ОБЕЗЛИЧИВАНИЕ МЕДИЦИНСКИХ ТЕКСТОВЫХ ДАННЫХ С ЦЕЛЬЮ РАЗРАБОТКИ И ВНЕДРЕНИЯ СИСТЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Макарова Е.А.[1], Андрейченко А.Е.[2], Казакова М.А.[3], Иванов Д.А.[4], Гусев А.В.[5]

***Ключевые слова**: детекция персональных данных, обработка естественного языка, обработка текстовых данных, поиск именованных сущностей, медицинские информационные системы, большие данные, языковые модели.*

Аннотация

Цель статьи: развитие и внедрение технологий обезличивания медицинских текстовых данных с целью разработки систем искусственного интеллекта для здравоохранения, что является одним из приоритетов цифровой трансформации.

Методы исследования: обзор и сравнение научных источников, посвященных обезличиванию медицинских данных и поиску именованных сущностей; валидация результатов работы программного комплекса на размеченном экспертами наборе данных.

Полученные результаты: описаны проблемы сбора и обработки больших объемов медицинских данных, которые необходимы как для обучения моделей искусственного интеллекта, так и для проведения научных исследований на основе реальной клинической практики; поставлена задача обезличивания данных из медицинских информационных систем; описаны персональные данные, накапливаемые в электронных медицинских картах, такие как ФИО, дата рождения, номера документов и т. д.; описаны сложности обезличивания подобных записей, связанные с тем, что готовые решения для поиска именованных сущностей не подходят для работы с медицинскими данными без дообучения моделей или дополнительной обработки результата; описана архитектура и ключевые программные компоненты разработанного сервиса Webiomed.Anonymizer, поддерживающего выявление и удаление персональной информации из текстовых медицинских записей.

Практическая ценность: приведены примеры работы сервиса с различными данными; проведена экспериментальная апробация разработанных алгоритмов путем валидации на выборке текстовых медицинских записей, содержащих персональные данные, из разных регионов Российской Федерации; приведены полученные на выборке метрики точности детекции атрибутов персональных данных сервиса, таких как ФИО, дата рождения, документы, сведения о работе и т. д.

DOI: 10.24682/1994-1404-2024-3-96-105

[1] ***Макарова Елена Андреевна,*** *кандидат технических наук, ООО «К-Скай», г. Петрозаводск, Российская Федерация. ORCID: 0000-0002-5410-5890.*
E-mail: emakarova@webiomed.ru

[2] ***Андрейченко Анна Евгеньевна,*** *кандидат физико-математических наук, ООО «К-Скай», г. Петрозаводск, Российская Федерация. ORCID: 0000-0001-6359-0763.*
E-mail: aandreychenko@webiomed.ru

[3] ***Казакова Мария Анатольевна,*** *кандидат медицинских наук, ООО «Ритейл», г. Казань, Российская Федерация. ORCID: 0000-0002-8669-3383.*
E-mail: mak26091986@gmail.com

[4] ***Иванов Дмитрий Александрович,*** *ООО «К-Скай», г. Петрозаводск, Российская Федерация. ORCID: 0009-0001-4079-2642.*
E-mail: divanov@webiomed.ru

[5] ***Гусев Александр Владимирович,*** *кандидат технических наук, ФГБУ «ЦНИИОИЗ Минздрава РФ», г. Москва, Российская Федерация. ORCID: 0000-0002-7380-8460.*
E-mail: agusev@webiomed.ai

Введение

Развитие и внедрение технологий искусственного интеллекта (далее — ИИ) является приоритетным направлением для цифровой трансформации российской социально-экономической сферы, в том числе для здравоохранения. Актуальная практика и перспективы применения ИИ в здравоохранении широкие: системы поддержки принятия врачебных решений, ИИ-ассистенты, анализ накопленных данных с целью поиска скрытых закономерностей, поддержка проведения клинических исследований и т. д.

Сбор и обработка больших объемов медицинских данных необходимы как для обучения и мониторинга моделей искусственного интеллекта, так и для проведения научных исследований. Одним из важных аспектов обеспечения безопасности и этичности этого процесса является обезличивание медицинских данных пациентов [1]. Даже в задаче обезличивания хорошо структурированных данных всё ещё изучается эффективность различных подходов [2]. Также и обезличивание неструктурированных, текстовых данных из медицинских документов является препятствием для обеспечения безопасного хранения и использования данных из медицинских информационных систем. Данные, позволяющие идентифицировать пациента, должны присутствовать в протоколах осмотра, эпикризах, инструментальных обследованиях. Речь идёт о данных, которые относятся к сведениям, позволяющие идентифицировать личность (Personally Identifiable Information, PII) [3]. Примерами таких данных могут служить ФИО, дата рождения, номера документов и т. д. Хранение и использование медицинских текстов с целью анализа без предварительного удаления PII повышает требования к безопасности информационной системе и создаёт риски утечки чувствительной информации. Перед использованием данных с целью обучения моделей искусственного интеллекта, эта информация должны быть удалена из документов. Однако из-за особенностей ведения электронных медицинских записей в различных МО и человеческого фактора, удаление этих данных может быть осложнено большим разнообразием как самих данных, так и структур их написания.

