Russian Technological Journal.
2023;11(5):7–18

ISSN 2782-3210 (Print)
ISSN 2500-316X (Online)

УДК 004.41, 004.89
https://doi.org/10.32362/2500-316X-2023-11-5-7-18

НАУЧНАЯ СТАТЬЯ

Автоматическое обезличивание конфиденциальной информации

Н.Г. Бабак [@, 1, 2],
Л.Ю. Белорыбкин [2],
Ш.А. Оцоков [3],
А.А. Теренин [2],
А.И. Шаброва [2]

[1] *Национальный исследовательский университет «МЭИ», Москва, 111250 Россия*
[2] *Публичное акционерное общество «Сбербанк России», Москва, 117312 Россия*
[3] *МИРЭА – Российский технологический университет, Москва, 119454 Россия*
[@] *Автор для переписки, e-mail: nikita.enrollee@gmail.com*

Резюме

Цели. В то время как объем персональных данных, передаваемых по сети, продолжает расти, законодательные органы все более жестко регулируют процессы хранения и обработки цифровой информации. В работе рассматривается проблема защиты персональных данных и другой конфиденциальной информации (КИ), например, банковской или врачебной тайны, физических лиц. Одним из способов защиты конфиденциальных данных является их обезличивание – преобразование, в результате которого становится невозможно установить принадлежность этих данных конкретному субъекту. Цель работы – построение автоматической системы, позволяющей быстро и безопасно обезличивать данные с помощью технологий машинного обучения.

Методы. Предлагается использовать модели искусственного интеллекта для реализации системы автоматического обезличивания КИ, т.к. это дает возможность распознавать КИ даже в неструктурированных данных с достаточно высокой точностью без привлечения человеческого труда. Для повышения точности всей системы обезличивания предлагается использовать алгоритмы на основе правил.

Результаты. На конфиденциальных данных, размеченных авторами для решения данной задачи, обучена модель распознавания именованных сущностей, которая в связке с алгоритмами на основе правил в результате имеет значение F_1-меры больше, чем 0.9. Реализовано несколько вариаций алгоритмов обезличивания, что позволяет выбирать между ними для каждой конкретной задачи.

Выводы. Разработанная система решает задачу автоматического обезличивания КИ. Это открывает возможность для безопасной обработки и передачи КИ во многих областях, например, в банковской деятельности, государственном управлении, рекламных кампаниях. Также автоматизация процесса обезличивания делает возможной передачу КИ в тех случаях, когда это необходимо, но невозможно в силу правовых ограничений. Отличительная особенность разработанного решения заключается в том, что обезличиваются как структурированные данные, так и неструктурированные, в т.ч. с сохранением контекста.

Ключевые слова: автоматизированная система, анонимизация, защита информации, кибербезопасность, конфиденциальная информация, машинное обучение, нейросети, обезличивание, персональные данные, распознавание именованных сущностей

• **Поступила:** 10.02.2023 • **Доработана:** 14.06.2023 • **Принята к опубликованию:** 13.07.2023

Для цитирования: Бабак Н.Г., Белорыбкин Л.Ю., Оцоков Ш.А., Теренин А.А., Шаброва А.И. Автоматическое обезличивание конфиденциальной информации. *Russ. Technol. J.* 2023;11(5):7–18. https://doi.org/10.32362/2500-316X-2023-11-5-7-18

Прозрачность финансовой деятельности: Авторы не имеют финансовой заинтересованности в представленных материалах или методах.

Авторы заявляют об отсутствии конфликта интересов.

RESEARCH ARTICLE

Automatic depersonalization of confidential information

Nikita G. Babak [@, 1, 2], Leonid Yu. Belorybkin [2], Shamil A. Otsokov [3], Alexey A. Terenin [2], Anastasia I. Shabrova [2]

[1] *National Research University "Moscow Power Engineering Institute," Moscow, 111250 Russia*
[2] *Sberbank of Russia, Moscow, 117312 Russia*
[3] *MIREA – Russian Technological University, Moscow, 119454 Russia*
[@] *Corresponding author, e-mail: nikita.enrollee@gmail.com*

Abstract

Objectives. As the scope of personal data transmitted online continues to grow, national legislatures are increasingly regulating the storage and processing of digital information. This paper raises the problem of protecting personal data and other confidential information such as bank secrecy or medical confidentiality of individuals. One approach to the protection of confidential data is to depersonalize it, i.e., to transform it so that it becomes impossible to identify the specific subject to whom the data belongs. The aim of the work is to develop a method for the rapid and safe automation of the depersonalization process using machine learning technologies.

Methods. The authors propose the use of artificial intelligence models to implement a system for the automatic depersonalization of personal data without the use of human labor to preclude the possibility of recognizing confidential information even in unstructured data with sufficient accuracy. Rule-based algorithms for improving the precision of the depersonalization system are described.

Results. In order to solve this problem, a model of named entity recognition is trained on confidential data provided by the authors. In conjunction with rule-based algorithms, an F_1 score greater than 0.9 is achieved. For solving specific depersonalization problems, a choice between several implemented anonymization algorithm variants can be made.

