ЛЕГАШЕВ Л.В.,
к.т.н., ФГБОУ ВО «Оренбургский государственный университет», Оренбург, Россия,
e-mail: silentgir@gmail.com

ШУХМАН А.Е.,
к.п.н., ФГБОУ ВО «Оренбургский государственный университет», Оренбург, Россия, e-mail: shukhman@gmail.com

БОЛОДУРИНА И.П.,
д.т.н., профессор, ФГБОУ ВО «Оренбургский государственный университет», Оренбург, Россия,
e-mail: ipbolodurina@yandex.ru

ГРИШИНА Л.С.,
ФГБОУ ВО «Оренбургский государственный университет», Оренбург, Россия, e-mail: zabrodina97@inbox.ru

ЖИГАЛОВ А.Ю.,
ФГБОУ ВО «Оренбургский государственный университет», Оренбург, Россия, e-mail: leroy137.artur@gmail.com

ОБРАБОТКА РУССКОЯЗЫЧНЫХ НЕСТРУКТУРИРОВАННЫХ МЕДИЦИНСКИХ ТЕКСТОВ И ВЕРОЯТНОСТНОЕ ПРОГНОЗИРОВАНИЕ ГРУПП ЗАБОЛЕВАНИЙ

DOI: 10.25881/18110193_2022_4_52

Аннотация.

Актуальность. Разработка и внедрение медицинских информационных систем позволило упростить и автоматизировать множество процессов в медицинских организациях. Вместе с тем, постоянно накапливаемый объём данных о здоровье пациентов позволяет решать множество задач, связанных с прогнозированием и диагностикой заболеваний.

Цель. Исследование подходов к обработке неструктурированных русскоязычных медицинских текстов и прогнозированию групп заболеваний на основе методов машинного обучения.

Материалы и методы. Исходные данные: Массив деперсонализированных данных медицинских организаций Оренбургской области, содержащий 119 780 записей. Исследуются три подхода к вероятностному прогнозированию групп медицинских заболеваний на основе неструктурированных медицинских текстов жалоб пациентов на русском языке: подход на основе правил, подход на основе логистической регрессии и подход с использованием моделей трансформеров BERT.

Результаты. Сравнительный анализ показывает, что подход с использованием логистической регрессии и метода TfidfVectorizer демонстрирует наилучшие результаты по метрикам Precision (0,8296), F1-score (0,8269) и Matthews's correlation coefficient (0,7695).

Выводы. Традиционный подход на основе правил является наименее эффективным (Precision = 0,7182) среди исследуемых методов, но при этом позволяет интерпретировать результаты работы классификатора в виде визуализации дерева решений. Подход с использованием логистической регрессии (Precision = 0,8296) и подход с использованием предобученных моделей BERT (Precision = 0,8164) показывают лучшие результаты классификации среди исследуемых методов и в дальнейшем могут послужить базисом для построения и развития систем поддержки принятия врачебных решений и найти применение в работе практикующих терапевтов.

Ключевые слова: *обработка естественного языка, цифровая медицина, электронные медицинские карты, логистическая регрессия, BERT.*

Для цитирования*: Легашев Л.В., Шухман А.Е., Болодурина И.П., Гришина Л.С., Жигалов А.Ю. Обработка русскоязычных неструктурированных медицинских текстов и вероятностное прогнозирование групп заболеваний. Врач и информационные технологии. 2022; 4: 52-63. doi: 10.25881/18110193_2022_4_52.*

LEGASHEV L.V.,
PhD, Orenburg State University, Orenburg, Russia,
e-mail: silentgir@gmail.com

SHUKHMAN A.E.,
PhD, Orenburg State University, Orenburg, Russia, e-mail: shukhman@gmail.com

BOLODURINA I.P.,
DSc, Prof., Orenburg State University, Orenburg, Russia,
e-mail: ipbolodurina@yandex.ru

GRISHINA L.S.,
Orenburg State University, Orenburg, Russia, e-mail: zabrodina97@inbox.ru

ZHIGALOV A.YU.,
Orenburg State University, Orenburg, Russia, e-mail: leroy137.artur@gmail.com

RUSSIAN UNSTRUCTURED CLINICAL TEXTS PROCESSING AND PROBABILISTIC CLASSIFICATION OF DISEASE GROUPS

DOI: 10.25881/18110193_2022_4_52

Abstract

Background. The development and implementation of medical information systems make it possible to simplify and automate many processes in medical organizations. At the same time, the amount of data on patients' health is constantly accumulating which allows solving many problems related to the prediction and diagnosis of diseases.

Aim. To study approaches to processing of Russian unstructured medical texts and to predicting certain groups of diseases based on machine learning methods.

Материалы и методы. Initial data consisted of an array of depersonalized data from medical organizations in the Orenburg region containing 119,780 records. Three approaches to probabilistic forecasting of groups of diseases based on unstructured medical texts of patient complaints in Russian were studied: rule-based approach, logistic regression-based approach and approach using BERT transformer models.

Results. Comparative analysis showed that показываем, logistic regression-based approach combined with TfidfVectorizer method had the best results in Precision (0,8296), F1-score (0,8269) and Matthews's correlation coefficient (0,7695).

Conclusion. Traditional rule-based approach was the least effective (Precision = 0,7182) among the studied methods, but at the same time it allowed to interpret the results of the classifier as visualization of the decision tree. Logistic regression-based approach (Precision = 0,8296) and approach using BERT transformer models (Precision = 0,8164) showed the best classification results and can be further used as a basis for building and developing medical decision support systems and find application in medical practice.