Задача поиска подобных данных в неструктурированных текстах и дальнейшего обезличивания (анонимизации) медицинских текстовых документов рассмотрена в данной статье.

Постановка задачи обезличивания

PII-атрибуты можно условно разделить на два типа: те, которые позволяют однозначную идентификацию пациента сами по себе и те, которые позволяют обеспечить однозначную идентификацию пациента по совокупности атрибутов.

К атрибутам, каждый из которых может однозначно идентифицировать пациента, относятся:

1. паспортные данные (серия, номер паспорта, кем и когда выдан);
2. серия и номер полиса ОМС;
3. СНИЛС (номер);
4. ИНН (номер);
5. телефонный номер (мобильный).

В качестве данных, позволяющих идентифицировать личность может выступать совокупность из следующих сведений:

- Ф.И.О.;
- дата и место рождения;
- адрес места регистрации;
- адрес проживания (реальный);
- гражданство;
- информация об образовании;
- место работы;
- должность;
- сведения о законных представителях, членах семьи, родственниках и т. д.;
- номер электронной медицинской карты (ЭМК);
- адрес электронной почты (не всегда позволяет идентифицировать персону).

В то же время часть информации, хранящейся в медицинских документах, не является PII, например:

1) название медицинской организации (МО);
2) адрес МО;
3) дата обращения в МО;
4) записи в ЭМК (врачебные осмотры, проведенное обследование, проведенное лечение, эпикризы, выписки);
5) Ф.И.О. сотрудников МО.

Такие сведения называют ICP (сведения, соотносимые с конкретной личностью — персоной, на основе которых невозможно однозначно определить их принадлежность конкретному лицу (персоне); то есть любой набор ICP-атрибутов является анонимным (anonymized data)).

Считается, что всегда существует хотя бы одна совокупность PII-атрибутов, обеспечивающая однозначную идентификацию пациента; заметим, что с точки зрения возможности идентификации пациента набор PII-атрибутов в записи может быть избыточным; некоторые рекомендации по составу атрибутов и алгоритмам идентификации пациентов приведены в ГОСТ ISO/TS22220. В таблице 1 приведены некоторые примеры совокупности атрибутов.

Таким образом, задачей обезличивания медицинских текстов является детекция и удаление персональных данных пациентов, с одновременным сохранением важной медицинской информации, необходимой для поставленных задач использования медицинских текстов.

Обзор исследований в области обезличивания медицинских данных

Для целей исследования был проведен поиск и анализ публикаций российских и зарубежных авторов,

Таблица 1

Совокупности PII-атрибутов

Список возможных совокупностей PII-атрибутов, обеспечивающих однозначную идентификацию пациента	Список возможных совокупностей PII-атрибутов, на основе которых невозможно однозначно идентифицировать пациента
1. Ф.И.О. +дата и место рождения	1. Ф.И.О. + информация об образовании
2. Ф.И.О. + адрес места регистрации	2. Ф.И.О. + место работы
3. Ф.И.О. + номер медицинской карты (ЭМК)	3. Ф.И.О. + адрес электронной почты
4. дата и место рождения + адрес места регистрации	4. адрес места регистрации + гражданство
5. Ф.И.О. + дата и место рождения+ адрес проживания (реальный)	5. адрес места регистрации + информация об образовании
6. Ф.И.О. + дата и место рождения+ гражданство	6. адрес места регистрации + адрес электронной почты
7. Ф.И.О. + дата и место рождения+ информация об образовании	7. адрес проживания (реальный)+ гражданство
8. Ф.И.О. + дата и место рождения+ место работы+ должность	8. адрес проживания (реальный)+ информация об образовании
9. Ф.И.О. + дата и место рождения+ сведения о законных представителях, членах семьи, родственниках и т. д.	9. адрес проживания (реальный)+ место работы
10. Ф.И.О. + дата и место рождения + номер медицинской карты (ЭМК)	10. адрес проживания (реальный) + сведения о законных представителях, членах семьи, родственниках и т. д.