Conclusions. The developed system solves the problem of automatic anonymization of confidential data. This opens an opportunity to ensure the secure processing and transmission of confidential information in many areas, such as banking, government administration, and advertising campaigns. The automation of the depersonalization process makes it possible to transfer confidential information in cases where it is necessary, but not currently possible due to legal restrictions. The distinctive feature of the developed solution is that both structured data and unstructured data are depersonalized, including the preservation of context.

Keywords: automated system, anonymization, information protection, cybersecurity, sensitive information, machine learning, neural networks, depersonalization, personal data, named entity recognition

• **Submitted:** 10.02.2023 • **Revised:** 14.06.2023 • **Accepted:** 13.07.2023

For citation: Babak N.G., Belorybkin L.Yu., Otsokov Sh.A., Terenin A.A., Shabrova A.I. Automatic depersonalization of confidential information. *Russ. Technol. J*. 2023;11(5):7–18. https://doi.org/10.32362/2500-316X-2023-11-5-7-18

Financial disclosure: The authors have no a financial or property interest in any material or method mentioned.

The authors declare no conflicts of interest.

ВВЕДЕНИЕ

В современном мире объем хранимых и обрабатываемых данных постоянно растет, а сами данные нуждаются все в более надежной защите. Особенно актуален вопрос защиты персональных данных, передаваемых через компьютерные сети и хранимых в информационных системах. Перечень и порядок обработки персональных данных зафиксирован в Федеральном законе № 152 «О персональных данных». Персональные данные – это любая информация, относящаяся к прямо или косвенно определенному или определяемому физическому лицу[1].

В настоящее время технические средства позволяют организациям проводить сбор и обработку больших объемов данных, необходимых для эффективного развития. Стремительное развитие информационных технологий дает возможность получать доступ к различным данным, что в свою очередь повышает риск утечки информации [1]. Высокий риск незаконного доступа к конфиденциальной информации (КИ) делает задачу обеспечения ее защиты особо актуальной и востребованной.

Одной из мер, направленных на минимизацию рисков причинения вреда человеку в случае утечки его персональных данных из автоматизированных систем (АС), является обезличивание согласно требованиям законодательства. Обезличивание персональных данных – это действия, в результате которых становится невозможным без использования дополнительной информации определить принадлежность персональных данных конкретному субъекту этих данных. Анонимизация позволяет снизить требования, устанавливаемые законом, к АС, обрабатывающим эти данные, что в свою очередь приводит к снижению затрат организаций на разработку таких систем. Таким образом, обезличивание персональных данных не только защищает людей от киберугроз, но и несет положительный экономический эффект. Данная проблема рассматривалась в некоторых работах [2–5], но при этом не были учтены особенности обработки данных на русском языке, имеющем более сложную морфологию. Также в этих работах не осуществлялась достаточная детализация распознаваемых сущностей КИ, что снижает качество обезличенных данных.

1. ПОСТАНОВКА ЗАДАЧИ

Чтобы перейти к обезличиванию, необходимо понять, что конкретно должно быть скрыто в данных. Поэтому можно сказать, что предварительным этапом обезличивания КИ (в частности персональных данных) является ее выделение из всей информации. Ручное выделение определенного вида информации не только сильно замедляет процесс, но и все еще подвержено риску человеческой ошибки.

Исходя из вышесказанного, возникает задача автоматического распознавания и последующего обезличивания КИ в данных, обрабатываемых и передаваемых в АС. Данные могут передаваться в виде файлов, информационного потока и т.д. В связи с этим необходимо предусмотреть возможность извлечения информации из файлов разного расширения и байтового представления.

2. РАСПОЗНАВАНИЕ КОНФИДЕНЦИАЛЬНОЙ ИНФОРМАЦИИ

Существует несколько основных автоматизированных способов распознать информацию – это поиск по словарю, регулярные выражения и алгоритмы машинного обучения. Если распознавание какого-либо вида информации в структурированных данных довольно часто решается с помощью систем, построенных на правилах (rule-based), то с неструктурированными данными не все так однозначно. Более того, существует большое количество разнообразных данных, прямо или косвенно идентифицирующих человека, например, фамилия, имя и отчество, серия и номер паспорта, номер телефона. Для каждого вида данных придется составлять и постоянно актуализировать большие словари, а также сложные правила.

Решить эти проблемы можно, воспользовавшись алгоритмами машинного обучения для распознавания персональных данных в структурированной и неструктурированной информации. В частности, задача распознавания персональных данных сводится к задаче распознавания именованных сущностей (named entity recognition, NER) [6].

[1] Федеральный закон от 27.07.2006 № 152-ФЗ «О персональных данных». [Federal Law No. 152-FZ dated July 27, 2006 “On Personal Data.” (in Russ.).] https://docs.cntd.ru/document/901990046. Дата обращения 09.02.2023. / Accessed February 09, 2023.

