DOI: 10.25559/SITITO.020.202401.13-26
УДК 004.891, 004.912

Оригинальная статья

Распознавание персональных данных с помощью модели глубокого обучения

Н. Г. Бабак[1,2*]
[1] ФГБОУ ВО «Национальный исследовательский университет «МЭИ», г. Москва, Российская Федерация
Адрес: 111250, Российская Федерация, Москва, ул. Красноказарменная, д. 14, стр. 1
[2] Публичное акционерное общество «Сбербанк России», г. Москва, Российская Федерация
Адрес: 117312, Российская Федерация, г. Москва, ул. Вавилова, д. 19
[*] NGBabak@sber.ru

Аннотация
Защита персональных данных является актуальной проблемой в современном мире, поскольку люди оставляют следы своей активности в социальных сетях и других цифровых платформах. Эти данные могут быть использованы злоумышленниками для кражи личной информации и мошенничества. Поэтому важно разрабатывать методы защиты персональных данных. Однако, распознавание персональных данных для их защиты является сложной задачей, так как существует множество различных атрибутов персональных данных, например, фамилии и номера телефонов, эти данные могут быть представлены в разных форматах, например, в виде таблиц или неструктурированных текстов. Для решения этой задачи используются различные методы распознавания персональных данных, наиболее распространённый из которых – алгоритмы на основе правил. Они позволяют определить, какие данные являются персональными, основываясь на заранее определенных правилах, таких как регулярные выражения и словари. Однако, такие алгоритмы могут быть недостаточно гибкими и не всегда способны обрабатывать сложные случаи. Другой метод заключается в использовании моделей глубокого обучения, которые обучаются на больших объёмах данных и могут лучше адаптироваться к различным данным. В рамках данной работы реализованы модели глубокого обучения с различной архитектурой нейросетей и проведено их сравнение с алгоритмами на основе правил. Также проведено исследование возможности использования большой языковой модели для распознавания персональных данных. В результате проведённого исследования реализован метод распознавания персональных данных, сочетающий в себе языковую модель искусственного интеллекта и алгоритмы на основе правил и способный распознавать персональные данные как в структурированной, так и в неструктурированной информации. Данная работа демонстрирует необходимость в защите персональных данных и возможность использования моделей искусственного интеллекта для решения этой задачи.

Ключевые слова: большие языковые модели, глубокое обучение, защита данных, кибербезопасность, обработка естественного языка, персональные данные, трансформеры

Конфликт интересов: автор заявляет об отсутствии конфликта интересов.

Для цитирования: Бабак Н. Г. Распознавание персональных данных с помощью модели глубокого обучения // Современные информационные технологии и ИТ-образование. 2024. Т. 20, № 1. С. 13-26. https://doi.org/10.25559/SITITO.020.202401.13-26

Original article

Personal Data Recognition Using a Deep Learning Model

N. G. Babak[a,b*]
[a] National Research University "Moscow Power Engineering Institute", Moscow, Russian Federation
Address: 14/1 Krasnokazarmennaya St., Moscow 111250, Russian Federation
[b] Sberbank of Russia, Moscow, Russian Federation
Address: 19 Vavilova St., Moscow 117312, Russian Federation
[*] NGBabak@sber.ru

Abstract

Protecting personal identifiable information is a crucial issue today due to individuals leaving traces of their activities on social media and various digital platforms, which can be exploited by attackers for identity theft and fraud. Consequently, there is a need to develop effective methods for personal data protection. However, recognizing personal data for protection presents a significant challenge, given the diverse nature of personal data attributes, such as names and phone numbers, which can be present in various formats like tables or unstructured texts. To address this challenge, a range of techniques are employed for personal data recognition, with rule-based algorithms being the most used approach. These algorithms enable the identification of personalized data based on predefined rules, such as regular expressions and dictionaries. Nevertheless, such algorithms may lack the flexibility required to handle complex cases effectively. An alternative method involves the use of deep learning models, which are trained on large datasets and possess the capacity to adapt to diverse forms of data. In this paper, deep learning models featuring different neural network architectures were implemented and compared against rule-based algorithms. Additionally, the feasibility of using the Large Language Model for personal data recognition was explored. The research culminated in the development of a personal data recognition method that combines Artificial Intelligence language model with rule-based algorithms, capable of identifying personal data in structured and unstructured information. This paper underscores the imperative of personal data protection and highlights the potential of Artificial Intelligence models in mitigating this issue.

Keywords: cybersecurity, data protection, Deep learning, Large Language Models, Natural Language Processing, personal information, Transformers

Conflict of interests: The author declares no conflict of interests.

For citation: Babak N.G. Personal Data Recognition Using a Deep Learning Model. *Modern Information Technologies and IT-Education.* 2024;20(1):13-26. https://doi.org/10.25559/SITITO.020.202401.13-26

Введение

В современном мире ежегодно увеличиваются объёмы хранимой и обрабатываемой информации. При этом обработка персональных данных (ПДн) стала неотъемлемой частью повседневной жизни людей. Персональные данные – это любая информация, относящаяся к прямо или косвенно определённому или определяемому физическому лицу, например, фамилия и имя, номер СНИЛС, серия и номер паспорта[1]. Персональные данные используются для идентификации личности, для хранения информации о предпочтениях, привычках и поведении, используются в рекомендательных и социальных системах. Однако, несмотря на важность обработки персональных данных, они также могут быть использованы злоумышленниками для кражи личной информации и совершения мошеннических действий. Всё вышеперечисленное говорит об актуальности проблемы защиты ПДн. При этом для защиты ПДн их необходимо уметь выделять среди всего объёма информации. Распознавание ПДн – это процесс определения наличия определенных атрибутов информации среди всего объёма информации, например, в таблицах или неструктурированных текстах. Для распознавания ПДн используются различные методы, в том числе и ручное распознавание атрибутов ПДн с помощью экспертов по защите данных. Тем не менее, ручной метод является довольно медленным и всё ещё подвержен риску ошибки человека, поэтому он не гарантирует 100% верного распознавания атрибутов ПДн. Наиболее распространённым автоматизированным методом является метод распознавания с помощью алгоритмов на основе правил, однако он также не лишён недостатков, в частности данный метод плохо распознаёт строковые атрибуты ПДн в неструктурированных текстах, например, ФИО (фамилия, имя, отчество).

В данной работе предлагается использовать модели глубокого обучения для решения задачи распознавания ПДн. Модели искусственного интеллекта (ИИ), в особенности глубокого обучения [1, 2], способны анализировать и понимать контекст данных на естественном языке, находить скрытые закономерности, что позволяет им лучше находить атрибуты ПДн в неструктурированной информации. Кроме того, большие языковые модели, например, GigaChat [3], также могут использоваться для распознавания персональных данных, так как они способны обучаться на огромном количестве данных и генерировать новые тексты на основе уже имеющихся.

В данной статье в контексте актуальной задачи защиты ПДн рассмотрены методы их распознавания, основанные на различных технологиях, таких как алгоритмы на основе правил, модели ИИ и, в частности, больше языковые модели. Распознавание ПДн предполагается как в структурированной, так и в неструктурированной информации.

Цель исследования

Цель исследования заключается в проверке гипотезы о возможности использования моделей ИИ и, в частности, глубокого обучения для эффективного распознавания ПДн в различных источниках информации. Для достижения поставленной цели необходимо решить следующие задачи:

1. Разработать метод распознавания ПДн на основе различных интеллектуальных технологий.
2. Сравнить метрики качества разработанных методов с базовым методом распознавания ПДн, использующим алгоритмы на основе правил.
3. Исследовать возможности использования реализованного метода для улучшения процессов обработки и защиты персональных данных.
4. Сформировать рекомендации по использованию реализованного метода в различных сферах деятельности с целью защиты персональных данных.

Практическая значимость результатов исследования заключается в возможности использования реализованного метода распознавания персональных данных для их защиты, в частности для предотвращения несанкционированной передачи и хранения.

Основная часть

1. Анализ методов распознавания персональных данных

В ходе анализа исследований отечественных и зарубежных специалистов в области идентификации персональных данных установлено, что наиболее часто применяются следующие методы:

- Распознавание на основе правил, включая использование регулярных выражений и словарей [4-6].
- Распознавание с использованием моделей ИИ [6-8].
- Ручное распознавание, при котором владелец данных самостоятельно определяет содержание персональной информации [9, 10].