посвященных обезличиванию данных в медицинских информационных системах. В первую очередь речь идёт об обезличивании структурированных данных. Под структурированными данными понимаются данные, записанные в определенные поля баз данных систем здравоохранения и содержащих персональную информацию определенных типов, таких как, фамилия, дата рождения, номер паспорта, СНИЛС, свидетельства ОМС и т. д. Для обезличивания подобных данных существуют различные методы, такие как анонимизация и псевдонимизация данных. Существуют международные стандарты, регламентирующие псевдонимизацию [4]. В РФ требования к обезличиванию данных для систем искусственного интеллекта в медицине описаны в ГОСТ Р 2023[6].

Необходимость разработки и стандартизации подходов, касающихся обезличивания текстовых данных из систем здравоохранения, подчеркивают и зарубежные исследователи [5]. В том числе обезличивать медицинские документы предлагается с целью привлечения новых групп исследователей к вопросам извлечения ценной информации из медицинских данных[7].

Так как текстовые данные, прошедшие процедуры обезличивания, в дальнейшем будут использованы для анализа, задача обезличивания рассматривается не только как отдельная задача обработки естественного языка, но и как метод предварительной обработки данных, который влияет на результат применения последующих методов обработки. Исследователи по итогам экспериментов над наборами данных на английском языке приходят к выводу, что не существует идеального способа обезличивания данных, оптимально подходящего под различные задачи дальнейшей обработки и позволяющему достигнуть полной анонимности данных, и что разработчики и исследователи в данной области должны будут подбирать методы под собственные задачи и специфику обрабатываемых данных, чтобы найти баланс между удалением потенциальных персональных данных и сохранением ценной информации [6].

Существуют различные программные решения, ориентированные на обезличивание медицинских текстовых данных, такие как, например, решение от John Snow Labs[8]. Это решение учитывает особенности текстовых данных из систем здравоохранения и имеет высокие показатели точности по всем извлекаемым сущностям (>0.9), и опережает качество, полученное с помощью ChatGPT[9] [7]. Однако в нём отсутствует поддержка русского языка.

Задача детекции персональных данных в неструктурированных медицинских текстах осложняется в том числе тем, что среди медицинских данных присутствует множество фамилий и имен собственных (болезнь Боткина, пучок Гиса, симптом Виленкина и т. д.), а запись персональных данных пациентов имеет большое разнообразие структур в зависимости от медицинской информационной системы или организации.

[6] Системы искусственного интеллекта в клинической медицине. Наборы данных в формате DICOM для тестирования алгоритмов. Методы деперсонализации набора данных и контроля набора данных на отсутствие персональных данных (1.11.164-1.273.23).

[7] Marciniak M., Mykowiecka A., Rychlik P. Medical text data anonymization. Journal of Medical Informatics and Technologies. 2010. № 16.

[8] URL: https://johnsnowlabs.com/

[9] URL: https://openai.com/chatgpt/

В связи с вышеперечисленными факторами, для решения этой задачи был разработан программный сервис обезличивания текстовых медицинских записей Webiomed.Anonymizer.

Предлагаемое решение

Сервис Webiomed.Anonymizer, включая детекцию персональных данных, был разработан с использованием технологии поиска именованных сущностей.

Именованная сущность (далее NER) — это элемент текста, который выделяет объекты с похожими характеристиками из множества других частей текста. Её называют жестким обозначением, атомарным элементом или членом семантического класса, который может варьироваться в зависимости от интересующей области[10]. Например, в области анализа медицинских текстов для фармакологических исследований объектами интереса являются названия лекарств, их дозировки и курсы приёма. В качестве NER-объектов персональных данных пациентов выступают ФИО, адрес проживания, дата рождения и т. д.

В текущей версии сервиса обезличивания поддерживается детекция и удаление следующих атрибутов персональных данных: серия и номер паспорта, СНИЛС, телефон, ИНН, полис ОМС, ФИО, номер медкарты, место работы (организация), должность, e-mail, гражданство, сведения об образовании, дата рождения, адрес.

Работа сервиса обезличивания состоит из нескольких этапов:

1. Валидация входных данных. Проверка соответствия структуры входного файла формата JSON ожидаемому.
2. Детекция персональных данных в текстах, получение списка найденных атрибутов персональных данных в виде NER-объектов. NER-объекты — это подстроки в тексте, имеющие координаты, и являющиеся описанием атрибута персональных данных, таких как ФИО, место работы, данные о документах и т. д.
3. Обработка результатов детекции: удаление или искажение атрибутов персональных данных в исходных текстах.
4. Формирование ответа сервиса, содержащего измененные тексты и метаданные.