Существует несколько основных способов решения этой задачи:

- с помощью статистических методов, например, по количеству определенных символов;
- с помощью правил на основе словарей и регулярных выражений;
- с помощью нейросетей.

Статистические методы не обеспечивают достаточного качества распознавания в данной задаче, особенно в неструктурированных данных. Системы на основе правил хоть и работают сравнительно быстро, но требуют более частой актуализации и подвержены ошибкам в более сложных данных, например, в названиях организаций, фамилиях и именах. Кроме того, подходы на основе статистических методов и правил не учитывают контекст. Справиться с этими недостатками позволяют нейросети. Для задач обработки естественного языка, и, в частности, распознавания именованных сущностей, наиболее передовыми являются нейросети с архитектурой типа трансформеров (transformer) [7]. 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.

Для распознавания некоторых видов конфиденциальной информации, в особенности в структурированных данных, нет необходимости использовать нейросети, а достаточно воспользоваться правилами и статистическими методами. Предварительный анализ и разделение данных на структурированные и неструктурированные позволяет подобрать подходящий алгоритм распознавания. Для распознавания числовых данных больше подходят регулярные выражения с проверкой контрольных разрядов, особенно в структурированных данных. Но стоит заметить, что при достаточно большом наборе неструктурированных данных некоторые числовые персональные данные довольно хорошо распознаются нейросетями. Чтобы воспользоваться алгоритмами машинного обучения, необходимо подготовить обучающую выборку.

2.1. Разметка данных

Обучающая выборка состоит из данных, представленных определенным образом и размеченных на различные атрибуты КИ. Текст разбивается на токены, представленные словами, которым в соответствие ставится тег (метка), обозначающий принадлежность к определенному виду информации.

Теги могут проставляться по одной из следующих схем:

- BIO/IOB, где B (begin) – начало сущности, I (inside) – продолжение сущности, O (outside) – не относится к сущности;
- BILUO/BILOU [8], 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.

Схема BIO является более распространенной, поэтому в данной работе используется именно она.

Разметка токенов тегами может различаться в зависимости от решаемой задачи. В задаче распознавания вложенных именованных сущностей (nested NER) [9, 10] каждому токену в соответствие ставится два тега – сводный и вложенный. Пример разметки представлен в табл. 1. Теги при разметке проставляются вручную квалифицированным в данной области экспертом и чаще всего содержат в названии сокращенное, осмысленное описание информации, содержащейся в размечаемом токене. Например, тег B-SNM получен сокращением фамилии на английском языке – Surname.

Таблица 1. Разметка токенов для распознавания вложенных именованных сущностей

Токен	Сводный тег	Вложенный тег
Сидоров	B-PERSON	B-SNM
Иван	I-PERSON	B-FNM
Петрович	I-PERSON	B-PNM
заключил	O	O
договор	O	O
с	O	O
ООО	B-ORG	B-OPF
Ромашка	I-ORG	B-ORG_NAME

При распознавании прерывистых именованных сущностей (discontinuous NER) [11] разметка тегами может быть представлена в виде таблицы, где количество колонок зависит от максимального количества разрывов для прерывистой сущности. Таким образом, первому слову в прерывистой сущности ставится тег с префиксом B, а все последующие теги смещаются на одну колонку вправо при каждом разрыве и имеют префикс I в случае использования схемы BIO.

В статье решается классическая задача распознавания именованных сущностей, поэтому обучающая выборка разбивается по словам. Каждому слову ставится в соответствие метка, означающая принадлежность к тому или иному виду КИ. Используемый набор данных содержит различные нормативные документы, служебные записки и прочие документы, участвующие в производственной деятельности организации, в которой в последующем будет осуществляться обезличивание.

Разметив данные и обучив модель искусственного интеллекта (ИИ), можно автоматически распознавать КИ, что, в свою очередь, открывает возможность для последующего автоматического обезличивания.

3. ОБЕЗЛИЧИВАНИЕ

После обнаружения в тексте КИ ее можно обезличить обратимым или необратимым способом. В большинстве случаев под обезличиванием понимается необратимая реализация, но при необходимости возможно сохранить таблицу замен в защищенном контуре, чтобы получить обратимое обезличивание.

При любой реализации возможны следующие алгоритмы обезличивания:

- обнуление – удаление всего исходного значения или его значимой части;
- замена константой – замена исходного значения ненулевой константой;
- замена значением из справочника – замена исходного значения случайным отличным значением из справочника, соответствующего заменяемому типу данных;
- замена набором символов – преобразование каждого символа исходного значения в случайный символ, соответствующий типу данных;
- перемешивание – перестановка отдельных значений или групп значений атрибутов персональных данных в массиве персональных данных;
- размытие суммы и даты – замена исходного значения случайным значением, близким к обезличиваемому;
- преобразование на основе заданного выражения – преобразование исходного значения по выражению, содержащему как константы, так и переменные величины;
- маскирование – замена части исходного значения специальным символом или набором символов (маской);
- замена случайным значением – замена исходного значения случайно сгенерированным значением;
- генерация псевдоосмысленных значений – создание текста на основе языковой модели или заданных выражений, позволяющее получить корректный текст с точки зрения основных языковых норм и параметров данных. Также к этому методу можно отнести генерацию фотографий, учитывающую пол и возраст человека.