***Keywords**: natural language processing, digital medicine, electronic health records, logistic regression, BERT.*

***For citation:** Legashev L.V., Shukhman A.E., Bolodurina I.P., Grishina L.S., Zhigalov A.Yu. Russian unstructured clinical texts processing and probabilistic classification of disease groups. Medical doctor and information technology. 2022; 4: 52-63. doi: 10.25881/18110193_2022_4_52.*

ВВЕДЕНИЕ

Методы обработки естественного языка (Natural Language Processing, NLP) активно применяются для решения различных задач обработки электронных медицинских карт (Electronic health records, EHRs), в частности для распознавания рассеянного склероза [1], аксиального спондилоартрита [2], гепатоцеллюлярного рака [3], сахарного диабета [4], биполярного расстройства [5], беременных с суицидальным поведением [6], выявления передозировок, связанных с опиоидами [7], диагностики инфекционных заболеваний [8], выявления пациентов с метастатическим раком молочной железы [9] и др.

Новейшие исследования в области здравоохранения посвящены использованию языковых моделей на основе трансформеров с поддержкой контекстуализированных эмбеддингов (contextual embeddings), которые используются для векторного представления многозначных слов с учетом контекста предложения, содержащего то или иное слово. Rasmy et al. в работе [10] описали модель контекстуализированных эмбеддингов Med-BERT, предварительно обученную на структурированном наборе данных, содержащем более 28 миллионов записей электронных медицинских карт пациентов. Nath et al. в работе [11] проводят исследования публичных моделей векторного пространства для решения задачи распознавания именованных сущностей (Named Entity Recognition, NER), при этом лучшие результаты показывает модель Bio + clinical BERT. Li et al. в работе [12] представляют обзор современных подходов к обработке неструктурированных медицинских текстов, которые отличаются от традиционных статистических систем и систем на основе правил. В исследовании [13] Syed et al. описывают гибридную архитектуру искусственной нейронной сети с комбинированными контекстуализированными эмбеддингами моделей BERT и FLAIR для решения задачи скрининга колоректального рака.

Проводятся активные исследования в области NLP для русского языка. В статье Ялунина и др. [14] представлены модели RuBioBERT и RuBioRoBERTa для анализа биомедицинских текстов на русском языке. В статье Блинова и др. [15] описывается бенчмарк понимания русского медицинского языка, частично решая проблему отсутствия универсального медицинского датасета. В задаче интеллектуального анализа клинического текста решают проблему обнаружения отрицаний [16] и автоматической коррекции орфографии [17].

В рамках данного исследования представлен сравнительный анализ нескольких подходов к решению задачи прогнозирования группы заболеваний на основе русскоязычных неструктурированных данных медицинских карт пациентов: традиционный подход на основе правил, подход с использованием логистической регрессии и подход с использованием предобученной модели EnRuDR-BERT.

Исходные данные: массив деперсонализированных формализованных данных электронных медицинских карт пациентов, проходивших обследование и лечение в медицинских организациях Оренбургского региона, содержащий 119780 записей. Выполнена предварительная обработка датасета: удалены пропущенные значения и записи, а также записи, в которых длина строки с жалобами пациента меньше 100 символов. Также выполнена разбивка датасета (Табл. 1) на шесть основных групп по коду МКБ (при этом использованы первые два символа кода МКБ). Для каждого текстового описания жалоб пациента в исходном массиве данных поставлен один диагноз по коду МКБ, в связи с чем в проводимом исследовании речь идет о задаче многоклассовой классификации.

Итоговый датасет для исследования содержит 16601 запись. На рисунке 1 представлены два примера неструктурированных текстов жалоб пациентов. Всего выделено пять специфических групп заполнения поля с жалобами пациентов:

- красным цветом выделены опечатки и грамматические ошибки;
- желтым цветом выделены медицинские аббревиатуры и сокращения терминов на английском и русском языке, при этом часть аббревиатур может иметь несколько значений (например, «ддп» может обозначать «давление в дыхательных путях», «дыхательные движения плода», «добавочная доля печени», «дегенеративно-дистрофическое поражение» и т.д.);
- зеленым цветом выделены числовые показатели, часть из которых может быть полезна в постановке диагноза, при этом в некоторых

Таблица 1 — Распределение данных по шести группам заболеваний по коду МКБ

№	Код МКБ	Название	Кол-во записей
1	I1	Болезни, характеризующиеся повышенным кровяным давлением	5243
2	I6	Цереброваскулярные болезни	4589
3	I2	Ишемическая болезнь сердца, легочное сердце и нарушения легочного кровообращения	4138
4	I4	Другие болезни сердца	1569
5	J0	Острые респираторные инфекции верхних дыхательных путей	630
6	U0	Временные обозначения новых диагнозов неясной этиологии	432

Рисунок 1 — Примеры неструктурированных текстовых данных жалоб пациентов.

случаях ключевые цифры могут указываться несколько раз в тексте (например, зафиксированное самое высокое давление и текущее давление на приёме у врача);

- голубым цветом выделены произвольные сокращения слов, при этом так же может возникать двусмысленность (например «с-м» может быть сокращением слов «симптом» и «синдром»);
- фиолетовым цветом выделена оценочная характеристика потенциальных ключевых слов и фраз, влияющих на постановку диагноза, при этом наличие того или иного признака может быть выражено как в текстовой форме («улучшилось», «отрицает, «не выражено» и т.д.), так и в произвольной символьной форме («(+)», «(-)», «(–)» и т.д.).