Детекция NER-объектов персональных данных проводится в 3 этапа:

1. Поиск персональных данных с помощью правил, построенных на основе регулярных выражений. Используется для детекции хорошо структурированных типов данных, таких как номера документов, телефонов и т. д.
2. Поиск слабоструктурированных персональных данных с помощью специализированной NER модели.
3. Поиск дат рождения, используя правила, построенные на основе наиболее часто встречаемых сочетаний, и информацию о найденных персональных данных типа «ФИО», а также инструментов для поиска дат в тексте.

Схема работы сервиса представлена на рисунке 1.

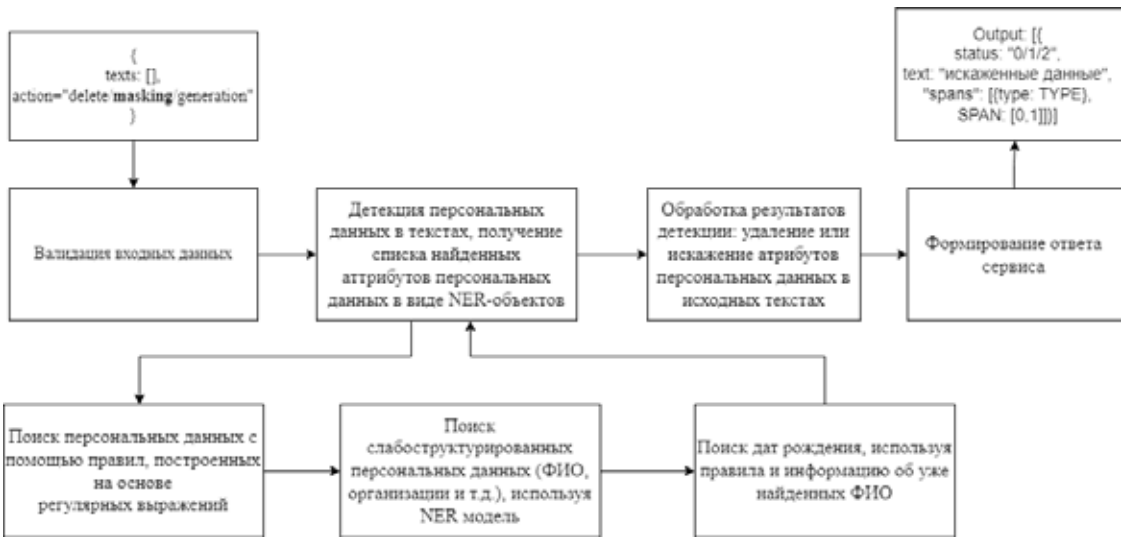

Рис.1. Схема работы сервиса обезличивания медицинских записей

[10] Nadeau D., Sekine S. A Survey of Named Entity Recognition and Classification. Lingvisticæ Investigationes. 2007. № 30. DOI: 10.1075/li.30.1.03nad.

В качестве входа сервис принимает JSON-объект, состоящий из массива текстов, которые необходимо обезличить, и указания, какие именно образом. В данный момент поддерживается два режима: удаление (замена символов на X) и маскировка (замена на теги, обозначающие тип атрибута). Выход сервиса содержит статус его работы (0 — успешное исполнение, ПДн не найдено, 1 — успешное исполнение, ПДн найдены и обработаны, 2 — в процессе выполнения произошла ошибка), обезличенные тексты, а также справочная информация о том, какие типы персональных данных были найдены с координатами в изначальном тексте.

На данном этапе было принято решение использовать готовую модель для выделения NER-объектов. Выбор модели для выделения именованных сущностей осуществлялся путем анализа открытых данных о работе различных решений на открытых наборах данных[11]. Исходя из комплексного анализа, для первой версии продукта была выбрана модель Slovnet[12]. Хотя некоторые показатели точности её работы отстают от Deepavlov-bert [9] в пределах 3%, модель Slovnet затрачивает меньше вычислительных ресурсов (205 Мб ОЗУ против 6144 Мб ОЗУ у Deepavlov-bert) и не требует использования GPU (графического процессора) для вычислений. Вычислительные затраты и скорость обработки критически важны, учитывая объемы данных, обрабатываемых платформой Webiomed (около 500 000 документов в сутки) и необходимость установки сервиса в защищенном контуре заказчика на ограниченных ресурсах. Архитектура программного комплекса была выстроена таким образом, что модель для выявления NER-сущностей может быть заменена на другую готовую модель с небольшими временными затратами.