При выборе подхода к обезличиванию персональных данных целесообразно учитывать рекомендации Роскомнадзора[2], в соответствии с которыми к основным методам обезличивания относятся: метод введения идентификаторов – замена части сведений идентификаторами с созданием таблицы соответствия идентификаторов исходным данным; метод изменения состава или семантики персональных данных путем их замены результатами статистической обработки, преобразования, обобщения или удаления части сведений; метод декомпозиции – разделение множества персональных данных на несколько подмножеств с последующим раздельным хранением подмножеств.

С учетом рекомендаций Роскомнадзора наиболее подходящими являются алгоритмы генерации псевдоосмысленных значений и замены константой. Для реализации обратимого обезличивания необходимо создавать таблицу соответствия исходным данным, при этом следует отметить, что сама таблица должна храниться отдельно от деперсонифицированных данных, а доступ к ней должен быть только у лиц, имеющих право работать с персональными данными в открытом виде.

В зависимости от решаемой задачи могут применяться различные алгоритмы. Например, если необходимо однозначно определить, что было произведено обезличивание и понять, какой вид информации удален, то лучше подойдет алгоритм замены константой. Если необходимо сохранить длину заменяемого значения и также определить, что была произведена анонимизация, то с этой задачей хорошо справится алгоритм частичного маскирования. В случае, когда обезличенные данные необходимо использовать практически также как исходные, например, для обучения моделей ИИ, оптимальным выбором будет алгоритм генерации псевдоосмысленных значений.

4. СИСТЕМА АВТОМАТИЧЕСКОГО ОБЕЗЛИЧИВАНИЯ

Чтобы работа с КИ осуществлялась максимально безопасно, необходимо разработать систему автоматического обезличивания. Процесс автоматического обезличивания с помощью реализованной авторами системы состоит из следующих задач (рис. 1):

1.1 Запрос к системе на обезличивание по API от сторонней автоматизированной системы (АС).
1.2 Запрос к системе на обезличивание через интерфейс от пользователя.
2 Определение типа данных и их предобработка.
3.1 Распознавание КИ в структурированных данных.
3.2 Распознавание КИ в неструктурированных данных.
4 Обезличивание распознанной КИ наиболее подходящим алгоритмом.
5 Возврат обезличенного документа или потока данных.

[2] Приказ Роскомнадзора от 05.09.2013 № 996 «Об утверждении требований и методов по обезличиванию персональных данных». [Order of Roskomnadzor dated September 05, 2013 No. 996 "On approval of requirements and methods for depersonalization of personal data" (in Russ.).]. https://rkn.gov.ru/docs/6_Trebovanija_i_metody_po_obezlichivaniju_personal6nykh_dannykh.docx. Дата обращения 09.02.2023. / Accessed February 09, 2023.