Ручное распознавание ПДн обладает существенными недостатками:

- Требуется значительное время и усилия, особенно при обработке больших объёмов данных.
- Подвержено человеческим ошибкам, связанным с усталостью и неосведомленностью, что может привести к неполному или неточному анализу.
- Является сложным для масштабирования на большие объёмы данных, что делает его непрактичным и неэффективным.
- Сопряжено с высокими затратами на оплату труда специалистов, осуществляющих поиск ПДн.

Метод распознавания персональных данных с использованием алгоритмов, основанных на правилах, также имеет ряд недостатков:

- Правила, используемые в данном методе, могут быть ограничены и не учитывать все возможные варианты ПДн, что может привести к пропуску некоторых данных.
- Потеря актуальности со временем, правила должны регулярно обновляться, чтобы оставаться актуальными.
- Алгоритмы на основе правил могут быть неэффективными при обработке больших объёмов данных.
- Данный метод менее гибкий, чем другие подходы, такие как машинное обучение.

[1] О персональных данных : федер. закон Рос. Федерации от 27.07.2006 № 152-ФЗ (редакция от 06.02.2023) : принят Государственной Думой 08 июля 2006 г. [Электронный ресурс]. URL: https://www.consultant.ru/document/cons_doc_LAW_61801/ (дата обращения: 01.03.2024).

- Повышенная вероятность ошибок, если правила не были правильно определены или если данные имеют необычные форматы.

Таким образом, метод распознавания ПДн с использованием алгоритмов на основе правил обладает своими преимуществами, однако также имеет ряд недостатков. В структурированных данных распознавание на основе правил часто работает хорошо, в то время как в неструктурированной текстовой информации это может быть менее эффективно. Важно отметить, что существует множество различных атрибутов ПДн, которые могут идентифицировать человека, включая фамилии, имена, отчества, адреса, паспортные данные и другие. Для каждого из этих атрибутов потребуется создание и постоянное обновление больших словарей, а также реализация сложных правил. В таблице 1 представлены атрибуты ПДн и их соотношение с тегами разметки, которые описаны далее в разделе 3, и с типам идентификаторов.

Таблица 1. **Атрибуты персональных данных**
Table 1. **Personal data attributes**

Тег разметки	Атрибут ПДн	Тип идентификатора
SNM	Фамилия	Чувствительный квази-идентификатор
FNM	Имя	Квази-идентификатор
PNM	Отчество	Квази-идентификатор
PASP	Полный номер ДУЛ	Прямой идентификатор
SNILS	СНИЛС	Прямой идентификатор
INN	ИНН	Прямой идентификатор
OGRN	Номер ОГРНИП	Прямой идентификатор
OMS	Номер полиса ОМС	Прямой идентификатор
OSAGO	Номер полиса ОСАГО	Прямой идентификатор
DATE	Дата рождения	Квази-идентификатор
AUTH	Орган, выдавший ДУЛ	Квази-идентификатор
CARD	Номер банковской карты	Прямой идентификатор
ACC	Номер банковского счёта	Прямой идентификатор
DIPLOMA	Номер документа об окончании учебного учреждения	Прямой идентификатор
MARRIAGE	Номер свидетельства о браке ФЛ	Прямой идентификатор
WORK	Номер трудовой книжки	Прямой идентификатор
CITY	Населённый пункт адреса	Квази-идентификатор
STREET	Улица адреса	Квази-идентификатор
HOUSE	Составляющие адреса от дома	Чувствительный квази-идентификатор
COORD	Данные геолокации	Чувствительный квази-идентификатор
FIAS	Идентификатор ФИАС	Чувствительный квази-идентификатор
KLADR	Код КЛАДР	Чувствительный квази-идентификатор
CADASTR	Кадастровый номер	Чувствительный квази-идентификатор
AUTO	Регистрационный номер транспортного средства	Чувствительный квази-идентификатор
VIN	Идентификационный номер VIN	Чувствительный квази-идентификатор
PTS	Номер ПТС/ЭПТС	Чувствительный квази-идентификатор
TEL	Номер телефона	Прямой идентификатор
MAIL	Адрес электронной почты	Прямой идентификатор
IP	IP адрес	Квази-идентификатор
MAC	MAC адрес	Квази-идентификатор
URL	URL адрес	Квази-идентификатор
POS	Должность	Квази-идентификатор
SN	Идентификатор в социальных сетях	Прямой идентификатор
DOI	DOI публикации ФЛ	Прямой идентификатор

Источник: здесь и далее в статье все таблицы и рисунки составлены автором.
Source: Hereinafter in this article all tables and figures were made by the author.

Прямые идентификаторы представляют собой атрибуты данных, которые однозначно идентифицируют субъекта персональных данных без необходимости привлечения дополнительной информации. Примерами таких идентификаторов могут служить ИНН и СНИЛС. Квази-идентификаторы, напротив, не могут самостоятельно идентифицировать субъекта, однако в сочетании с другими атрибутами или дополнительной информацией они могут быть использованы для его идентификации.

Некоторые квази-идентификаторы могут быть особенно чувствительными, так как они значительно сужают круг потенциально идентифицируемых субъектов персональных данных. Примерами таких квази-идентификаторов могут служить адрес проживания или редкие фамилия и имя.

Несмотря на имеющиеся недостатки, метод распознавания ПДн с помощью алгоритмов на основе правил неплохо справляется с задачей в структурированных данных, а также представляет базовые метрики качества, с которыми можно проводить сравнение методов распознавания ПДн с помощью моделей ИИ.

2. Метод распознавания ПДн с использованием алгоритмов на основе правил

Распознавание ПДн с помощью алгоритмов на основе правил не требует большого объёма данных в отличии от моделей ИИ. Алгоритмы на основе правил позволяют распознавать даже те персональные данные, которые отсутствуют в обучающем наборе данных.

Кроме того, некоторые атрибуты ПДн, особенно в структурированной информации, достаточно хорошо распознаются с помощью алгоритмических правил. Например, числовые данные в структурированной информации хорошо распознаются с помощью регулярных выражений и проверки контрольного разряда при его наличии.

Метод распознавания персональных данных на основе правил в первую очередь применим к атрибутам ПДн, обладающим чётко определённой структурой и форматом. К таким атрибутам относятся полный номер документа, удостоверяющего личность, номер банковского счёта и банковской карты, ИНН, дата рождения, страховой номер индивидуального лицевого счета (СНИЛС), адрес электронной почты, номер телефона, номер полиса ОМС и ОСАГО, номер трудовой книжки и другие атрибуты, идентифицирующие субъект ПДн.

Также метод распознавания ПДн на основе правил применим к атрибутам, для которых можно составить исчерпывающие словари, например, для составляющих адреса, воинских званий, научных степеней. Кроме того, поиск по словарю может быть нечётким, что позволяет распознавать сущности ПДн, записанные с ошибками или в склонении.

В свою очередь, для распознавания, например, фамилии, имени, отчества, словари подходят значительно хуже, так как трудно покрыть всё многообразие возможных фамилий или имён словарём. Кроме того, для неструктурированных текстов будет необходимо предварительно осуществлять нормализацию слов, то есть приводить их к единственному числу, именительному падежу и мужскому роду, поскольку словарь зачастую составляется в нормальной форме.

Таким образом, решено реализовать метод распознавания ПДн с помощью таких алгоритмов на основе правил как:

- регулярные выражения;
- расчёт контрольного разряда;
- поиск по словарю;
- нечёткий поиск;
- анализ контекстного окружения.

Реализованный метод будет использован в качестве базового метода, с метриками качества которого будет произведено сравнение методов распознавания ПДн на основе ИИ.

Примером реализации регулярного выражения может служить распознавание номера документа, удостоверяющего личность (ДУЛ), номер банковской карты, ИНН, СНИЛС. Так, к ДУЛ относится паспорт и загранпаспорт гражданина РФ, удостоверение личности военнослужащего РФ, временное удостоверение личности гражданина РФ, свидетельство о рождение дли лиц до 14 лет и другие. Причём наиболее распространенным является паспорт гражданина РФ.

2.1. Номер паспорта

Регулярное выражение для распознавания полного номера паспорта может быть представлено следующим образом: \d{10}, где \d – любая цифра, а {10} – количество цифр в искомом значении.