В качестве альтернативы использованию NER-моделей, основанных на глубоком обучении, рассматривался словарный подход с использованием словарей ФИО русского языка[13]. В процессе предварительного тестирования на двух подходах ложноположительные срабатывания на медицинских данных были выявлены при использовании как обученных NER-моделей, так и словарей ФИО, примеры представлены в таблице 2.

NER-модель создаётся с использованием технологий глубокого обучения на предварительно размеченной выборке. Для защиты от ложноположительных срабатываний на медицинских данных используется ряд правил «достоверности» найденных атрибутов персональных данных, основанных на проверке исключений и использовании информации о ранее найденных атрибутах персональных данных в анализируемом тексте. В качестве примеров правил можно привезти то, что для удаления/маскировки подстроки, распознанной как «Организация», необходимо явное указание на отношение к месту работу/учебы пациента или к медицинскому учреждению. Ограничение было введено с целью того, что многие медицинские аббревиатуры, написанные заглавными буквами, распознаются NER-моделями как потенциальное указание на организацию. А при проверке подстрок, помеченных NER-моделью как «ФИО», происходит проверка на пересечение с медицинскими эпонимами — терминами, образованными от фамилий людей.

Таблица 2

Примеры ложноположительных срабатываний

Примеры ложноположительных срабатываний на категорию «ФИО» с использованием NLP-модели	Примеры ложноположительных срабатываний на категорию «Фамилия» с использованием словарей ФИО
Врач-уролог	Фактор
Врач-эндокринолог	Класс
Аутоиммунный	Москва
Клексан	Контроль
Н. Полиморфизмы	Последняя
Диализный	Число
Апгар	Неделя
	Девочка

Кроме того, для использования NER-модели, необходима модель векторного представления слов. В данном случае использовалась модель, обученная с помощью алгоритма GloVe[14]. Для создания подобной модели, первым шагом собирается статистика совпадения слов в формате матрицы совпадения слов X.

Каждый элемент X_{ij} такой матрицы представляет, как часто слово i появляется в контексте слова j. Анализ корпуса происходит следующим образом: для каждого термина мы ищем термины контекста в некоторой области, определяемой window_size перед термином и window_size после термина.

1. Определяются ограничения для каждой пары слов:

$$w_i^T w_j + b_i + b_j = log(X_{ij}) \quad (1)$$

где w_i - вектор главного слова, w_j - вектор контекстного слова, b_i, b_j - скалярные смещения для основного и контекстного слов.

2. Производительность функции будет вычисляться следующим образом:

$$J = \sum_{i=1}^{V} \sum_{j=1}^{V} f(X_{ij})(w_i^T w_j + b_i + b_j - \log X_{ij})^2 \quad (2)$$

Весовая функция, которая помогает нам предотвратить обучение только по чрезвычайно распространенным парам слов:

[11] URL: https://github.com/natasha/slovnet?tab=readme-ov-file#ner-1

[12] URL: https://github.com/natasha/slovnet

[13] URL: https://github.com/cybermatt/russian-names

[14] Pennington J., Socher R., Manning Ch. Glove: Global Vectors for Word Representation. EMNLP. 2014. № 14, pp. 1532–1543. DOI: 10.3115/v1/D14-1162 .

$$f(X_{ij}) = \begin{cases} (\frac{X_{ij}}{x_{max}})^{\alpha}, \text{если } X_{ij} < X_{max} \\ 1\ , \text{если } X_{ij} \geq X_{max} \end{cases} \quad (3)$$

После получения результатов обработки моделью, результаты проходят фильтрацию по подобранным в процессе разработки и валидации правилам.

Пример результата работы сервиса — в таблице 3.

Таблица 3

Пример работы сервиса Webiomed.Anonymizer

Текст на вход	Жалобы: Симоненков Андрей Юрьевич, дата рождения: 27.05.1978 пол: М код: 700000 СНИЛС: 123-456-789 91, 89123456789 Москва, улица ленина, д. 2? работает в ОООКомуслуги. Должность: слесарь. ИНН: 143517991154 Страховой полис: 1234400893001234 Выдан: 01.01.2015 Давление 120/80, кашель. Анамнез: туберкулез, болезнь Боткина отрицает
Текст на выходе	Жалобы: <PER>, дата рождения: <DATABIRTH> пол: М код: 705775 <SNILS>, <PHONE> <LOC>? <ORG> . Должность: <JOB>. <INN> <OMS> Давление 120/80, кашель. Анамнез: туберкулез, болезнь Боткина отрицает

При разработке Webiomed.Anonymizer были использованы следующие фреймворки и библиотеки для языка Python 3.9 с открытым исходным кодом: веб-фреймворк для создания API FastAPI, библиотека для моделирования языка на основе глубокого обучения для русского языка SlovNet (лицензия MIT License); библиотека предварительно обученных компактных моделей векторного представления слов для русского языка Navec[15]; Библиотека для токенизации русских предложений и слов Razdel,[16]; библиотека для преобразования типов данных Pydantic[17]; модуль для расширения стандартной библиотеки datetime, предназначенный в том числе для парсинга дат в строковом формате, Python-dateutil[18].