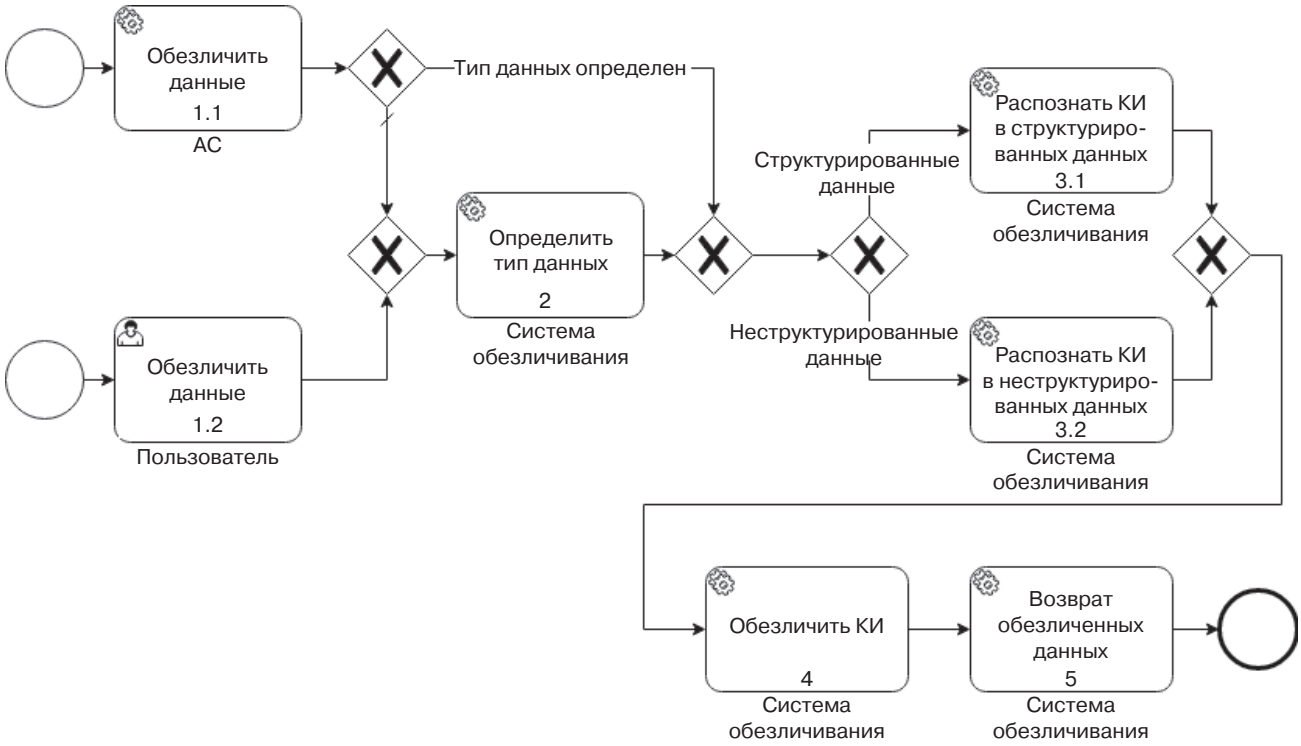

Рис. 1. Процесс обезличивания КИ

Необходимость выделения в системе отдельно распознавания персональных данных в структурированной и неструктурированной информации вызвана тем, что используются различные алгоритмы машинного обучения, в частности учитывающие и не учитывающие синтаксические особенности.

4.1. Подготовка данных

Всего для обучения моделей авторами вручную размечено около миллиона токенов, представленных отдельными словами. Для разметки использованы служебные документы, содержащие персональные данные, банковскую тайну и другую КИ. В качестве схемы разметки выбрана BIO-схема, где первый токен в рамках конфиденциальной сущности имеет префикс B, а все последующие – I. Такой подход позволяет сравнивать и использовать большинство предобученных архитектур, что упрощает процесс создания модели ИИ, как минимум с точки зрения сокращения времени на ее обучение.

Полученный набор размеченных данных разделен на 3 части, где 80% данных используется для обучения модели, 10% – для ее валидации и 10% – для расчета метрик обученной модели. Используется именно такое соотношение, а не 60/20/20, потому что некоторых видов КИ в наборе данных мало, и будет нерационально еще сильнее уменьшать их количество в обучающем наборе.

При разделении текста на токены необходимо сохранить индексы границ разбиения, чтобы в последующем после распознавания КИ обезличить ее строго в пределах заданных границ.

4.2. Обучение модели

Самые передовые результаты в задачах распознавания именованных сущностей показывают нейросети на основе архитектуры трансформеров. Трансформеры, предварительно обученные (pretrained) на большом корпусе данных, хорошо переиспользуются в задачах обработки естественного языка [12]. Для этого достаточно произвести дообучение (fine-tuning) модели на собственных данных, тем самым скорректировав веса для того, чтобы лучше учитывать семантику входных данных.

В качестве основы используется предобученная модель rubert-base-cased [13], использование других подходящих архитектур существенно не влияет на показатели работы модели. В первую очередь это объясняется схожестью различных трансформеров, используемых для решения задачи NER, например, BERT [13], RoBERTa [14] и spaCy [15]. Предобученная модель представляет из себя токенизатор (tokenizer) и кодировщик (encoder), к которым добавлен NER-классификатор. Также для повышения точности и уменьшения ложно положительных срабатываний используются алгоритмы распознавания

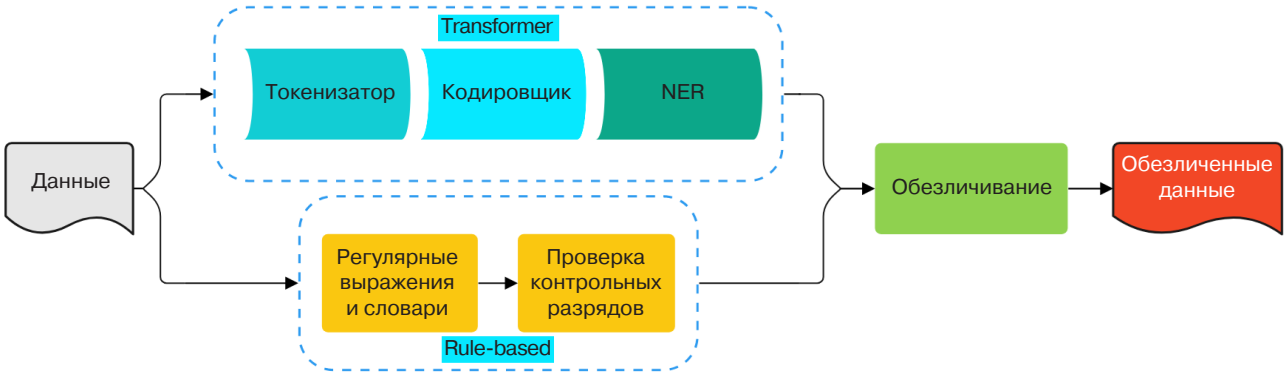

Рис. 2. Обработка данных системой обезличивания

на основе правил (rule-based), в частности, регулярные выражения и проверка контрольных разрядов [16]. Затем результаты обработки данных нейросетями и правилами сводятся в общее предположение о принадлежности текста к одному из видов конфиденциальной информации. Схематичное изображение процесса обезличивания данных предлагаемой системой представлено на рис. 2.