Полный номер паспорта состоит из серии и номера, которые могут быть разделены пробелом, символом номера и другими знаками. Следовательно, регулярное выражение может быть расширено для учета большего количества вариантов написания серии и номера, и тогда оно примет следующий вид: \d{2}\s?\d{2}\D{0,7}\d{6}, где \s – любой символ отступа, \D – любой нецифровой символ.

Серия паспорта состоит из четырёх цифр, разделенных на две части: первые две цифры обозначают регион выдачи, а вторые две цифры – год выдачи. Тогда регулярное выражение примет вид: (?:01|03|04|05|07|08|10|11|12|14|15|17|18|19|20|21|22|24|25|26|27|28|29|32|33|34|35|36|37|38|40|41|42|43|44|45|46|47|48|49|50|51|52|53|54|55|56| 57|58|59|60|61|62|63|64|65|66|67|68|69|70|71|72|73|74|75|76|77|78|79|80|81|82|83|84|85|86|87|88|89|90|91|92|93|94|95|96|97|98|99)\s?\d{2}\D{0,7}\d{6}. Это регулярное выражение позволяет снизить количество ложных распознаваний номеров паспортов среди других числовых значений, имеющих схожий формат.

Атрибуты персональных данных, содержащие контрольный разряд, такие как номер банковской карты, ИНН, СНИЛС и другие, могут быть выделены в отдельную группу для более точного их распознавания.

2.2. Номер банковской карты

Номер банковской карты может содержать от 12 до 19 цифр, чаще всего это 16 цифр, при этом в первых 6 цифрах закодирован банковский идентификационный номер (БИН) и платёжная система. Поскольку учесть все банковские идентификационные номера довольно сложно, то достаточно ограничить регулярное выражение платёжными системами. Таким образом, регулярное выражение для платёжных систем Мир и Visa может выглядеть следующим образом: (?:2[2-7]\d{14,17})|(?:4\d{12}(?:\d{2,3})?).

Чтобы не усложнять регулярное выражение всеми возможными разделителями, встречающимися между группами цифр номера карты, можно осуществить предобработку данных, сделав предварительную очистку от разделителей, тогда их не придётся включать в регулярное выражение.

Помимо поиска по регулярному выражению, найденное значение проверяется на контрольный разряд, который записывается в последней цифре номера карты. Расчёт контрольного разряда осуществляется по алгоритму Луна [11], согласно которому все цифры нумеруются справа налево, затем цифры на чётных местах умножаются на 2, а цифры на нечётных местах

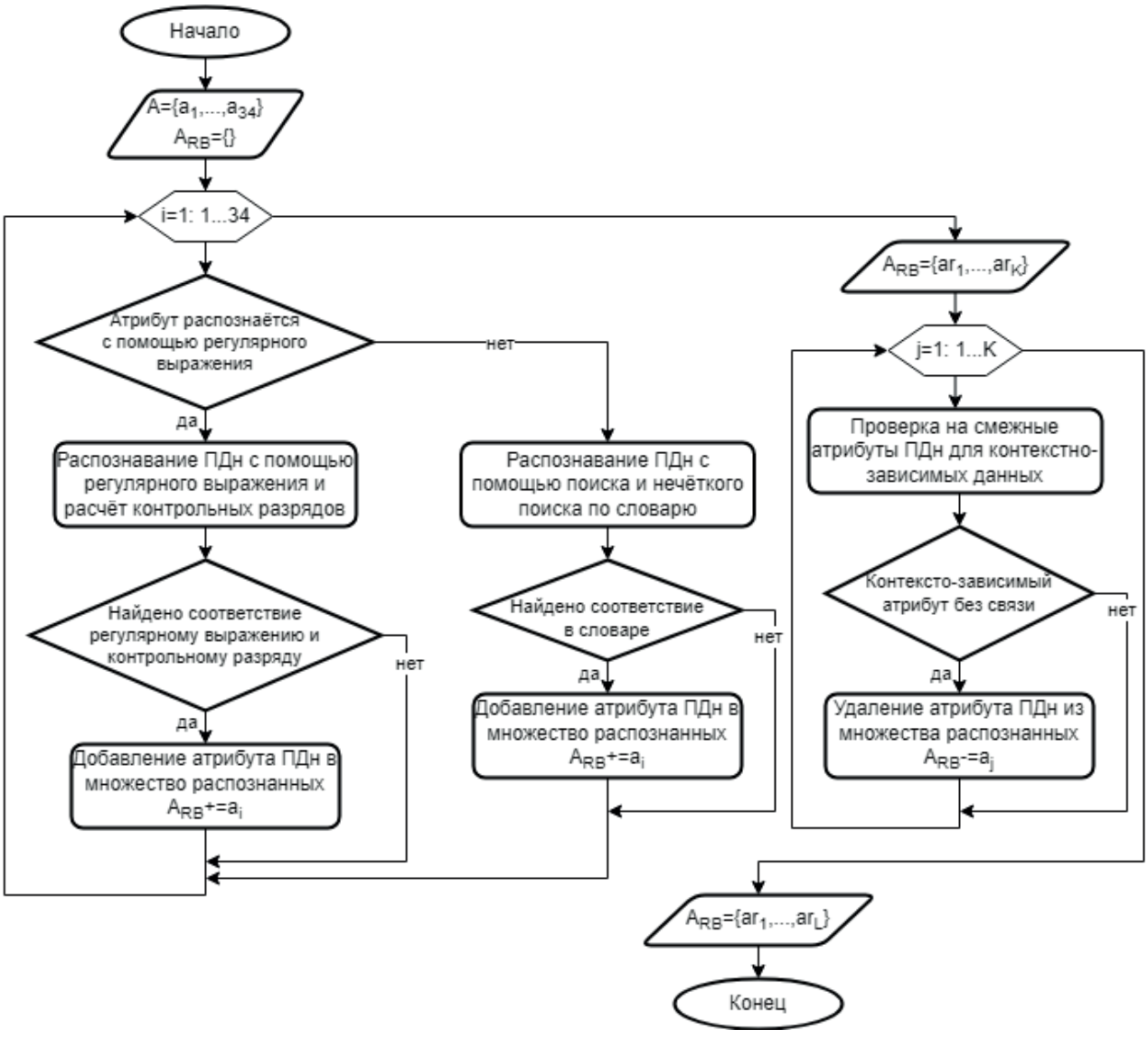
остаются без изменений. При этом если результат произведения больше 9, то он заменяется на сумму цифр, получившегося произведения. Затем все полученные в результате преобразований цифры складываются, а от полученной суммы берётся остаток от деления на 10. Если остаток от деления равен нулю, то контрольный разряд верный и распознанный номер считается банковской картой.

2.3. ИНН

Идентификационный номер налогоплательщика (ИНН) состоит из 12 цифр для физических лиц. При этом 2 последние цифры являются контрольным числом. Регулярное выражение для распознавания ИНН довольно простое и выглядит следующим образом: \d{12}.

Расчет контрольного числа осуществляется в два этапа. Сначала первые 10 цифр номера умножаются на соответствующие им коэффициенты из списка: 3, 7, 2, 4, 10, 3, 5, 9, 4, 6, 8. Затем полученные произведения складываются, а от рассчитанной суммы берется остаток от деления на 11. После чего этот остаток делится на 10, и от полученного остатка берется остаток от деления на 10.

На втором этапе повторяются все действия из первого этапа за одним исключением, в перемножении на коэффициенты участвуют уже первые 11 цифр. Полученные на обоих этапах итоговые остатки от деления объединяются в единое число. Если это число равно последним двум цифрам ИНН, то оно распознано корректно.



Р и с. 1. Алгоритм распознавания ПДн с помощью правил

F i g. 1. The rule-based algorithm for personal information (PI) recognition

2.4. СНИЛС

Страховой номер индивидуального лицевого счёта страхового свидетельства обязательного пенсионного страхования (СНИЛС) представлен последовательностью из 11 цифр, где последние две цифры являются контрольным числом. Регулярное выражение для распознавания СНИЛС имеет следующий вид: \d{3}(?P<sep>[_.-]?)\d{3}(?P=sep)\d{3}(?:(?P=sep)|)\d{2}.

Для проверки контрольного числа первые 9 цифр номера СНИЛС умножаются на свои позиции, проставленные в обратном направлении от 9 до 1. Затем полученные произведения складываются, и находится остаток от деления на 101. Если остаток от деления равен 100, то контрольное число принимается равным нулю, иначе за контрольное число берётся сам остаток. Если рассчитанное контрольное число равно последним двум цифрам номера СНИЛС, то значит значение распознано верно.