Методы экспериментального исследования

Для валидации качества работы сервиса обезличивания производится сравнение данных NER-объектов, полученных путём разметки в программе Medtator (https://medtator.ohnlp.org/), и данных, полученных в результате обработки текста сервисом.

Сравнение выполняется с помощью метрики F1-мера, адаптированной под задачи NER, которая оценивает совпадение извлеченных NER объектов из текста по двум параметрам:

1) Правильно ли модель идентифицировала тип NER объекта.
2) Найдены ли точные границы NER объекта в тексте.

Каждый текст может содержать от 0 до нескольких NER объектов. В каждом NER объекте есть поле type (TAG) и поле text с координатами.

Точность вычисляется по формуле:

$$Precision = \frac{correct}{actual} \quad (4)$$

где: *correct* — количество правильных прогнозов (совпадение NER объектов по обоим полям — type и text); *actual* — количество фактических прогнозов.

Полнота вычисляется по формуле:

$$Recall = \frac{correct}{possible} \quad (5)$$

где: *correct* — количество правильных прогнозов (совпадение NER объектов по обоим полям); *possible* — количество возможных прогнозов (количество NER объектов в разметке, умноженное на 2 — два поля).

F1-мера вычисляется по формуле:

$$F1\ score = \frac{2}{\left(\frac{1}{Precision} + \frac{1}{Recall}\right)} = \frac{2 \times Precision \times Recall}{(Precision + Recall)} \quad (6)$$

Выборка и разметка

Разметка проводилась экспертами с доменными знаниями в области медицины в программе Medtator. Схема разметки состояла из 14 тегов, некоторые из которых подразделялись на атрибуты, обозначающие принадлежность. Так, например, тег «ФИО» мог относится как к пациенту, так и к его лечащему врачу, или его представителю. Таким образом, в одном документе могли содержаться персональные данные сразу нескольких человек: врача, пациента, родственников пациента, чьи ФИО и контактные данные он указал.

Выборка для валидации собиралась, исходя из разнообразия регионов РФ, семантики текстов, а также наличия в них нескольких атрибутов персональных данных, исходя из предварительного анализа текстов. В итоге была создана выборка из 124 медицинских документов, которые были предварительно размечены

[15] URL: https://github.com/natasha/navec/, лицензия MIT License.

[16] URL: https://github.com/natasha/razdel/, лицензия MIT License.

[17] URL: https://docs.pydantic.dev/, лицензия MIT License.

[18] URL: https://pypi.org/project/python-dateutil/, лицензия на программное обеспечение Apache, лицензия BSD.

Рис. 2. Распределение частоты встречаемости атрибутов персональных данных в документах выборки

экспертами, а затем результат обработки этих документов через API был сопоставлен с результатами разметки. Распределение частоты встречаемости атрибутов персональных данных представлена на рисунке 2. Частота встречаемости рассчитывалась путем определения доли документов, в которых присутствовал минимум один атрибут заданного типа, от общего числа проанализированных документов.

По результатам валидации были получены метрики, представленные в таблице 4.

Таблица 4

Метрики работы сервиса

Признак	F1-мера	Precision	Recall
Паспорт	0,95	0,95	0,95
СНИЛС	0,98	1	0,96
Телефон	0,92	0,89	0,95
ИНН	1	1	1
Полис ОМС	0,95	0,93	0,98
Номер медкарты	0,78	0,93	0,67
Место работы (Организация)	0,41	0,46	0,36
ФИО	0,81	0,85	0,77
Должность	0,61	0,63	0,59
E-mail	0,9	0,93	0,88
Гражданство	0,83	0,83	0,83
Сведения об образовании	0,82	0,88	0,77
Дата рождения	0,76	0,64	0,95
Адрес	0,49	0,54	0,44

В рамках оценки производительности сервиса Webiomed.Anonymizer были проведены измерения скорости обработки запросов в однопоточном режиме. Этот тест позволил оценить базовую эффективность системы без учета параллельной обработки. Полученные результаты, отражающие время выполнения различных операций, систематизированы и представлены в табл. 5. Эти данные служат важным ориентиром для оптимизации работы сервиса и планирования его масштабирования в будущем.

