–37.
15. Зарубина Т. В. Единая государственная информационная система – основа цифровизации здравоохранения. *Информационные технологии в медицине и здравоохранении* 2020: 22–35.
16. Юсупова Н. И., Гаянова М. М., Богданов М. Р. Извлечение информации об использовании информационных технологий для поддержки принятия решений в медицинской диагностике. *Вестник Южно-Уральского государственного университета. Серия: Компьютерные технологии, управление, радиоэлектроника* 2022; 22 (1): 14–27. DOI: 10.14529/ctcr220102
17. Алпатов А. Н., Попов К. С., Чесалин А. Н. Анализ точности моделей машинного обучения с использованием методов векторизации для задач классификации разнородных текстовых данных. *International Journal of Open Information Technologies* 2022: 10 (7): 47–53.
18. Harrison CJ, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction to natural language processing. *BMC Medical Research Methodology* 2021; 21: 158. DOI: 10.1186/s12874-021-01347-1
19. Зацман И. М. Проблемно-ориентированная актуализация словарных статей двуязычных словарей и медицинской терминологии: сопоставительный анализ. *Информатика и ее применения* 2021; 15 (1): 94–101. DOI: 10.14357/19922264210113
20. Guo Y, Li C, Roan C, Pakhomov S, Cohen T. Crossing the «Cookie Theft» Corpus Chasm: Applying What BERT Learns From Outside Data to the ADReSS Challenge Dementia Detection Task. *Frontiers in Computer Science* 2021; 3: 642517. DOI: 10.3389/fcomp.2021.642517
21. Землянский С. А., Аксенов С. В., Лызин И. А., Берестнева О. Г. Тематическое моделирование в контексте медицинских текстов. *Доклады Томского государственного университета систем управления и радиоэлектроники* 2021; 24 (4): 58–64. DOI: 10.21293/1818-0442-2021-24-4-58-64
22. Мингазов Д. Р. Обзор применения медицинских экспертных систем. *Вопросы устойчивого развития общества* 2022; (4): 1502–1505.

23. Семенов Н. А., Бурдо Г. Б., Лебедев С. Н., Лебедева Ю. В. Интеллектуальная поддержка принятия решений в экспертных системах при диагностике заболеваний полости рта. *Программные продукты и системы* 2021; (3): 484–488. DOI: 10.15827/0236-235X.135.484-488
24. Chollet F. Deep Learning with Python, Second Edition. New York: Manning Publications; 2021. 504 p.
25. Mosolova AV., Fomin VV., Bondarenko IYu. Text augmentation for neural networks. *CEUR Workshop Proceedings* 2018; (2268): p. 104–109.
26. Cui M, Bai R, Lu Zh, Li X, Aickelin U, Ge P. Regular Expression Based Medical Text Classification Using Constructive Heuristic Approach. *IEEE Access* 2019; 7: p. 147892–147904. DOI: 10.1109/access.2019.2946622
27. Астанин П. А., Ронжин Л. В., Раузина С. Е., Зарубина Т. В. Алгоритмы семантического анализа данных и возможности их применения в разработке медицинских информационных систем. Цифровая статистика. Новые задачи и траектория движения: Материалы IV Съезда медицинских статистиков Москвы. 2022. С. 6–9.
28. Chawla N, Bowyer K, Hall L, Kegelmeyer W. SMOTE: Synthetic Minority Over-sampling Technique. *ArXiv* 2002; 16: p. 321–357. DOI: 10.1613/jair.953
29. Delvin J, Chang M, Lee K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv* 2019: n. pag.
30. Гусев А. В., Владзимирский А. В., Шарова Д. Е., Арзамасов К. М., Храмов А. Е. Развитие исследований и разработок в сфере технологий искусственного интеллекта для здравоохранения в Российской Федерации: итоги 2021 года. *Digital Diagnostics* 2022. - Т. 3. - №3. - С. 178-194. doi: 10.17816/DD107367