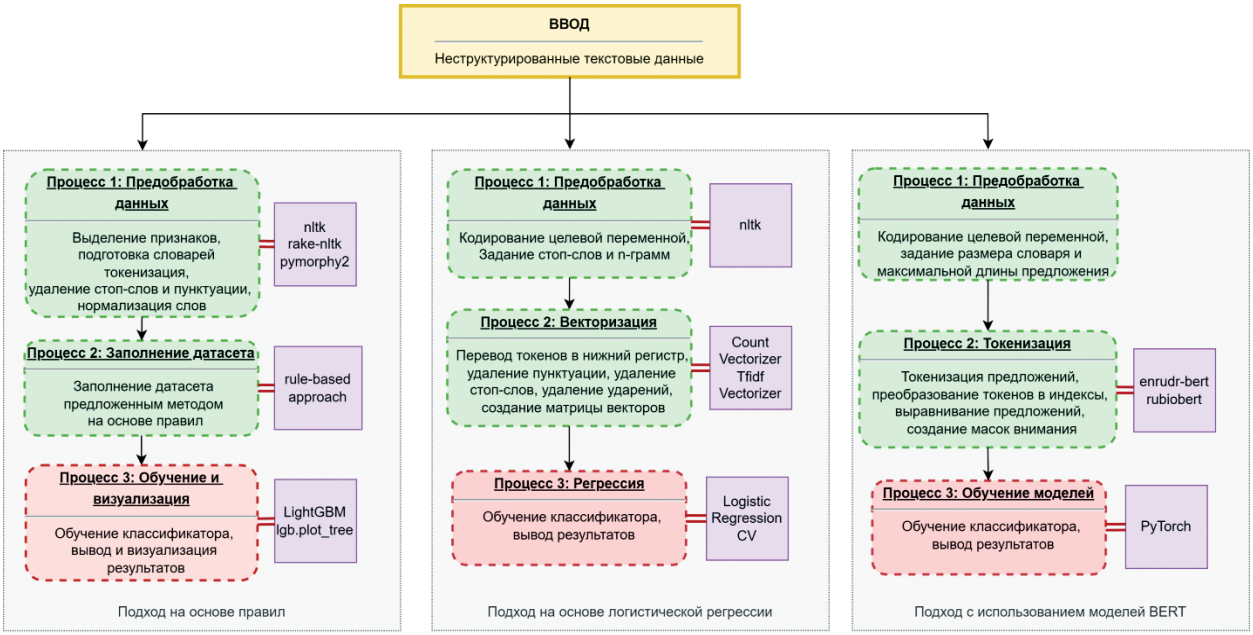

Рисунок 2 — Общая схема исследования трех подходов к обработке неструктурированных текстовых данных.

Данная работа организована следующим образом: в первом разделе рассматривается метод генерации численных признаков с использованием подхода на основе правил и классификатора LightGBM. Во втором разделе рассматривается подход с использованием логистической регрессии для построения матрицы количества токенов из коллекции текстовых медицинских документов. В третьем разделе рассматривается подход на основе контекстуальных эмбеддингов с использованием обученных русскоязычных моделей BERT. В четвертом разделе приводятся результаты сравнительного анализа трёх методов и их обсуждение. Общая схема исследования представлена на рисунке 2.

РАЗДЕЛ 1: ПОДХОД НА ОСНОВЕ ПРАВИЛ (RULE-BASED APPROACH, RBA)

Подход построения датасета из неструктурированных медицинских текстов на основе правил состоит в последовательном решении двух задач. Первая задача заключается в конструировании признаков, вторая задача заключается в обучении классификатора для определения группы заболеваний по коду МКБ.

Задача 1. Конструирование признаков

Процесс 1.1. Распознавание именованных сущностей.

В рамках первого этапа необходимо решить задачу распознавания именованных сущностей: выделить ключевые слова и ключевые фразы, наиболее часто встречающиеся в тексте. Для этого все жалобы пациентов объединяются в единый текстовый файл, который обрабатывается методом extract_keywords_from_text библиотеки rake-nltk.

В качестве параметров используется стандартный словарь стоп-слов из русскоязычного корпуса библиотеки nltk и задается минимальная и максимальная длина *n*-грам — последовательностей из *n* слов — от 1 до 4.

Процесс 1.2. Создание заголовков датасета.

В рамках второго этапа полученные ключевые фразы ранжируются в порядке убывания и выбирается несколько наиболее часто встречаемых признаков. Всего выделено три группы признаков (ключевых терминов):

- **complaint_synonyms**. Упоминание признака в тексте: «Слабость», «Туберкулез отрицает», «Симптом Пастернацкого отрицательный» и т.д.

№	Признак	Значение
1	Weakness	Слабость
2	Dizziness	Головокружение
3	Rheum	Насморк
4	Tuberculosis	Туберкулез
5	Noise	Шум в ушах
6	Staggering_Walk	Пошатывание при ходьбе
7	Numbness	Онемение
8	Dyspnea	Одышка
9	Nausea	Тошнота
10	Smoking	Курение
11	Dry_Mouth	Сухость во рту
12	Wheezing	Хрипы
13	HIV	ВИЧ
14	Irradiation	Иррадиация
15	Hepatitis	Гепатит
16	Speech_Disorder	Нарушение речи
17	Tremor	Дрожь
18	Bleeding	Кровотечение
19	Pasternatsky	Симптом Пастернацкого

- **complaint_synonyms_adjectives**. Признак с оценочной характеристикой в тексте: «Боль острая», «Мягкий живот», «Сердечные тоны не приглушены», «Слизистые без особенностей» и т.д.