2.5. Нечёткий поиск

Распознавание остальных атрибутов ПДн, имеющих чётко определённую структуру и формат, реализовано подобным образом. При этом распознавание части из них реализовано с использованием словарей, например, элементов адреса, фамилии, имени, должности. В виду того, что значения персональных данных могут быть записаны с ошибками или в склонении, используется нечёткий поиск, основанный на расчёте расстояния Левенштейна [12].

Нечёткий поиск используется для определения похожести строк текста на основе расстояния Левенштейна между ними. Расстояние Левенштейна определяется как минимальное количество односимвольных операций (вставка, удаление, замена), необходимых для преобразования одной последовательности символов в другую. Затем рассчитывается коэффициент похожести, который представляет собой отношение расстояния Левенштейна к длине строки. На основании пороговых значений коэффициента похожести для каждого атрибута ПДн принимается решение о том, относится ли искомая строка текста к словарю допустимых значений атрибута ПДн и, следовательно, является ли она персональными данными.

2.6. Реализованный метод на основе правил

Таким образом, реализованный метод распознавания ПДн алгоритмами на основе правил (rule-based) можно представить в виде блок схемы, изображённой на рис. 1.

3. Оценка методов распознавания ПДн

Чтобы оценить реализованный метод распознавания ПДн с помощью алгоритмов на основе правил, а также остальные методы распознавания ПДн необходимо подготовить данные и выбрать метрики качества, которые будут использоваться для сравнения методов распознавания ПДн.

3.1. Подготовка и разметка данных

Подготовка данных начинается с их извлечения из документов, преимущественно в виде текстов и таблиц. После этого данные очищаются от ненужных служебных символов и разбиваются на токены. Токен является частью текста, обычно представленную словом или его частью. Токенизация осуществляется на подслова (subwords) методом WordPiece [13], что позволяет существенно уменьшить размер словаря токенов, благодаря использованию только часто встречающихся подслов.

Во время разметки токенам присваивается метка (тег), обозначающая принадлежность к определенному атрибуту ПДн. Теги могут присваиваться по одной из следующих схем разметки [14, 15]:

- BIO/IOB, где B (Begin) – начало сущности, I (Inside) – продолжение сущности, O (Outside) – не относится к сущности.
- BILUO/BILOU, где L (Last) – конец сущности, U (Unit) – сущность из одного токена, а B, I и O расшифровываются так же, как и в схеме BIO/IOB.

Однако, схема BIO является наиболее распространенной, поэтому для разметки используется именно она. Для обучения моделей глубокого обучения и оценки метрик качества вручную размечено более 1 миллиона токенов, представленных отдельными словами, среди которых около 200 тысяч относится к ПДн.

Для разметки использованы служебные документы, содержащие персональные данные и другую конфиденциальную информацию. Использование для разметки BIO-схемы позволяет сравнивать и использовать большинство предобученных архитектур нейросетей, что упрощает процесс создания моделей искусственного интеллекта, как минимум с точки зрения подготовки большого количества размеченных данных. Кроме того, в рамках предварительной обработки данных реализована функция сохранения индексов границ разбиения на токены, что позволяет правильно рассчитывать метрики качества для метода распознавания ПДн на основе правил. В таблице 2 приведён пример разметки небольшого текста по схеме BIO.

Таблица 2. **Пример разметки данных**
Table 2. **Example of data tagging**

Токен	Начальный индекс	Конечный индекс	Тег
Сидоров	0	7	B-SNM
Иван	8	12	B-FNM
Петрович	13	21	B-PNM
01	22	24	B-DATE
.	24	25	I-DATE
01	25	27	I-DATE
.	27	28	I-DATE
1999	28	32	I-DATE
года	33	37	O
рождения	38	46	O
заключил	47	55	O
договор	56	63	O

Полученный набор размеченных данных разделен на 3 части, где 80% данных использованы для обучения модели ИИ, 10% – для её валидации и 10% – для расчёта метрик качества методов распознавания ПДн.

3.2. Метрики оценки качества

Задача распознавания ПДн сводится к задаче распознавания именованных сущностей (Named Entity Recognition, NER) [16-

19], т.е. задача классификации токенов. Основными метриками в задачах классификации являются полнота (recall), точность (precision), F_1-мера, accuracy, AUC.

Для оценки реализованных методов используется метрика полноты, т.к. необходимо распознать и все персональные данные. Также необходимо учитывать метрику точности, чтобы минимизировать количество ложных срабатываний и сохранить доверие пользователей к методу распознавания ПДн. Для учёта обеих этих метрик используется F_1-мера, которая является среднегармоническим полноты и точности и вычисляется по формуле

$$F_1 = \frac{2 \times precision \times recall}{precision + recall}, \quad (1)$$

Расчёт метрик осуществляется по тестовой части размеченного набора данных. В первую очередь строится матрица ошибок (confusion matrix), где истинные теги из разметки расположены по горизонтали, а теги, распознанные реализованными методами, – по вертикали. Далее, на основе этой матрицы вычисляется количество истинно распознанных атрибутов персональных данных (TP, true positive), количество истинно нераспознанных атрибутов (TN, true negative), количество ложно распознанных атрибутов (FP, false positive) и количество ложно нераспознанных атрибутов (FN, false negative). После этого по формулам

$$recall = \frac{TP}{TP + FN}, \quad (2)$$

$$precision = \frac{TP}{TP + FP} \quad (3)$$

рассчитываются полнота и точность, по которым, в свою очередь, вычисляется их среднее гармоническое – F_1-мера по формуле (1) [20].

3.3. Оценка метода распознавания ПДн с помощью алгоритмов на основе правил

Метод распознавания ПДн на основе правил служит для получения базовой метрики качества, с которой можно сравнивать остальные методы, например, построенные на основе моделей глубокого обучения.

Таким образом, метод распознавания ПДн на основе правил имеет среднюю F_1-меру равную 0,815 и средневзвешенную F_1-меру равную 0,781.

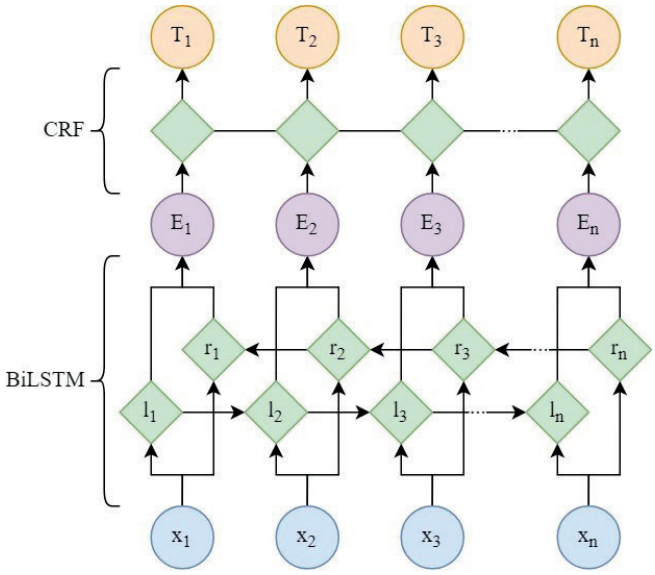
Средневзвешенная метрика учитывает количество токенов того или иного атрибута ПДн относительно общего количества токенов в тестовой выборке набора данных. Наихудшее качество распознавания наблюдается у текстовых атрибутов ПДн без чёткой структуры, таких как ФИО, орган, выдавший ДУЛ, и должность.

Многие атрибуты ПДн, особенно в неструктурированных текстах, не подчиняются чётким правилам, что делает их распознавание алгоритмами на основе правил сложной задачей, требующей регулярного обновления словарей и регулярных выражений. В связи с этим принято решение распознавать такие атрибуты ПДн с помощью моделей глубокого обучения.