References

1. Kalugina EA. Sistema jelektronnogo dokumentooborota, ee preimushhestva i perehod na jelektronnyj dokumentooborot [Electronic document management system, its advantages and changeover to the electronic document management system]. *Vestnik Nacional'nogo Instituta Biznesa* 2019; (37): 110–113. (In Rus.).
2. Choloyan SB, Ekimov AK, Baigazina EN, Molodtsov NS, Kalinina EA, Posnov AA. Sovremennye podhody k resheniju zadach upravlenija medicinskih organizacij [Modern approachesto managing a medical organization]. *Menedzher zdravoohranenija* 2021; (10): 4–13. DOI: 10.21045/1811-0185-2021-10-4-13. (In Rus.).
3. Tuchkova PA. Primenenie metodov obrabotki estestvennogo jazyka dlja analiza tekstovyh i rechevyh dannyh v medicine [Application of natural language processing methods for analysing of text and speech data in medicine]. *Naukosfera* 2021; (5-1): 174–179. DOI: 10.5281/zenodo.4771893. (In Rus.).
4. Shulayev AV, Galiautdinov GS, Biryukov DM, Marapov DI, Garipov RZ, Gornaeva LI, et al. Formalizacija medicinskih dannyh pacientov s arterial'noj gipertenziej [Formalization of medical data in patients with arterial hypertension]. *Ural'skij medicinskij zhurnal* 2020; (8): 21–26. DOI: 10.25694/URMJ.2020.08.07. (In Rus.).
5. Pervyshin NA, Lebedeva IV, Lebedeva EA. Formalizacija i informatizacija ambulatornogo priema pacientov s saharnym diabetom [Outpatient care formalization and informatization for patients with diabetes mellitus]. *Profilakticheskaja medicina* 2021; 24 (3): 14–21. DOI: 10.17116/profmed20212403114. (In Rus.).
6. Moskalev IV, Krotova OS, Khvorova LA. Avtomatizacija processa izvlechenija strukturirovannyh dannyh iz nestrukturirovannyh medicinskih vypisok s primeneniem tehnologij intellektual'nogo analiza [Automation of the process of extraction of structured data from unstructured medical statements using intellectual text analysis technologies]. Vyso*koproizvoditel'nye vychislitel'nye sistemy i tehnologii* 2020; 4 (1): 163–167. (In Rus.).
7. Krotova OS, Moskalev IV, Khvorova LA, Nazarkina OM. Realizacija jeffektivnyh modelej klassifikacii medicinskih dannyh metodami intellektual'nogo analiza tekstovoj informacii [Implementation of effective models for classifying medical data using text mining]. *Izvestija Altajskogo gosudarstvennogo universiteta* 2020; (1): 99–104. DOI: 10.14258/izvasu(2020)1-16. (In Rus.).
8. Zulkarneev RH., Yusupova NI, Smetanina ON, Gayanova MM, Vulfin AM. Metody i modeli izvlechenija znanij iz medicinskih dokumentov [Методы и модели извлечения знаний из медицинских документов]. *Informatika i avtomatizacija* 2022; 21 (6): 1169–1210. DOI: 10.15622/ia.21.6.4. (In Rus.).
9. Andrianova MG, Kudryavtsev ND, Petraikin AV. Razrabotka tezaurusa rentgenologicheskih terminov dlja golosovogo zapolnenija protokolov diagnosticheskih issledovanij [Thesaurus of radiology terms for preparing reports using speech recognition technology]. *Digital Diagnostics* 2022; 3 (S1): 21–22. DOI: 10.17816/DD105703. (In Rus.).
10. Morozov SP, Vladzymyrskyy AV, Shulkin IM, Ledikhova NV, Arzamasov KM, Andreychenko AE, et al. Celesoobraznost' primenenija tehnologij iskusstvennogo intellekta v luchevoj diagnostike (rezul'taty pervogo goda Moskovskogo jeksperimenta po komp'juternomu zreniju) [Feasibility of using artificial intelligence in radiology (first year of Moscow experiment on computer vision)]. *Vrach i informacionnye tehnologii* 2022; (1): 12–29. DOI: 10.25881/18110193_2022_1_12. (In Rus.).
11. Gusev AV, Vladzimirskii AV, Golubev NA, Zarubina TV. Informatizacija zdravoohranenija Rossijskoj Federacii: istorija i rezul'taty razvitija [Informatization of healthcare in the Russian Federation: history and results of development]. *Nacional'noe zdravoohranenie* 2021; 2 (3): 5–17. DOI: 10.47093/2713-069X.2021.2.3.5-17. (In Rus.).
12. Maslovskaya LYu. Osobennosti medicinskoj terminologii i puti ejo popolnenija [The pecularities of medical terminology and its development paths]. *The Scientific Heritage* 2021; (63): 41–43. DOI: 10.24412/9215-0365-2021-63-5-41-43. (In Rus.).
13. Gapparova DA, Yusupova SH, Iskandarov DF. Problemy leksikograficheskogo opisanija medicinskih terminov [Modern medical literature and problems of lexicographic description of medical terms]. *Open innovation* 2019: 142–144. (In Rus.).
14. Abayev JuK. Horoshij doktor. Chast' 9. Terminy i «terminotvorchestvo» v medicine [Good doctor. Part 9. Terms and creation of terminology in medicine]. *Zdravoohranenie (Minsk)* 2020; (878): 28–37. (In Rus.).
15. Zarubina TV. Edinaja gosudarstvennaja informacionnaja sistema – osnova cifrovizacii zdravoohranenija [Unified state information system – the basis for digitalization of healthcare]. *Informacionnye tehnologii v medicine i zdravoohranenii* 2020: 22–35. (In Rus.).