При обработке структурированных данных предпочтение отдается модулю распознавания по правилам, т.к. контекст в таких данных практически отсутствует.

Также в системе обезличивания для сравнения реализована модель распознавания КИ на основе правил без нейросетей и модель PyTorch на основе рекуррентной нейронной сети (RNN, recurrent neural network) [17].

Поскольку большинство существующих систем обезличивания работают на основе правил и имеют сходную между собой реализацию, модель по распознаванию КИ на основе правил без нейросетей служит для получения базовой метрики, с которой можно сравнивать остальные решения. Сравнение реализованной системы обезличивания с другими реализациями будет заведомо представлять предлагаемое решение в более выгодном свете, т.к. сторонние реализации разрабатывались под другой, чаще всего структурированный, набор данных [18–20]. Например, некоторые сторонние системы работают только с персональными данными и не поддерживают обезличивание банковской тайны.

В контексте данной задачи метрика полноты (recall) важна, т.к. необходимо распознать и обезличить все персональные данные, но также важна точность (precision), чтобы количество ложных срабатываний не подрывало доверие к системе. Именно поэтому используется F_1-мера, которая учитывает обе эти метрики и вычисляется по формуле

$$F_1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}.$$

Расчет метрик осуществляется по тестовой части размеченного набора данных, описанного в п. 4.1. Для начала строится матрица ошибок (confusion matrix), в которой по горизонтали располагаются истинные теги из разметки, а по вертикали – теги, предсказанные моделью ИИ. Затем по матрице ошибок подсчитывается количество истинно распознанных атрибутов конфиденциальной информации (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

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

и

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

вычисляются полнота и точность, а затем определяется их среднее гармоническое – F_1-мера [21].

В табл. 2 приведены основные атрибуты КИ и рассчитанная средневзвешенная F_1-мера по разным моделям: модель на основе правил, рекуррентная нейросеть и модель BERT. Стоит отметить, что реализация rule-based работает только на основе правил, а все остальные используют нейросети вместе с регулярными выражениями и другими алгоритмами на основе правил.

Таблица 2. Значения метрики F_1-меры моделей ИИ

Атрибут КИ	F_1 (rule-based)	F_1 (RNN)	F_1 (BERT)
Фамилия	0.804	0.911	0.931
Имя	0.819	0.876	0.929
Отчество	0.874	0.883	0.943
Серия и номер паспорта	0.883	0.907	0.906

Таблица 2. Окончание

Атрибут КИ	F_1 (rule-based)	F_1 (RNN)	F_1 (BERT)
Орган, выдавший паспорт	0.701	0.794	0.899
Номер телефона	0.959	0.969	0.967
E-mail	0.955	0.959	0.964
IP-адрес	0.929	0.932	0.930
Геолокация	0.904	0.919	0.922
Адрес	0.809	0.810	0.912
Дата рождения	0.813	0.837	0.915
ИНН	0.918	0.915	0.919
СНИЛС	0.931	0.935	0.934
Номер полиса ОМС	0.921	0.914	0.921
Номер банковского счета	0.937	0.929	0.936
Номер банковской карты	0.967	0.959	0.965
Номер военного билета	0.892	0.880	0.889
Номер ОГРНИП	0.910	0.909	0.919
Должность	0.812	0.820	0.873
Название организации	0.817	0.899	0.951
Средневзвешенная F_1-мера	0.878	0.898	0.926

Кроме того, проводилось сравнение моделей RoBERTa и spaCy, но они показали метрики аналогичные модели BERT с разбросом значений F_1-меры менее 0.01. В связи с этим выбрана именно модель BERT, т.к. она меньше модели RoBERTa по размеру и имеет более гибкую настройку, чем модель spaCy, что важно для внедрения промышленной версии модели в систему.

Как видно из табл. 2, решение на основе правил значительно уступает моделям машинного обучения по значениям метрики F_1-меры. Особенно эффект заметен в строковых типах данных, где значительную роль играет контекст. Из-за разнородного набора документов рекуррентная нейросеть RNN также справляется с задачей хуже, чем BERT. Исходя из значений метрики F_1-меры, представленных в табл. 2, и того, что модели-трансформеры имеют широкий потенциал развития и переиспользования, в качестве финального решения выбрана модель BERT, которая превосходит остальные решения в среднем на 4%.