№	Признак	Значение
20	Pain	Боль
21	Cough	Кашель
22	Stool	Стул
23	Memory	Память
24	Sleeping	Сон
25	Abdomen	Живот
26	Breath	Дыхание
27	Heart_Sounds	Тоны сердца
28	Skin	Кожные покровы
29	Lymph_Nodes	Лимфоузлы

- **complaint_synonyms_numerical**. Количественные признаки в тексте: «Температура 38.7», «гипертония 160/100», “ЧДД 23” и т.д.

№	Признак	Значение
30	Heart_Rate	Частота сердечных сокращений
31	Respiratory_Rate	Частота дыхательных движений
32	Temperature	Температура
33	Blood_Pressure_Lower	Артериальное давление нижняя граница
34	Blood_Pressure_Upper	Артериальное давление верхняя граница

В дополнение к вышеуказанным признакам в рамках данного исследования выделены также признаки, дополняющие симптоматику: ‘Pain_Frequency’ содержит показатель частоты упоминания слова «боль» и его вариаций в тексте, ‘Good_Frequency’ содержит показатель частоты упоминания «положительных» описательных характеристик признаков (например, ‘нормальный’, ‘обычный’, ‘ровный’, ‘безболезненный’ и т.д.), а ‘Bad_Frequency’ содержит показатель частоты упоминания «отрицательных» описательных характеристик признаков (например, ‘вздутый’, ‘болезненный’, ‘давящий’, ‘отёчный’ и т.д.). Итоговый датасет содержит 37 новых признаков.

Процесс 1.3. Предобработка данных.

На третьем этапе в соответствии с тремя группами признаков подготавливаются три словаря, содержащие всевозможные вариации признака (например, словарь для температуры содержит следующие элементы: [‘температура’, ‘т’, ‘т.’, ‘t’, ‘темп’, ‘°’]). Далее выполняются стандартные операции предобработки текстовых данных: токенизация, удаление стоп-слов, удаление пунктуации, а также нормализация слов.

Процесс 1.4. Заполнение датасета.

Заполнение датасета на основе созданных словарей признаков происходит на четвертом этапе. Для группы **complaint_synonyms** при обнаружении ключевого слова в тексте и отсутствии отрицательных частиц и слов («нет», «без», «отсутствует», «отрицает») запись признака принимает значение 1, в противном случае запись признака принимает значение -1. Для группы **complaint_synonyms_adjectives** при обнаружении ключевого слова в тексте оценивается «окрестность» слова на наличие «положительных» (запись признака принимает значение -1) или «отрицательных» (запись признака принимает значение 1) описательных характеристик, либо на отсутствие того и другого

(запись признака принимает значение 0). Для группы **complaint_synonyms_numerical** используется метод nums_from_string.get_nums языка Python для выделения всех чисел из строки, после чего выполняется обработка в соответствии с типом признака. В частности, показатель давления разбивается на два значения: верхняя и нижняя границы.

Процесс 1.5. Постобработка заполненного датасета.

В рамках пятого этапа осуществляется постобработка датасета в соответствии со следующими операциями: незаполненные значения заменяются на стандартные, в соответствии с видом признака, чтобы отразить отсутствие жалоб пациента на данный показатель. Например, если в тексте не указаны жалобы на боли в животе, соответствующий признак будет содержать значение -1. Если не указаны жалобы на высокий пульс или не выполнялись измерения, соответствующий признак будет содержать значение 60, и т.д.; для числовых показателей значения за пределами возможных также заменяются на дефолтные. Таким образом, решением первой задачи выступает полученный набор данных, характеризующий структурированное описание жалоб пациента на приеме.

Задача 2. Вероятностная классификация

Процесс 2.1. Классификация.

Полученный датасет разбивается на тренировочную и тестовую выборки (train_size = 0,7, test_size = 0,3), для обучения используется классификатор LightGBM со стандартными параметрами. На следующем этапе выполняется подсчет метрик. Результаты сравнения подходов представлены в секции 4. Метод predict_proba используется для получения предсказанной вероятности принадлежности выборки к одной из шести групп заболеваний по МКБ.

Процесс 2.2. Удаление избыточных признаков.

Дополнительно реализован алгоритм случайного поиска (random search algorithm, RSA), который перебирает подмножества исходного множества признаков в интервале от 3 до 37, а также варьирует параметры классификатора LightGBM, такие как max_depth, min_data_in_leaf, n_estimators, num_leaves, reg_alpha, и subsample с целью улучшения метрик. Результаты выполняемой оптимизации будут представлены в секции 4 как RBA+RSA подход.

Процесс 2.3. Визуализация результатов.

Основное достоинство представленного подхода к прогнозированию группы диагноза заболеваний пациента состоит в возможности интерпретации полученных результатов, которая очень важна при решении задач в области цифровой медицины. В частности, для обученного классификатора LightGBM визуализировать дерево принятия решений можно с помощью различных библиотек Python (lgb.plot_tree, dtreeviz и др.). Пример фрагмента итогового дерева принятия решений представлен на рисунке 3. Результаты исследования дерева решения показали, что признак верхней границы давления является наиболее значимым при определении группы заболевания.