Таблица 5

Скорость работы разработанного сервиса

Запросов (шт)	Текстов (шт)	Символов в одном тексте	Время в секундах
1	1	10	0,05
1	1	1000	0,12
1	100	10	0,18
1	100	1000	1,53
1	100	10000	16
100	1	10000	25,7

Выводы

Для дальнейшего развития и внедрения технологий ИИ в здравоохранении необходимо накапливать, обрабатывать и анализировать большие массивы данных, в том числе текстовых. В неструктурированных текстовых медицинских записях могут встречаться атрибуты персональных данных пациентов, поэтому перед их загрузкой в системы искусственного интеллекта необходимо произвести удаление из текстов этих атрибутов, которое не снизит общее качество и информативность записей для дальнейшего анализа.

В то же время, модели, разработанные для решения «общих» задач анализа и генерации естественного языка, показывают более низкое качество на медицинских текстах, включая большее количество ложноположительных срабатываний. Исследования, касающиеся анонимизации медицинских записей на английском языке, показывают, что сервисы, разработанные для медицинских данных, выигрывают в качестве по сравнению с использованием общедоменных больших языковых моделей.

Для решения этой задачи в условиях российского здравоохранения был разработан программный комплекс для обезличивания медицинских текстовых данных. Преимуществами разработанного комплекса являются: использование комбинированного подхода, который позволяет обнаруживать как структурированные, так и слабоструктурированные данные; использование предобученных моделей, таких как Slovnet, позволяет сэкономить время на разработку. Гибкая архитектура позволяет легко заменять модели. В результате экспериментальной апробации программного комплекса на наборе медицинских документов, была показана высокая (>0,92) точность работы с атрибутами, позволяющими однозначную идентификацию пациента и более низкую (0,41—0,9) для работы с атрибутами, позволяющими идентификацию пациента в совокупности. Этой точности будет недостаточно, чтобы гарантировать полное обезличивание текстов из медицинских документов, но эта точность позволяет выявлять шаблоны и документы, в которых чаще всего попадает персональная информация, и в дальнейшем предотвращать её появление на уровне МИС. Дальнейшими направлениями работы над решением задачи являются: расширение списка поддерживаемых атрибутов персональных данных, применение методов активного обучения для постоянного улучшения модели, калибровка по результатам дальнейшего внедрения с целью повышения качества детекции и удаления персональных данных.

Литература

1. Гусев А.В., Шарова Д.Е. Этические проблемы развития технологий искусственного интеллекта в здравоохранении // Общественное здоровье. 2023. № 3 (1). С. 42—50. DOI: 10.21045/2782-1676-2023-3-1-42-50 .
2. A question of trust for AI research in medicine. Nature Machine Intelligence. 2024. No. 6 (739). URL: https://doi.org/10.1038/s42256-024-00880-0
3. Kulkarni P., Cauvery K. Personally Identifiable Information (PII) Detection in the Unstructured Large Text Corpus Using Natural Language Processing and Unsupervised Learning Technique. International Journal of Advanced Computer Science and Applications. 2021. No. 12(9). DOI: 10.14569/IJACSA.2021.0120957 .
4. Столбов А.П. Обезличивание персональных данных в здравоохранении // Врач и информационные технологии. 2017. № 3. С. 76—91.
5. Li X., Qin J. Anonymizing and Sharing Medical Text Records. Inf Syst Res. 2017, 28(2), pp. 332-352. DOI: 10.1287/isre.2016.0676 .
6. Larbi I., Burchardt A., Roller R. Clinical Text Anonymization, Its Influence on Downstream NLP Tasks and the Risk of Re-Identification. Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop. 2023, pp.105–111. DOI: 10.18653/v1/2023.eacl-srw.11 .

7. Kocaman V., Hasham U.H., Talby D. Beyond Accuracy: Automated De-Identification of Large Real-World Clinical Text Datasets. 2023. DOI: 10.48550/arXiv.2312.08495 .
8. Goyal A., Gupta V., Kumar M. Recent Named Entity Recognition and Classification Techniques: A Systematic Review. Computer Science Review. 2018. No. 29. Pp. 21–43. DOI: 10.1016/j.cosrev.2018.06.001 .
9. Kuratov Y., Arkhipov M. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. 2019. DOI: 10.48550/arXiv.1905.07213.