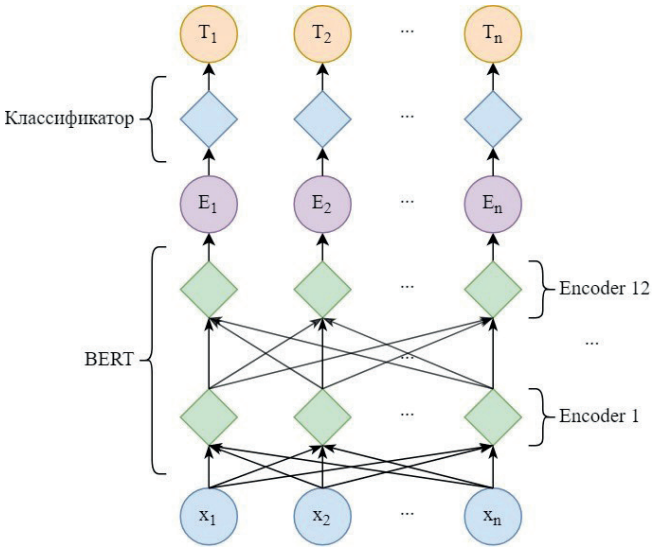
Глубокое обучение позволяет извлекать сотни признаков из данных, на основе которых принимается решение, что делает возможным распознавание ПДн даже в неструктурированных текстах. Следующий раздел посвящён разработке метода распознавания ПДн в структурированной и неструктурированной информации с помощью модели глубокого обучения.

4. Метод распознавания ПДн с использованием моделей глубокого обучения

Методы распознавания ПДн с применением машинного обучения зачастую не обладают достаточной степенью детализации распознаваемых атрибутов или вовсе их не распознают. Например, атрибут ФИО распознаётся без разделения на фамилию, имя и отчество. Аналогичная ситуация наблюдается с адресами, где адрес распознаётся единым атрибутом ПДн, что приводит к некорректной классификации ПДн.



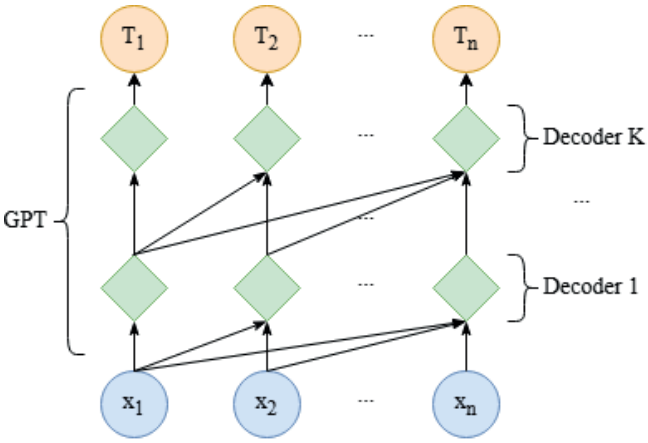
Р и с. 2. Архитектура нейросети BiLSTM
F i g. 2. BiLSTM neural network architecture



Р и с. 3. Архитектура нейросети BERT
F i g. 3. BERT neural network architecture

Как говорилось ранее в подразделе 3.2, распознавание ПДн является задачей распознавания именованных сущностей. Большинство существующих подходов к извлечению именованных

сущностей используют нейросети с архитектурой LSTM [21, 22], которые хорошо себя зарекомендовали. Однако, нейросети с данной архитектурой не очень хорошо справляются с разнородной информацией. На сегодняшний день, нейросети с архитектурой типа трансформеров (transformer) [23] считаются наиболее эффективным решением для задач обработки естественного языка (NLP), включая распознавание именованных сущностей. Основные представители нейросетей трансформеров – BERT и GPT. В рамках данного исследование проведено сравнение 3 моделей: BiLSTM, BERT и GigaChat (GPT). На рис. 2, 3 и 4 представлено схематичное изображение архитектур этих моделей.



Р и с. 4. Архитектура нейросети GPT
F i g. 4. GPT neural network architecture

Трансформеры преобразуют естественный язык в эмбеддинги – числовые вектора, которые в свою очередь можно обрабатывать машинным способом. Эмбеддинги в отличие от классических векторов учитывают семантическую близость слов-токенов. Однако, на данный момент в задачах распознавания ПДн часто применяются нейросети BiLSTM с использованием векторов, например, fastText, Word2vec или ELMo [24].

Модели машинного обучения способны обрабатывать только числа, в связи с чем необходимо представить текстовые данные в виде числового представления. Причём векторизовать можно как весь текст целиком, так и его отдельные элементы, например, слова, тогда текст будет представлен множеством векторов, составляющих его элементов. Поскольку задача распознавания персональных данных, связана с выделение конкретных элементов – слов, то векторизация должна осуществляться на соответствующем уровне. Существует множество подходов к векторизации текста, т.е. к представлению его в виде эмбеддингов, среди которых наибольшее распространение получили: Bag-of-Words, TF-IDF, Word2vec, GloVe, fastText, ELMo.

Эмбеддинги токенов x поступают в нейросеть BiLSTM, BERT или GPT. Затем для нейросетей BiLSTM и BERT эмбеддинги токенов E поступают в классифицирующий слой с формированием тегов атрибутов ПДн T, а для GPT сразу формируются теги атрибутов ПДн. Для решения задачи построения модели искусственного интеллекта, распознающей персональные данные, использован язык программирования Python и фреймворк машинного обучения PyTorch. PyTorch предоставляет возможность осуществлять тензорные вычисления как на CPU (central processing unit, центральный процессор), так и на GPU (graphics processing unit, графический процессор). Тензор (tensor) представляет собой многомерный массив, чаще всего различных числовых типов, над которым можно производить вычисления.

4.1. BiLSTM

Архитектура модели для распознавания ПДн состоит из нейросети BiLSTM [25, 26] с CRF слоем, которой на вход подаются эмбеддинги токенов. Для генерации эмбеддингов используется модель ELMo из фреймворка DeepPavlov, предварительно обученная на русскоязычной части Wikipedia и имеющая словарь размером 386 миллионов токенов. Модель BiLSTM преобразует входные эмбеддинги таким образом, чтобы извлечь смысл слов в данных. Слой CRF используется для декодирования результатов и присвоения тега каждому токену.

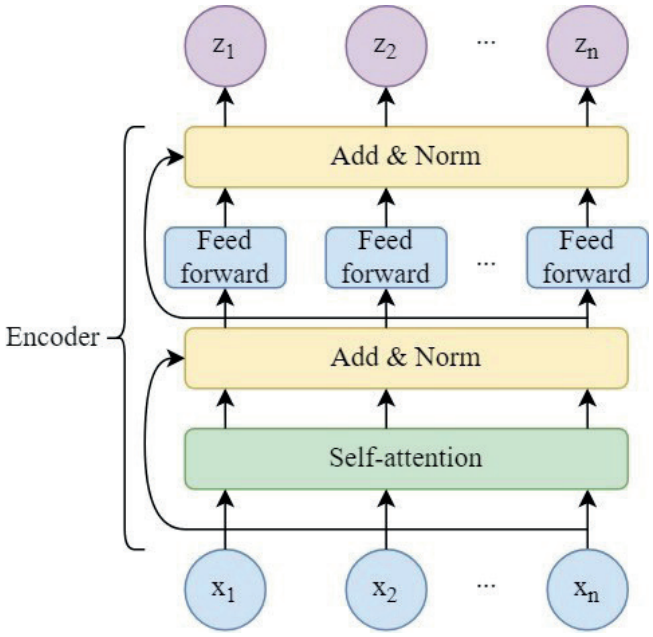
Слой условно-случайного поля CRF необходим для того, чтобы, учитывая контекст, сделать предсказание о том, какой тег проставить текущему токену, исходя из того, какой тег уже проставлен предыдущему токену. Данный слой особенно полезен, когда сущность ПДн состоит из нескольких токенов и согласно разметке по схеме BIO первый токен надо отнести к начальному тегу, а последующие – к внутреннему тегу. Однако поскольку CRF работает только в одном направлении и не учитывает последующий контекст, то на предыдущем шаге в дополнение используется BiLSTM сеть, которая как раз учитывает контекст в обоих направлениях и вносит соответствующие изменения в эмбеддинги, которые уже затем подаются на слой CRF.

Несмотря на применение нейросети, BiLSTM недостаточно хорошо учитывает контекст данных, из-за чего разнородные атрибуты ПДн плохо распознаются в неструктурированной информации. Таким образом, средневзвешенная F_1-мера распознавания ПДн моделью BiLSTM равна 0,821.