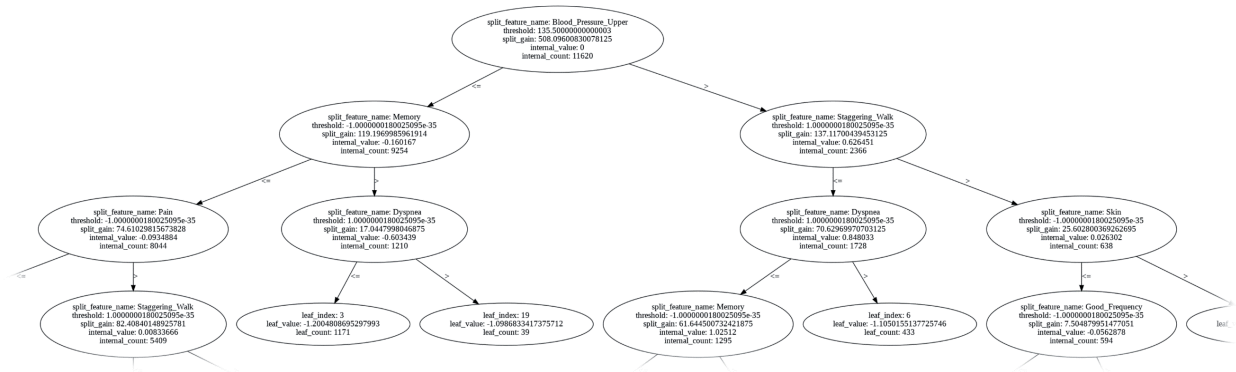

Рисунок 3 — Фрагмент дерева принятия решений классификатора LightGBM.

РАЗДЕЛ 2: ПОДХОД НА ОСНОВЕ ЛОГИСТИЧЕСКОЙ РЕГРЕССИИ (LOGISTIC REGRESSION, LOGREG)

Альтернативный подход, реализованный для решения задачи диагностирования групп заболеваний, — использование логистической регрессии, которая основана на линейном разделении объектов и является классическим методом машинного обучения. Важно отметить, что основное отличие данного метода от предыдущего подхода состоит в предварительном применении метода преобразования текста в его векторное представление.

Процесс 1: Предобработка данных.

На первом этапе выполняется числовое кодирование целевой переменной — названия шести групп заболеваний по МКБ. Задается словарь стоп-слов из русскоязычного корпуса библиотеки nltk и задается минимальная и максимальная длина n-грам от 1 до 3.

Процесс 2: Векторизация.

Предварительно выполняются операции перевода токенов в нижний регистр, удаления пунктуации, удаления стоп-слов, удаление ударений и др. Далее необходимо выполнить векторизацию — преобразование документа в числовой вектор [18]. На этом этапе использовались два разных объекта из библиотеки SkLearn. Объект CountVectorizer использует модель мешка слов (bag-of-words). В этом случае формируется словарь n-грамм длины m, и каждый текст представляется вектором длины m, в котором каждый элемент соответствует количеству вхождений соответствующей n-граммы в текст. В результате получаются разреженные векторы, в которых большинство элементов равны нулю. Объект TfidfVectorizer вместо количества вхождений для n-граммы сохраняет коэффициент TF-IDF [18, с. 116–117], который принимает наибольшие значения для n-грамм, которые часто встречаются в кодируемом документе, но редко — в остальных документах коллекции.

Выведем список наиболее часто встречающихся ключевых терминов, который достаточно близко пересекается с множеством признаков, выделенных в подходе на основе правил в первом разделе статьи:

(‘боли’, 7307), (‘ад’, 5490), (‘слабость’, 4690), (‘головокружение’, 4554), (‘головные’, 4173), (‘головные боли’, 4159), (‘повышение’, 3740), (‘головокружение ,’, 3708), (‘боли ,’, 3562), (‘, головокружение’, 3290), (‘слабость ,’, 2945), (‘одышку’, 2729), (‘повышение ад’, 2628), (‘ходьбе’, 2416), (‘, слабость’, 2407), (‘сердца’, 2403), (‘боль’, 2393), (‘жалобы’, 2154), (‘шум’, 2137), (‘нагрузке’, 2064), (‘, повышение’, 2056), (‘области’, 2000), (‘, одышку’, 1845), (‘ад ,’, 1817), (‘, шум’, 1807), (‘периодические’, 1759), (‘головную’, 1729), (‘головную боль’, 1723), (‘, боли’, 1695), (‘боль ,’, 1629), (‘сердцебиение’, 1495), (‘ходьбе ,’, 1493), (‘области сердца’, 1455), (‘памяти’, 1442), (‘голове’, 1439), (‘утомляемость’, 1412), (‘нагрузке ,’, 1382), (‘, головные’, 1322), (‘грудиной’, 1320), (‘100’, 1320), (‘снижение’, 1306), (‘сердца ,’, 1278), (‘ушах’, 1220), (‘боли области’, 1191), (‘одышка’, 1185), (‘сердце’, 1184), (‘физической’, 1158), (‘, снижение’, 1148), (‘голове ,’, 1120), (‘общую слабость’, 1075), (‘утомляемость ,’, 1025)

Процесс 3: Тренировка классификатора.

Полученные векторные текстовые эмбеддинги разбиваются на тренировочную и тестовую выборки (train_size = 0,7, test_size = 0,3), для обучения используется классификатор LogisticRegressionCV с поддержкой кросс-валидации. На следующем этапе выполняется подсчет метрик. Результаты сравнения подходов представлены в секции 4. Метод predict_proba используется для получения предсказанной вероятности принадлежности выборки к одной из шести групп заболеваний по МКБ.