Преимущество системы обезличивания с использованием модели BERT перед другими решениями заключается в использовании механизма внимания на себя (self-attention), который позволяет лучше обнаруживать КИ благодаря анализу контекста и важности слов в тексте. Применяемый в модели механизм внимания можно выразить формулой

$$\text{attention} = \text{softmax}\left(\frac{\mathbf{QK}^T}{\sqrt{d_k}}\right)\mathbf{V},$$

где **Q** – вектор запроса, **K** – вектор ключа, **V** – вектор значения, d_k – размерность векторов. Векторы **Q**, **K** и **V** получаются путем перемножения эмбеддинга токена на соответствующие матрицы, полученные при предварительном обучении взятой за основу модели rubert-base-cased. Поскольку в действительности вычисления производятся над векторными представлениями нескольких токенов, то **Q**, **K** и **V** являются матрицами, и перед расчетом произведения **Q** и **K** матрицу **K** необходимо транспонировать [7]. В практической реализации вектор ключа и вектор значения являются одним и тем же вектором и служат для представления токена, а вектор запроса показывает значимость данного токена относительно других.

Функция softmax, выраженная формулой

$$\sigma(\mathbf{Z})_i = \frac{e^{z_i}}{\sum_{j=1}^{N} e^{z_j}},$$

где i и j – индексы элемента вектора в диапазоне от 1 до N, служит для нормализации, т.е. преобразует вектор **z** размерности N к вектору **σ** той же размерности, где все координаты нормированного вектора $\boldsymbol{\sigma}_i$ выражены числом в интервале от 0 до 1, а их сумма равна единице.

Распознавание КИ в неструктурированных данных, представленных изображениями и аудиозаписями, сводится к обработке неструктурированных текстов. Для этого предварительно решается задача OCR (optical character recognition) – оптического распознавания символов [22], а также задача ASR (automatic speech recognition) – автоматического распознавания речи [23].

4.3. Обезличивание

Распознав КИ, система ее обезличивает одним из выбранных алгоритмов. Причем выбор алгоритма возможен как для всего документа или набора данных, так и для отдельного вида КИ. В системе, представленной авторами, реализованы алгоритмы обезличивания, работающие на основе следующих методов:

- замена константой (плейсхолдером) вида {Атрибут_КИ};
- маскирование на символ «*»;
- генерация псевдоосмысленных значений, включающая в себя замену значением из справочника, преобразование на основе заданного выражения и размытие даты.

Например, распознав моделью ИИ в предложении «Сидоров Александр Викторович (ИНН 503199560259) получил перевод на карту 4561 2612 1234 5467» конфиденциальную информацию, представленную фамилией, именем, отчеством, идентификационным номером налогоплательщика (ИНН) и номером банковской карты, пользователю системы можно выбрать один из описанных выше алгоритмов обезличивания. При замене на плейсхолдер рассматриваемое предложение примет следующий вид: «{Фамилия} {Имя} {Отчество} (ИНН {ИНН}) получил перевод на карту {Номер банковской карты}», где КИ заменена на константы, показывающие, какой вид информации ранее находился в предложении. При частичном маскировании КИ заменится на маску, и рассматриваемое предложение примет следующий вид: «С****** ********* ********** (ИНН 50**********) получил перевод 4561 26** **** 5467», где сохранены части слов, не представляющие опасности для идентификации субъекта данных, но позволяющие определить косвенные признаки, например, банк, выдавший карту. При замене на псевдоосмысленные значения рассматриваемое предложение примет следующий вид: «Лазарев Владислав Алексеевич (ИНН 503195234624) получил перевод на карту 4561 2698 5513 5467». Последний алгоритм в отличие от двух предыдущих работает медленнее, т.к. генерируются псевдоосмысленные данные, но формирует полностью осмысленный текст, который можно применять, например, в задачах машинного обучения.

Используемый алгоритм обезличивания зависит от решаемой задачи и его выбор остается за пользователем или АС.

ЗАКЛЮЧЕНИЕ

Всего для обучения модели ИИ размечено около миллиона токенов, благодаря чему покрыто большое количество способов представления данных, содержащих КИ. В случае, когда количество типов обезличиваемых документов невелико, для дообучения модели достаточно разметить небольшой набор данных, который включает все необходимые виды КИ. Кроме того, благодаря использованию моделей-трансформеров в большинстве случаев дообучение модели не требуется, что позволяет переиспользовать разработанную систему в различных организациях «как есть» или с корректировкой на небольшом объеме данных. Использование нейросетей позволяет избавиться от составления огромных справочников фамилий и имен, а также других сущностей, идентифицирующих человека. Регулярные выражения в свою очередь учитывают особенности структуры, например, существующие серии, коды и банковские идентификационные номера, что позволяет обнаруживать даже те данные, на которых модель ранее не обучалась.

Отличительным преимуществом представленной авторами системы обезличивания от существующих является поддержка как структурированных, так и неструктурированных данных. Кроме того, в большинстве известных систем обезличивание осуществляется разрушающим способом, после чего данные становятся непригодными для многих задач, например, для машинного обучения.

Средняя взвешенная F_1-мера реализованной модели распознавания КИ превысила 0.9, что говорит о высоком качестве системы обезличивания. Благодаря этому исчезает потребность в привлечении человеческого труда для обнаружения КИ.