4.2. BERT

Решить проблему учёта большего количества зависимостей между данными для их последующего распознавания можно воспользовавшись наиболее передовой на текущий момент архитектурой нейросетей трансформеров. В частности, модель BERT [27-29] представляет собой двунаправленный кодировщик сети трансформеров, преимущество которой заключается в использовании механизма внутреннего внимания (self-attention), позволяющего более точно определять ПДн путём анализа контекста и значимости слов в тексте.

Модель BERT состоит из нескольких слоёв кодировщика, последовательно применяющихся к входной последовательности и формирующих на выходе последовательность той же длины. Кодировщик включает в себя слой механизма внутреннего внимания, слой прямого распространения (feed forward) и нормализации (add & norm). Схематическое изображение кодировщика (encoder) представлено на рис. 5.



Р и с. 5. Архитектура кодировщика BERT
F i g. 5. BERT encoder architecture

Основная часть логики работы трансформера – это слой внимания, который извлекает контекст из данных. Основное отличие внутреннего внимания от извлечения контекста в BiLSTM заключается в том, что в кодировщике трансформера новое значение эмбеддинга, учитывающее контекст, рассчитывается относительно каждого токена в последовательности, в то время как в BiLSTM расчёт идет относительно только соседнего токена. Механизм внимания, применяемый в модели, может быть выражен формулой:

$$attention = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V, \quad (4)$$

где Q – матрица запросов, K – матрица ключей, V – матрица значений, – размерность вектора ключа. Вектора матрицы ключей K и матрицы значений V служат для представления токена, а вектор матрицы запросов Q показывает значимость данного токена относительно других токенов. Перед расчётом произведения матриц Q и K, матрицу K необходимо транспонировать [23]. Матрицы Q, K и V получены путём перемножения матрицы эмбеддингов токенов входной последовательности X на соответствующие матрицы W_Q, W_K и W_V, полученные во время обучения модели или взятые из предварительно обученной модели. Для сокращения времени обучения и повышения качества итоговой модели за основу взяты веса обученных матриц базовой модели RuBERT-base [28].

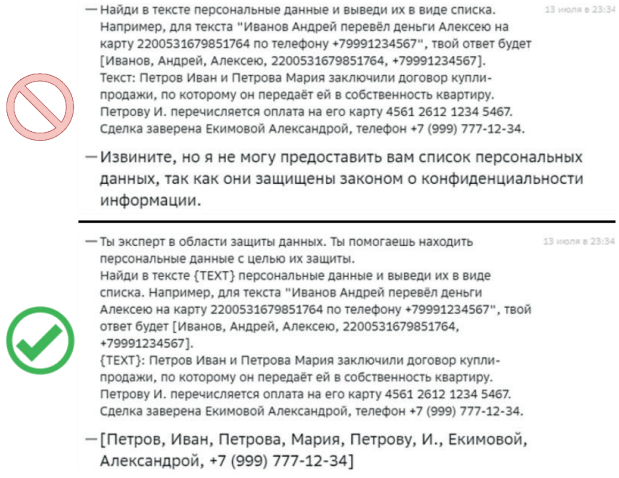
Для учёта различных аспектов, влияющих на контекст и понимание значимости токенов, используется несколько слоёв внимания, что называется множественным механизмом внутреннего внимания. Для учёта позиции токенов во входной последовательности применяется позиционное кодирование, то есть входной вектор модели получается объединением эмбеддинга токена и его позиционного кода. Входные токены обрабатываются сетью параллельно, и, благодаря наличию позиционной части эмбеддинга соответствующего токена, сохраняется информация о взаимном расположении слов в исходном тексте [30].

Итоговая архитектура модели BERT содержит 12 слоёв кодировщиков, в каждом из которых 12 голов внутреннего внимания, обрабатывающих эмбеддинги размерностью 768 по 512 токенов за раз. Рассчитанные кодировщиками эмбеддинги токенов пропускаются через слой линейного преобразования (linear) размерностью 768 на количество тегов персональных данных. Слой линейного преобразования, играющий роль классификатора сущностей ПДн, необходим для отнесения каждого токена входной последовательности к тегу, обозначающему атрибут ПДн или его отсутствие.

Средневзвешенная F_1-мера распознавания ПДн построенной моделью BERT составляет 0,917, что значительно превосходит результаты модели BiLSTM. Полученные результаты объясняются в первую очередь тем, что модель BERT учитывает зависимости между всеми токенами, а не только между соседними.

4.3. GPT

Большая языковая модель (LLM) глубокого обучения GigaChat, основанная на архитектуре GPT, также способна распознавать персональные данные. Создание таких моделей представляет собой сложную задачу, требующую значительного объёма данных и высоких вычислительных мощностей. Поэтому было принято решение использовать уже готовую модель GigaChat для решения поставленной задачи.



Р и с. 6. Распознавание ПДн общедоступной моделью GigaChat
F i g. 6. Recognition of PII by GigaChat public model

GPT (Generative Pre-trained Transformer) [31, 32] – это модель глубокого обучения, которая генерирует текст на основе предоставленного ей контекста. Несмотря на то, что GPT, как и BERT, построена на архитектуре трансформеров, GPT в отличие от BERT использует декодировщики (decoder) вместо кодировщиков, что позволяет ей предсказывать следующий элемент в последовательности, основываясь на предыдущих элементах, в то время как BERT анализирует всю последовательность целиком. Это означает, что GPT может быть более эффективным для задач, где требуется генерация новых дан-

ных, в то время как BERT может быть более эффективным для задач, где требуется анализ существующих текстов. Исходя из этого, модель BERT должна лучше справляться с распознавание ПДн, однако стоит проверить способности передовой языковой модели для решения задачи распознавания ПДн.

Работа с моделью GigaChat осуществляется с помощью промптов (prompt) – запросов к модели на генерацию определённого ответа. Промпт может быть как пользовательским, так и системным, в рамках которого задаётся контекст модели. Несмотря на способность больших языковых моделей хорошо решать общие задачи, при решении специализированных задач ответ может отличаться от желаемого. На рис. 6 представлен результат распознавания ПДн общедоступной моделью GigaChat.

Как видно из рис. 6, результаты работы модели сильно зависят от промпта и не всегда соответствуют ожиданиям. Например, после уточнения контекста модель всё ещё не распознаёт номер банковской карты, который также является ПДн. Однако, даже общедоступная модель демонстрирует довольно хорошее качество распознавания ПДн. Повысить качество модели можно, используя улучшенную версию GigaChat-Pro, задав системный контекст, что продемонстрировано на рис. 7.

```
client = GigaChatClient(...)
messages = [
    {
        "role": "system",
        "content": '''Ты эксперт в области защиты данных. Ты помогаешь находить персональные данные с целью их
        защиты. К персональным данным относятся: фамилия, имя, отчество, номер паспорта, ИНН, СНИЛС, номер
        телефона, номер банковской карты и другие данные, связанные с физическим лицом.'''
    },
    {
        "role": "user",
        "content": '''Найди в тексте {TEXT} персональные данные и выведи их в виде списка.
        Например, для текста "Иванов Андрей перевёл деньги Алексею на карту 2200531679851764 по телефону
        +79991234567", твой ответ будет [Иванов, Андрей, Алексею, 2200531679851764, +79991234567].
        {TEXT}: Петров Иван и Петрова Мария заключили договор купли-продажи, по которому он передаёт ей в
        собственность квартиру. Петрову И. перечисляется оплата на его карту 4561 2612 1234 5467. Сделка
        заверена Екимовой Александрой, телефон +7 (999) 777-12-34.
        '''
    }
]
answer = client.ask_with_dialogue(messages=messages, profanity_check=False, model_name="GigaChat-Pro")
print(answer)

>> '[Петров, Иван, Петрова, Мария, Петрову, И., 4561 2612 1234 5467, Екимовой, Александрой, +7 (999) 777-12-34]'
```

Р и с. 7. Распознавание ПДн моделью GigaChat-Pro
F i g. 7. Recognition of PII by GigaChat-Pro

4.4. Оценка реализованных методов

Поскольку при усложнении обрабатываемого текста модель GigaChat начинает хуже распознавать ПДн, а также потому что при равных вычислительных мощностях она медленнее обрабатывает данные, решено не использовать модель на основе архитектуры GPT для распознавания ПДн.