РАЗДЕЛ 3: ПОДХОД НА ОСНОВЕ МОДЕЛЕЙ ТРАНСФОРМЕРОВ BERT

В заключение, рассмотрим подход с использованием русскоязычных моделей трансформеров BERT на неструктурированных медицинских текстах, который заключается в дообучении предварительно обученной нейронной сети с дополненными слоями классификатора на размеченном наборе данных.

Процесс 1: Предобработка данных.

На первом этапе выполняется числовое кодирование целевой переменной — названия шести групп заболеваний по МКБ. Задается максимальный размер словаря num_words = 15000 и максимальная длина сообщения max_len = 200

в токенах, происходит выравнивание предложений исходного датасета до одинаковой длины (padding='post').

Процесс 2: Токенизатор.

Выполняется токенизация обучающей выборки с помощью модели EnRuDR-BERT [19], предварительно обученной на коллекции отзывов потребителей о приеме лекарств, и модели RuBioBERT [14], предварительно обученной на корпусе свободно доступных текстов в области биомедицины.

Модель EnRuDR-BERT имеет общий размер словаря 119547, включает в себя последовательность следующих блоков: входной слой вложений, который формирует 768-байтовое векторное представление токена; кодировщик, состоящий из 12 блоков трансформеров, включая слой внимания, полносвязные слои и слои нормализации; последний полносвязный слой — пулер. Модель RuBioBERT имеет общий размер словаря 120138. Стоит отметить, что изначально RuBERT включает выходной слой, который предсказывает замаскированные слова в тексте (Masked Word Prediction). Для решения задачи классификации группы заболеваний он заменен выходным слоем с шестью выходами в соответствии с таблицей 1. Создается маска внимания для каждого примера обучающей выборки. Единицами заполняются те токены, которые нужно учитывать при обучении и вычислении градиентов, нулями заполняются те токены, которые следует пропустить.

Процесс 3: Тренировка модели

Векторные представления формируются с помощью входного слоя нейронной сети на основе списка словарных номеров текстовых токенов. Выполняется обучение и тестирование модели. Количество эпох подбирается экспериментально (epoch = 2). В результате ошибка на обучающем и проверочном датасете имеет следующие значения — train_loss: 0,5425, val_loss: 0,5644. Функция softmax библиотеки torch используется для получения предсказанной вероятности принадлежности выборки к одной из шести групп заболеваний по МКБ.

РАЗДЕЛ 4: СРАВНЕНИЕ ПОДХОДОВ И ОЦЕНКА МОДЕЛЕЙ

Результаты сравнения трех исследуемых подходов к прогнозированию групп заболеваний по метрикам Precision, F1-score и Matthews correlation coefficient (MCC) представлены в таблице 2.

Подход на основе логистической регрессии показывает лучшие результаты по всем трем метрикам (Precision = 0,8296, F1-score = 0,8269, MCC = 0,7695) среди исследуемых методов. При этом подход на основе предобученных моделей BERT имеет точность в среднем меньше на 1,6%, а подход на основе правил имеет точность в среднем меньше на 12,3%. Отметим, что применение алгоритма случайного поиска в подходе на основе правил не оказало существенного влияния на метрики классификации (наблюдается повышение значения метрик в среднем на 1,47%). Основные достоинства и недостатки предложенных подходов представлены в таблице 3.

Примеры распределения вероятностей классов тремя подходами для нескольких жалоб пациентов представлены на рисунке 4. На рисунках 4(a)-4(d) жалобы выбраны из исходного датасета случайным образом, на рисунках 4(e) и 4(f) представлен произвольный текст с возможными признаками ОРВИ и COVID-19. По комментарию терапевта разброс классов вероятностей на примерах 4(c), 4(d) и 4(f) возникает по причине того, что перечисленные жалобы могут относиться к нескольким группам заболеваний и на практике врач принимает окончательное решение, ориентируясь на личный опыт и возможные результаты дополнительных обследований.

Таблица 2 — Сравнение метрик для исследуемых подходов

Подход	Precision	F1-score	MCC
RBA	0,6956	0,6898	0,5856
RBA + RSA	0,7182	0,7034	0,5936
Logistic Regression + CountVectorizer	0,8187	0,8161	0,7551
Logistic Regression + TfidfVectorizer	0,8296	0,8269	0,7695
EnRuDR-BERT	0,8095	0,8088	0,7450
RuBioBERT	0,8164	0,8137	0,7508

Таблица 3 — Преимущества и недостатки исследуемых подходов

Подход	Преимущества	Недостатки
RBA	Хорошая интерпретация и визуализация решений.	Сильно зависим от вида входных данных, требует ручной донастройки пространства признаков и словарей.
RBA+RSA	Позволяет незначительно улучшить значения метрик производительности за счет снижения размерности признакового пространства и подбора гиперпараметров классификатора.	При сильном увеличении объёма исходных данных RSA работает достаточно медленно.
Logistic Regression	Позволяет проводить отбор признаков по анализу коэффициентов модели и выделить наиболее значимые термины для предметной области.	Отсутствует возможность визуализации принятых решений.
BERT	Позволяет учитывать контекст слов в предложениях, следовательно точнее определять значения токенов для предметной области.	Требует дообучения на специализированных медицинских текстах, работает по принципу «черного ящика» (blackbox), т.е. отсутствует возможность интерпретации и визуализации принятых решений, ограниченный размер вводимого текста.