16. Yusupova NI, Gayanova MM, Bogdanov MR. Izvlechenie informacii ob ispol'zovanii informacionnyh tehnologij dlja podderzhki prinjatija reshenij v medicinskoj diagnostike [Retrieving information about the use information technology to support decision-making in medical diagnostics]. *Vestnik Juzhno-Ural'skogo gosudarstvennogo universiteta. Serija: Komp'juternye tehnologii, upravlenie, radiojelektronika* 2022; 22 (1): 14–27. DOI: 10.14529/ctcr220102. (In Rus.).
17. Alpatov AN, Popov KS, Chesalin AN. Analiz tochnosti modelej mashinnogo obuchenija s ispol'zovaniem metodov vektorizacii dlja zadach klassifikacii raznorodnyh tekstovyh dannyh [Accuracy analysis of machine learning models using vectorization methods for heterogeneous text data classification tasks]. *International Journal of Open Information Technologies* 2022: 10 (7): 47–53. (In Rus.).
18. Harrison CJ, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction to natural language processing. *BMC Medical Research Methodology* 2021; 21: 158. DOI: 10.1186/s12874-021-01347-1
19. Zatsman IM. Problemno-orientirovannaja aktualizacija slovarnyh statej dvujazychnyh slovarej i medicinskoj terminologii: sopostavitel'nyj analiz [Problem-oriented updating of dictionary entries of bilingual dictionaries and medical terminology: comparative analysis]. *Informatika i ee primenenija* 2021; 15 (1): 94–101. DOI: 10.14357/19922264210113. (In Rus.).
20. Guo Y, Li C, Roan C, Pakhomov S, Cohen T. Crossing the «Cookie Theft» Corpus Chasm: Applying What BERT Learns From Outside Data to the ADReSS Challenge Dementia Detection Task. *Frontiers in Computer Science* 2021; 3: 642517. DOI: 10.3389/fcomp.2021.642517
21. Zemlyansky SA, Axyonov SV, Lyzin IA, Berestneva OG. Tematicheskoe modelirovanie v kontekste medicinskih tekstov [Topic modeling in the context of medical texts]. *Doklady Tomskogo gosudarstvennogo universiteta sistem upravlenija i radiojelektroniki* 2021; 24 (4): 58–64. DOI: 10.21293/1818-0442-2021-24-4-58-64. (In Rus.).
22. Mingazov DR. Obzor primenenija medicinskih jekspertnyh sistem [Rewiev of medical expert systems usage]. *Voprosy ustojchivogo razvitija obshhestva* 2022; (4): 1502–1505. (In Rus.).
23. Semenov NA, Burdo GB, Lebedev SN, Lebedeva YuV. Intellektual'naja podderzhka prinjatija reshenij v jekspertnyh sistemah pri diagnostike zabolevanij polosti rta [Intelligent decision support in expert systems in the diagnosis of oral cavity diseases]. *Programmnye produkty i sistemy* 2021; (3): 484–488. DOI: 10.15827/0236-235X.135.484-488. (In Rus.).
24. Chollet F. Deep Learning with Python, Second Edition. New York: Manning Publications; 2021. 504 p.
25. Mosolova AV., Fomin VV., Bondarenko IYu. Text augmentation for neural networks. CEUR Workshop Proceedings 2018; (2268): p. 104–109.
26. Cui M, Bai R, Lu Zh, Li X, Aickelin U, Ge P. Regular Expression Based Medical Text Classification Using Constructive Heuristic Approach. IEEE Access 2019; 7: p. 147892–147904. DOI: 10.1109/access.2019.2946622
27. Astanin PA, Ronzhin LV, Rauzina SE, Zarubina TV. Algoritmy semanticheskogo analiza dannyh i vozmozhnosti ih primenenija v razrabotke medicinskih informacionnyh system [Semantic analysis algorithms for data processing and possibilities of their usage in medical information systems development]. *Cifrovaja statistika. Novye zadachi i traektorija dvizhenija: Materialy IV S'ezda medicinskih statistikov Moskvy* 2022: 6–9. (In Rus.).
28. Chawla N, Bowyer K, Hall L, Kegelmeyer W. SMOTE: Synthetic Minority Over-sampling Technique. ArXiv 2002; 16: p. 321–357. DOI: 10.1613/jair.953
29. Delvin J, Chang M, Lee K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv 2019: n. pag.
30. Gusev AV, Vladzymyrskyy AV, Sharova DE, Arzamasov KM, Khramov AE. Razvitie issledovanij i razrabotok v sfere tehnologij iskusstvennogo intellekta dlja zdravoohranenija v Rossijskoj Federacii: itogi 2021 goda [Evolution of research and development in the field of artificial intelligence technologies for healthcare in the Russian Federation: results of 2021]. *Digital Diagnostics*. 2022: 3 (3): 178–194. DOI: 10.17816/DD107367. (In Rus.).

« Пред. След. »