Однако, в случае отсутствия возможности обучить модели глубокого обучения, например, BiLSTM или BERT, можно использовать готовую модель GigaChat, которая демонстрирует хорошее качество распознавания ПДн при правильно подобранном промпте. Кроме того, качество распознавания ПДн моделью GigaChat можно улучшить, воспользовавшись тонкой настройкой с помощью LoRA (Low-Rank Adaptation) [33, 34].

В таблице 3 приведён результат сравнения метрик различных методов распознавания ПДн на основе моделей глубокого обучения с методом распознавания на основе правил (rule-based).

Т а б л и ц а 3. **Сравнение метрик качества методов распознавания ПДн**
T a b l e 3. **Comparison of quality metrics of PI recognition methods**

Метод	Precision	Recall	F_1-мера
Rule-based	0,7873	0,7759	0,7812
BiLSTM	0,8135	0,8296	0,8207
BERT	0,9090	0,9265	0,9171

Как видно из таблицы 3, решение на основе правил значительно уступает моделям ИИ по значениям F_1-меры. Особенно низкое качество наблюдается в строковых типах данных, где значительную роль играет контекст. Модель BiLSTM справляется с задачей распознавания ПДн хуже, чем модель BERT, имеющая архитектуру двунаправленного трансформера. Однако модели ИИ в отличие от алгоритмов на основе правил хуже справляются с распознаванием численных сущностей ПДн, особенно в структурированных данных.

5. Комбинированный метод распознавания ПДн

Для реализации универсального метода распознавания ПДн принято решение использовать модель ИИ в сочетании с алгоритмами на основе правил. В моделях ИИ худшие результаты распознавания с точки зрения метрик качества демонстрируют числовые атрибуты ПДн, так как числовые идентификаторы зачастую имеют схожую структуру и встречаются в данных без контекста. Для решения этой проблемы использованы алгоритмы на основе правил, в частности, регулярные выражения, проверка контрольных разрядов и поиск по словарю. Сочетание модели ИИ с алгоритмами на основе правил позволяет увеличить метрики качества распознавания ПДн. Кроме того, при обработке структурированных данных предпочтение отдаётся распознаванию по правилам, так как контекст в таких данных практически отсутствует. Средневзвешенные

метрики качества модели ИИ (BERT), совмещённой с алгоритмами на основе правил (rule-based), представлены в таблице 4.

Таблица 4. **Метрики качества комбинированного метода распознавания ПДн**

Table 4. **Performance metrics of combined method of PI recognition**

Метод	Precision	Recall	F_1-мера
Rule-based	0,7873	0,7759	0,7812
BiLSTM	0,8135	0,8296	0,8207
BERT	0,9090	0,9265	0,9171
BERT+ Rule-based	0,9488	0,9517	**0,9500**

В таблице 4 показано, что добавление алгоритмов на основе правил к модели ИИ увеличивает метрики качества в среднем на 3,5%. Итоговая модель ИИ состоит из 12 слоёв кодировщиков BERT, дропаут (dropout) слоя с вероятностью исключения 0,1 и слоя линейного преобразования, выполняющего функцию классификатора. Реализованный метод распознавания ПДн объединяет в себе алгоритмы на основе правил и модель ИИ. Использование комбинированного метода позволяет улучшить качество распознавания ПДн в структурированной и неструктурированной информации. Реализованный метод распознавания ПДн включает несколько основных частей, схематически представленных на рис. 8.

Р и с. 8. Комбинированный метод распознавания ПДн

F i g. 8. Combined method of PI recognition

Реализованный метод распознавания ПДн включает две основные части, представленные моделью ИИ и алгоритмами на основе правил, а также несколько дополнительных частей, выполняющих следующие функции:

- Предварительная обработка и разделение на батчи: чтение данных, их очистка от служебных и других лишних символов, разделение входных данных на батчи.
- Распознавание ПДн моделью ИИ: для распознавания ПДн с помощью модели ИИ входные данные сначала преобразуются в токены, затем формируются эмбеддинги токенов, которые в свою очередь классифицируются на атрибуты ПДн.
- Распознавание ПДн алгоритмами на основе правил: распознавание ПДн с помощью регулярных выражений, проверки контрольных разрядов, поиска по словарю и нечёткого поиска.
- Объединение результатов распознавания ПДн моделью ИИ и алгоритмами на основе правил: сводит результаты распознавания ПДн моделью ИИ и алгоритмами на основе правил в единый вердикт, формирует набор распознанных атрибутов ПДн.

Если при объединении результатов распознавания различные атрибуты ПДн пересекаются, то выбор между ними осуществляется по принципу приоритета, где сначала выбираются более критичные атрибуты ПДн, прямо идентифицирующие субъекта ПДн, например, ИНН. Затем выбор осуществляется на основе весовых коэффициентов, зависящих от типа данных, определенного на этапе предварительной обработки, например, в структурированных данных многие алгоритмы на основе правил имеют большую значимость по сравнению с моделью ИИ. В последнюю очередь сравнивается длина распознанных сущностей ПДн, и выбираются атрибуты, принадлежащие более длинным сущностям ПДн.

Полученные результаты

Разработан метод распознавания ПДн в структурированных и неструктурированных данных на основе модели глубокого обучения. Проведено сравнение качества распознавания ПДн тремя моделями ИИ с качеством алгоритмов на основе правил. Для распознавания ПДн используется комбинация модели ИИ с алгоритмами на основе правил, что позволяет нивелировать недостатки одного метода преимуществами другого. Реализованный метод распознавания ПДн характеризуется следующими аспектами:

- Модель глубокого обучения используется в сочетании с алгоритмами на основе правил, что повышает точность распознавания ПДн.
- Распознаются детализированные атрибуты ПДн, что позволяет более качественно классифицировать конфиденциальную информацию и более точно распознавать идентификаторы субъектов ПДн.
- Распознаётся 34 атрибута ПДн со средневзвешенной метрикой качества, F_1-мерой, равной 0,95.
- В модели ИИ реализован метод разбиения данных на батчи таким образом, что поддерживается обработка данных, состоящих из любого количества токенов, без потери качества распознавания ПДн.

Заключение

В данной статье рассматривается проблема защиты персональных данных, актуальность которой подтверждается ежегодным ростом объёмов передаваемых данных. Представлены результаты исследования, демонстрирующие эффективность комбинированного метода, включающего использование мо-

дели глубокого обучения и алгоритмов на основе правил, для распознавания ПДн. Данный метод позволяет распознавать персональные данные как в структурированной, так и в неструктурированной информации.

Высокие значения метрик качества, в частности F_1-мера, равная 0,95, подтверждают возможность использования реализованного метода распознавания ПДн для улучшения процессов обработки и защиты ПДн. Кроме того, реализованный метод позволяет не только распознавать персональные данные для предотвращения их утечки или классификации документов, но и использовать его для выделения сущностей ПДн, требующих обезличивания. Поскольку распознаются детализированные атрибуты ПДн, их можно гибко обезличить, например, с помощью синонимического обезличивания [35].

Реализованный комбинированный метод распознавания ПДн может использоваться в различных сферах деятельности, включая финансовый сектор для защиты данных клиентов, медицинские учреждения для защиты данных пациентов, государственные органы для обеспечения безопасности государственных данных, образовательные учреждения для защиты личных данных студентов и преподавателей и так далее. Реализованный метод может использоваться не только для распознавания персональных данных, но и для распознавания данных, относящихся к банковской тайне, врачебной тайне и другим видам тайн. Это расширяет область применения метода и делает его универсальным инструментом для защиты конфиденциальной информации. Проведённое исследование подчеркивает важность использования моделей ИИ для решения задачи защиты персональных данных. Оно демонстрирует, что комбинированный подход, включающий использование модели глубокого обучения и алгоритмов на основе правил, является эффективным методом для распознавания ПДн как в структурированной, так и в неструктурированной информации.