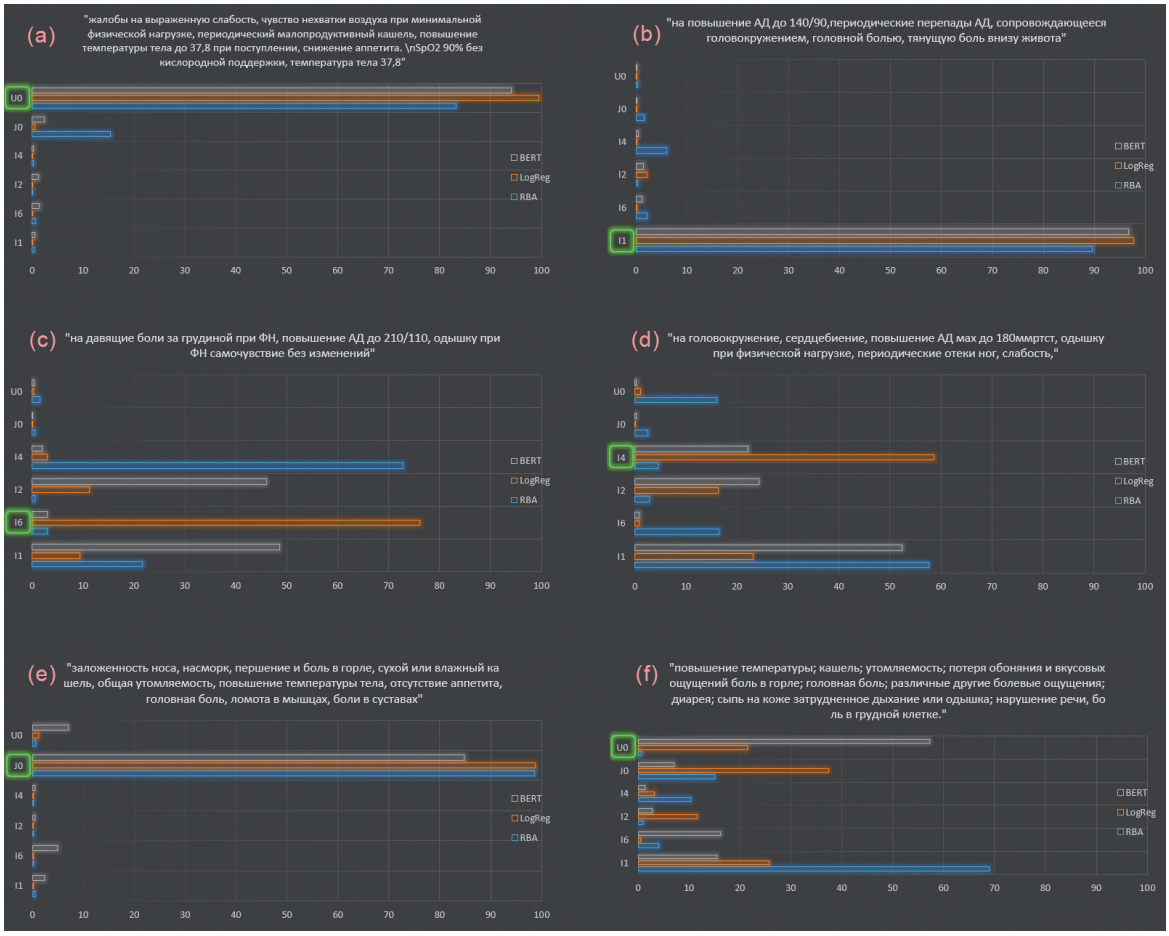

Рисунок 4 — Примеры распределения вероятностей классов для русскоязычных неструктурированных медицинских текстов.

ОБСУЖДЕНИЕ

Ограничения проведённого исследования заключаются в лимитированной подвыборке нозологий, относящихся к группам сердечно-сосудистых заболеваний, острых респираторных инфекций верхних дыхательных путей и новых диагнозов неясной этиологии (COVID-19). Деперсонализированные данные электронных медицинских карт получены для пациентов Оренбургской области и могут быть неприменимы к другим входным условиям. Используемая в исследовании модель EnRuDR-BERT обучена на текстовом корпусе отзывов потребителей на русском языке о фармацевтических продуктах, что может не совсем идеально соответствовать задаче описания жалоб пациентов и постановке диагноза. При этом использование более релевантной модели RuBioBERT показало лишь незначительный прирост по исследуемым метрикам (в среднем на 0,59%). Вместе с тем следует отметить, что исследуемая модель прогнозирования диагноза заболеваний на основе логистической регрессии может быть обобщена для других групп заболеваний при условии расширения исходного набора данных. Проведенное исследование может послужить базисом для построения и развития систем поддержки принятия врачебных решений и найти применение в работе практикующих терапевтов.

В результате выполнения исследования достигнуты следующие результаты:

1. Разработан подход построения датасета из неструктурированных медицинских текстов на основе правил, который состоит в последовательном конструировании признаков и обучении классификатора LightGBM для определения группы заболеваний по коду МКБ. Кроме того, реализована модификация данного метода в части эффективного выбора подмножества признаков и параметров классификатора алгоритмом случайного поиска. Данный метод продемонстрировал приемлемую точность классификации (Precision = 0,7182) с возможностью визуальной интерпретации результатов.
2. Реализован подход на основе преобразования коллекции неструктурированных текстовых документов с жалобами пациентов в матрицу в рамках модели мешка слов и последующего прогнозирования группы заболеваний методом логистической регрессии. Данный подход показал наиболее высокую точность классификации (Precision = 0,8296) среди рассматриваемых методов, однако результаты модели не поддаются визуализации.
3. Применение предобученных моделей EnRuDR-BERT и RuBioBERT на текстовом корпусе отзывов потребителей на русском языке о фармацевтических продуктах и на корпусе свободно доступных текстов в области биомедицины показало высокую точность классификации (Precision = 0,8164), однако не самое эффективное среди рассматриваемых методов. Для повышения качества решения задачи описания жалоб пациентов и постановке диагноза необходимо расширять исходный набор данных и дообучать модель на специализированных медицинских данных.