SECTION:
INFORMATION AND ELECTRONIC TECHNOLOGIES IN THE LEGAL SPHERE

DEPERSONALISATION OF MEDICAL TEXT DATA WITH A VIEW TO DEVELOPING AND IMPLEMENTING ARTIFICIAL INTELLIGENCE SYSTEMS

Elena Makarova, *Ph.D. (Technology), OOO (PLC) K-SkAI, Petrozavodsk, Russian Federation. ORCID: 0000-0002-5410-5890.*
E-mail: emakarova@webiomed.ru

Anna Andreichenko, *Ph.D. (Physics & Mathematics), OOO (PLC) K-SkAI, Petrozavodsk, Russian Federation. ORCID: 0000-0001-6359-0763.*
E-mail: aandreychenko@webiomed.ru

Mariia Kazakova, *Ph.D. (Medicine), OOO (PLC) Retail, Kazan, Russian Federation. ORCID: 0000-0002-8669-3383.*
E-mail: mak26091986@gmail.com

Dmitrii Ivanov, *OOO (PLC) K-SkAI, Petrozavodsk, Russian Federation. ORCID: 0009-0001-4079-2642.*
E-mail: divanov@webiomed.ru

Aleksandr Gusev, *Ph.D. (Technology), Central Research Institute for Organisation and Informatisation in Healthcare of the Ministry of Health of the Russian Federation, Moscow, Russian Federation. ORCID: 0000-0002-7380-8460.*
E-mail: agusev@webiomed.ai

Keywords: *personal data detection, natural language processing, text data processing, search for named entities, medical information systems, big data, language models.*

Abstract

Purpose of the paper: developing and implementing technologies for depersonalising medical text data for developing artificial intelligence (AI) systems for healthcare, which is one of the digital transformation priorities.

Methods used in the study: review and comparison of research sources on medical data depersonalisation and search for named entities, validation of the results of processing of an expert tagged data set by a software system.

Study findings: problems are described of collecting and processing large volumes of medical data necessary both for training AI models and performing research based on real clinical practice. The task of depersonalising data got from medical information systems is put forward. Personal data of electronic medical records, such as full name, date of birth, document numbers, etc., are described, as well as the difficulties of depersonalising such records, since ready-made solutions for search for named entities are not suitable for processing medical data without additional training of models or additional processing of the result. The architecture and key software components of the developed Webiomed.Anonymizer service supporting the identifying and removal of personal information from medical text records are described.

Practical value: examples are given of processing of different data by the service. Experimental testing of the developed algorithms was carried out by validating a sample of medical text records with personal data from different regions of the Russian Federation. Metrics of the attributes accuracy detection for personal data used by the service, such as full name, date of birth, documents, employment information, etc., obtained from the sample are given.

References

1. Gusev A.V., Sharova D.E. Eticheskie problemy razvitiia tekhnologii iskusstvennogo intellekta v zdravookhranenii. Obshchestvennoe zdorov'e. 2023. No. 3 (1). Pp. 42–50. DOI: 10.21045/2782-1676-2023-3-1-42-50 .
2. A question of trust for AI research in medicine. Nature Machine Intelligence. 2024. No. 6 (739). URL: https://doi.org/10.1038/s42256-024-00880-0
3. Kulkarni P., Cauvery K. Personally Identifiable Information (PII) Detection in the Unstructured Large Text Corpus Using Natural Language Processing and Unsupervised Learning Technique. International Journal of Advanced Computer Science and Applications. 2021. No. 12(9). DOI: 10.14569/IJACSA.2021.0120957 .
4. Stolbov A.P. Obezlichivanie personal'nykh dannykh v zdravookhranenii. Vrach i informatsionnye tekhnologii. 2017. No. 3. Pp. 76–91.
5. Li X., Qin J. Anonymizing and Sharing Medical Text Records. Inf Syst Res. 2017, 28(2), pp. 332-352. DOI: 10.1287/isre.2016.0676 .
6. Larbi I., Burchardt A., Roller R. Clinical Text Anonymization, Its Influence on Downstream NLP Tasks and the Risk of Re-Identification. Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop. 2023, pp.105–111. DOI: 10.18653/v1/2023.eacl-srw.11 .
7. Kocaman V., Hasham U.H., Talby D. Beyond Accuracy: Automated De-Identification of Large Real-World Clinical Text Datasets. 2023. DOI: 10.48550/arXiv.2312.08495 .
8. Goyal A., Gupta V., Kumar M. Recent Named Entity Recognition and Classification Techniques: A Systematic Review. Computer Science Review. 2018. No. 29. Pp. 21–43. DOI: 10.1016/j.cosrev.2018.06.001 .
9. Kuratov Y., Arkhipov M. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. 2019. DOI: 10.48550/arXiv.1905.07213 .