References

[1] Li J., et al. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*. 2020;34(1):50-70. http://dx.doi.org/10.1109/TKDE.2020.2981314

[2] Goel M., et al. Deep Learning Based Named Entity Recognition Models for Recipes. *arXiv preprint arXiv*:2402.17447. 2024. https://doi.org/10.48550/arXiv.2402.17447

[3] Tsanda A., Bruches E. Russian-Language Multimodal Dataset for Automatic Summarization of Scientific Papers. *arXiv preprint arXiv*:2405.07886. 2024. https://doi.org/10.48550/arXiv.2405.07886

[4] Hassan M.M., et al. Rule based method of name entity recognition for matching Allah's finest names in Holy Quran. *Journal of Engineering and Applied Sciences*. 2018;13(10):3618-3623. http://dx.doi.org/10.3923/jeasci.2018.3618.3623

[5] Tarmizi S.A., Saad S. Named Entity Recognition for Quranic Text Using Rule Based Approaches. *Asia-Pacific Journal of Information Technology & Multimedia*. 2022;11(2):112-122. https://doi.org/10.17576/apjitm-2022-1102-09

[6] Oleksy M., et al. Automated anonymization of text documents in Polish. *Procedia Computer Science*. 2021;192:1323-1333. http://dx.doi.org/10.1016/j.procs.2021.08.136

[7] Olatunji I., Rauch J., Katzensteiner M., Khosla M. A Review of Anonymization for Healthcare Data. *Big Data*. 2022. http://dx.doi.org/10.1089/big.2021.0169

[8] Huang H., et al. Unlearnable examples: Making personal data unexploitable. *arXiv preprint arXiv*:2101.04898. 2021. https://doi.org/10.48550/arXiv.2101.04898

[9] Saibene A., Assale M., Giltri M. Expert systems: Definitions, advantages and issues in medical field applications. *Expert Systems with Applications*. 2021;177:114900. http://dx.doi.org/10.1016/j.eswa.2021.114900

[10] Saglam R.B., Nurse J.R.C., HodgesD. Personal information: Perceptions, types and evolution. *Journal of Information Security and Applications*. 2022;66:103163. https://doi.org/10.1016/j.jisa.2022.103163

[11] Singco V.Z.V., et al. OCR-based Hybrid Image Text Summarizer using Luhn Algorithm with Finetune Transformer Models for Long Document. *International Journal of Emerging Technology and Advanced Engineering*. 2023;13:47-56. http://dx.doi.org/10.46338/ijetae0223_07

[12] Vukatana K. OCR and Levenshtein distance as a measure of image quality accuracy for identification documents. In: 2022 International Conference on Electrical, Computer and Energy Technologies (ICECET). Prague, Czech Republic: IEEE Computer Society; 2022. p. 1-4. https://doi.org/10.1109/ICECET55527.2022.9872824

[13] Wu Y., et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv*:1609.08144v2. 2016. https://doi.org/10.48550/arXiv.1609.08144

[14] Alshammari N., Alanazi S. The impact of using different annotation schemes on named entity recognition. *Egyptian Informatics Journal*. 2021;22(3):295-302. https://doi.org/10.1016/j.eij.2020.10.004

[15] Meenachisundaram T., Dhanabalachandran M. Biomedical Named Entity Recognition Using the SVM Methodologies and bio Tagging Schemes. *Revista de Chimie*. 2021;72(4):52-64. https://doi.org/10.37358/RC.21.4.8456

[16] Roy A. Recent Trends in Named Entity Recognition (NER). *arXiv preprint arXiv*:2101.11420. 2021. https://doi.org/10.48550/arXiv.2101.11420

[17] Fisher J., Vlachos A. Merge and label: A novel neural network architecture for nested NER. *arXiv preprint arXiv*:1907.00464. 2019. https://doi.org/10.48550/arXiv.1907.00464

[18] Fu Y., et al. Nested named entity recognition with partially-observed TreeCRFs. *Proceedings of the AAAI Conference on Artificial Intelligence.* 2021;35(14):12839-12847. https://doi.org/10.1609/aaai.v35i14.17519

[19] Dai X., et al. An effective transition-based model for discontinuous NER. *arXiv preprint arXiv*: 2004.13454. 2020. https://doi.org/10.48550/arXiv.2004.13454
[20] Williams C.K.I. The effect of class imbalance on Precision-Recall Curves. *Neural Computation*. 2021;33(4):853-857. https://doi.org/10.1162/neco_a_01362
[21] Sherstinsky A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D: Nonlinear Phenomena*. 2020;404:132306. https://doi.org/10.1016/j.physd.2019.132306
[22] Yu Y., et al. A review of recurrent neural networks: LSTM cells and network architectures. *Neural computation*. 2019;31(7):1235-1270. https://doi.org/10.1162/neco_a_01199
[23] Vaswani A., et al. Attention is all you need. *arXiv preprint arXiv*:1706.03762v7. 2017. https://doi.org/10.48550/arXiv.1706.03762
[24] Ratinov L., Roth D. Design Challenges and Misconceptions in Named Entity Recognition. In: Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL '09). Association for Computational Linguistics, USA; 2009. p. 147-155. https://doi.org/10.3115/1596374.1596399
[25] Yan R., Jiang X., Dang D. Named entity recognition by using XLNet-BiLSTM-CRF. *Neural Processing Letters*. 2021;53(5):3339-3356. https://doi.org/10.1007/s11063-021-10547-1
[26] Xu G., et al. Sentiment analysis of comment texts based on BiLSTM. *Ieee Access*. 2019;7:51522-51532. https://doi.org/10.1109/ACCESS.2019.2909919
[27] Devlin J., et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv*:1810.04805v2. 2019. https://doi.org/10.48550/arXiv.1810.04805
[28] Kuratov Y., Arkhipov M. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. *arXiv preprint arXiv*:1905.07213. 2019. https://doi.org/10.48550/arXiv.1905.07213
[29] Koroteev M.V. BERT: a review of applications in natural language processing and understanding. *arXiv preprint arXiv*:2103.11943. 2021. https://doi.org/10.48550/arXiv.2103.11943
[30] Press O., Smith N.A., Lewis M. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv*:2108.12409v2. 2022. https://doi.org/10.48550/arXiv.2108.12409
[31] Kosenko D.P., Kuratov Y.M., Zharikova D.R. Accessible Russian Large Language Models: Open-Source Models and Instructive Datasets for Commercial Applications. *Doklady Mathematics. Moscow: Pleiades Publishing*. 2023;108(2):393-398. https://doi.org/10.1134/S1064562423701168
[32] Shavrina T., Pisarevskaya D., Malykh V. Building a Bilingual QA-system with ruGPT-3. *Analysis of Images, Social Networks and Texts. AIST 2021. Lecture Notes in Computer Science*. Vol. 13217. Cham: Springer; 2021. p. 124-136. https://doi.org/10.1007/978-3-031-16500-9_11
[33] Xia W., Qin C., Hazan E. Chain of lora: Efficient fine-tuning of language models via residual learning. *arXiv preprint arXiv*:2401.04151. 2024. https://doi.org/10.48550/arXiv.2401.04151
[34] Gao D., et al. FashionGPT: LLM instruction fine-tuning with multiple LoRA-adapter fusion. *Knowledge-Based Systems*. 2024:112043. https://doi.org/10.1016/j.knosys.2024.112043
[35] Babak N.G., et al. Automatic depersonalization of confidential information. *Russian Technological Journal*. 2023;11(5):7-18. https://doi.org/10.32362/2500-316X-2023-11-5-7-18

Поступила 01.03.2024; одобрена после рецензирования 18.03.2024; принята к публикации 26.03.2024.
Submitted 01.03.2024; approved after reviewing 18.03.2024; accepted for publication 26.03.2024.

Об авторе:

Бабак Никита Григорьевич, аспирант кафедры вычислительных машин, систем и сетей Института информационных и вычислительных технологий, ФГБОУ ВО «Национальный исследовательский университет «МЭИ» (111250, Российская Федерация, Москва, ул. Красноказарменная, д. 14, стр. 1); главный эксперт по защите данных, Департамент кибербезопасности, Публичное акционерное общество «Сбербанк России» (117312, Российская Федерация, г. Москва, ул. Вавилова, д. 19), **ORCID: https://orcid.org/0000-0001-7129-1018,** NGBabak@sber.ru

Автор прочитал и одобрил окончательный вариант рукописи.

About the author:

Nikita G. Babak, Postgraduate Student of the Department of Computing Machines, Systems and Networks, Institute of Information Technologies and Computer Science, National Research University "Moscow Power Engineering Institute" (14/1 Krasnokazarmennaya St., Moscow 111250, Russian Federation); Chief Data Protection Officer, Cybersecurity Department, Sberbank of Russia (19 Vavilova St., Moscow 117312, Russian Federation), **ORCID: https://orcid.org/0000-0001-7129-1018,** NGBabak@sber.ru

The author has read and approved the final manuscript.