ЗАКЛЮЧЕНИЕ

В работе выполнена задача вероятностного прогнозирования классов шести основных групп заболеваний по коду МКБ на основе неструктурированных медицинских данных электронных карт пациентов. Реализовано три метода решения задачи: традиционный подход на основе правил, подход с использованием логистической регрессии и методов TfidfVectorizer и CountVectorizer, а также подход с использованием предобученных моделей EnRuDR-BERT и RuBioBERT. Проведенный анализ по метрикам Precision, F1-score и Matthews correlation coefficient показывает, что подход с использованием логистической регрессии дает лучшие результаты (Precision = 0,8296, F1-score = 0,8269, MCC = 0,7695) среди исследуемых методов, а подход на основе предобученных моделей BERT имеет точность в среднем меньше на 1,6%. Традиционный подход на основе правил является наименее эффективным (Precision = 0,7182), но при этом позволяет интерпретировать результаты работы классификатора в виде визуализации дерева решений.

ЛИТЕРАТУРА/REFERENCES

1. Chase HS, Mitrani LR, Lu GG, Fulgieri DJ Early recognition of multiple sclerosis using natural language processing of the electronic health record. BMC medical informatics and decision making. 2017; 17(1): 1-8.
2. Zhao SS, Hong C, Cai T, Xu C, Huang J, Ermann J et al. Incorporating natural language processing to improve classification of axial spondyloarthritis using electronic health records. Rheumatology. 2020; 59(5): 1059-1065.
3. Sada Y, Hou J, Richardson P, El-Serag H, Davila J Validation of case finding algorithms for hepatocellular cancer from administrative data and electronic health records using natural language processing. Medical care. 2016; 54(2): 1-15.
4. Zheng L, Wang Y, Hao S, Shin AY, Jin B, Ngo AD et al. Web-based real-time case finding for the population health Management of Patients with Diabetes Mellitus: a prospective validation of the natural language processing–based algorithm with statewide electronic medical records. JMIR medical informatics. 2016; 4(4): 1-13.
5. Castro VM, Minnier J, Murphy SN, Kohane I, Churchill SE, Gainer V et al. Validation of electronic health record phenotyping of bipolar disorder cases and controls. American Journal of Psychiatry. 2015; 172(4): 363-372.
6. Zhong QY, Mittal LP, Nathan MD, Brown KM, Knudson González D, Cai T et al. Use of natural language processing in electronic medical records to identify pregnant women with suicidal behavior: towards a solution to the complex classification problem. European journal of epidemiology. 2019; 34(2): 153-162.
7. Hazlehurst B, Green CA, Perrin NA, Brandes J, Carrell DS, Baer A et al. Using natural language processing of clinical text to enhance identification of opioid- related overdoses in electronic health records data. Pharmacoepidemiology and drug safety. 2019; 28(8): 1143-1151.
8. Wang M, Wei Z, Jia M, Chen L, Ji H Deep learning model for multi-classification of infectious diseases from unstructured electronic medical records. BMC medical informatics and decision making. 2022; 22(1): 1-13.
9. Ling AY, Kurian AW, Caswell-Jin JL, Sledge Jr GW, Shah NH, Tamang SR Using natural language processing to construct a metastatic breast cancer cohort from linked cancer registry and electronic medical records data. JAMIA open. 2019; 2(4): 528-537.
10. Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ digital medicine. 2021; 4(1): 1-13.
11. Nath N, Lee SH, McDonnell MD, Lee I The quest for better clinical word vectors: Ontology based and lexical vector augmentation versus clinical contextual embeddings. Computers in Biology and Medicine. 2021; 134: 1-11.
12. Li I, Goldwasser J, et al. Neural natural language processing for unstructured data in electronic health records: A review. Computer Science Review. 2022; 46: 1-29.
13. Syed S, Angel AJ, Syeda HB, Jennings CF, VanScoy J, Syed M et al. The h-ANN Model: Comprehensive Colonoscopy Concept Compilation Using Combined Contextual Embeddings. NIH Public Access, 2022; 5: 1-24.
14. Yalunin A, Nesterov A, Umerenkov D RuBioRoBERTa: a pre-trained biomedical language model for Russian language biomedical text mining. arXiv preprint arXiv:2204.03951. 2022.
15. Blinov P, Reshetnikova A, Nesterov A, Zubkova G, Kokh V RuMedBench: A Russian Medical Language Understanding Benchmark. arXiv preprint arXiv:2201.06499. 2022.
16. Funkner AA, Balabaeva K, Kovalchuk SV Negation Detection for Clinical Text Mining in Russian. MIE. 2020: 342-346.
17. Balabaeva K, Funkner AA, Kovalchuk SV Automated Spelling Correction for Clinical Text Mining in Russian. MIE. 2020: 43-47.
18. Батура Т.В. Математическая лингвистика и автоматическая обработка текстов. — Новосибирск: РИЦ НГУ, 2016. [Batura TV. Mathematical linguistics and automatic text processing. Novosibirsk: RIC NSU. 2016. (In Russ.)]
19. 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. 2021; 37(2): 243-249.