Реализованные алгоритмы обезличивания, основанные на методе замены константой, маскирования и генерации псевдоосмысленных значений, покрывают все основные задачи обезличивания: обезличивание с возможностью однозначного определения факта маскирования, синонимическое обезличивание, необратимое и обратимое обезличивание и другие. Также указанные алгоритмы позволяют автоматически обезличивать распознанные конфиденциальные данные. Практическая ценность разработанной авторами системы автоматического обезличивания заключается в том, что обезличенные с ее помощью конфиденциальные данные можно использовать аналогично исходным данным, но без риска нарушения требований кибербезопасности. При этом затраты на процедуры обезличивания практически отсутствуют, т.к. процесс автоматизирован.

Система обезличивания конфиденциальных данных содержит, по меньшей мере, один процессор и одну память, соединенную с процессором, которая содержит машиночитаемые инструкции. Кроме того, система обезличивания может выполняться на сервере программируемым логическим контроллером и любыми другими устройствами, способными выполнять заданную последовательность инструкций.

Предложенную систему автоматического обезличивания можно встроить практически в любой цикл, связанный с передачей и обработкой КИ, что позволит автоматически распознавать ее и обезличивать. Благодаря этому снижается риск раскрытия личности в случае утечки данных. Например, автоматическое обезличивание может применяться в банковской сфере, государственных услугах, Data Science сообществе [24] и прочей деятельности, связанной с обработкой КИ, в частности – персональных данных.

Вклад авторов. Все авторы в равной степени внесли свой вклад в исследовательскую работу.

Authors' contribution. All authors equally contributed to the research work.

СПИСОК ЛИТЕРАТУРЫ

1. Шаброва А.И., Теренин А.А, Бабак Н.Г. Методика оценки риска от разглашения конфиденциальной информации в источниках данных с использованием интеллектуального анализа данных. *Современные информационные технологии и ИТ-образование*. 2022;18(3):666–679. https://doi.org/10.25559/SITITO.18.202203.666-679
2. Столбов А.П. Обезличивание персональных данных в здравоохранении. *Врач и информационные технологии*. 2017;3:76–91. URL: https://elibrary.ru/zgyvot
3. Спеваков А.Г., Калуцкий И.В., Никулин Д.А., Шумайлова В.А. Обезличивание персональных данных при обработке в автоматизированных информационных системах. *Телекоммуникации*. 2016;10:16–20. URL: https://www.elibrary.ru/wwvxmt
4. Oleksy M., Ropiak N., Walkowiak T. Automated anonymization of text documents in Polish. *Procedia Computer Science*. 2021;192(1):1323–1333. https://doi.org/10.1016/j.procs.2021.08.136
5. Saluja B., Kumar G., Sedoc J., Callison-Burch C. Anonymization of Sensitive Information in Medical Health Records. In: *CEUR Workshop Proceedings*. 2019;2421:647–653. URL: https://ceur-ws.org/Vol-2421/MEDDOCAN_paper_2.pdf
6. Roy A. *Recent Trends in Named Entity Recognition* (*NER*). arXiv. 2021. https://doi.org/10.48550/arxiv.2101.11420
7. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I. Attention is all you need. In: *Advances in Neural Information Processing Systems*. 2017. https://doi.org/10.48550/arXiv.1706.03762
8. Ratinov L., Roth D. Design Challenges and Misconceptions in Named Entity Recognition. In: *Proceedings of the Thirteenth Conference on Computational Natural Language Learning* (*CoNLL 2009*). 2009. P. 147–155. URL: https://aclanthology.org/W09-1119.pdf
9. Fisher J., Vlachos A. *Merge and label: A novel neural network architecture for nested NER*. arXiv. 2019. https://doi.org/10.48550/arXiv.1907.00464
10. Fu Y., Tan C., Chen M., Huang S., Huang F. Nested named entity recognition with partially-observed TreeCRFs. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2021;35(14):12839–12847. https://doi.org/10.1609/aaai.v35i14.17519
11. Dai X., Karimi S., Hachey B., Paris C. *An effective transition-based model for discontinuous NER*. arXiv. 2020. https://doi.org/10.48550/arXiv.2004.13454
12. Lothritz C., Allix K., Veiber L., Klein J., Bissyande T.F.D.A. Evaluating pretrained transformer-based models on the task of fine-grained named entity recognition. In: *Proceedings of the 28th International Conference on Computational Linguistics*. 2020. P. 3750–3760. http://doi.org/10.18653/v1/2020.coling-main.334
13. Kuratov Y., Arkhipov M. *Adaptation of deep bidirectional multilingual transformers for Russian language*. arXiv. 2019. https://doi.org/10.48550/arXiv.1905.07213
14. Conneau A., Khandelwal K., Goyal N., Chaudhary V., Wenzek G., Guzman F., Grave E., Ott M., Zettlemoyer L., Stoyanov V. *Unsupervised cross-lingual representation learning at scale*. arXiv. 2020. https://doi.org/10.48550/arXiv.1911.02116
15. Patel A.A., Arasanipalai A.U. *Applied Natural Language*

REFERENCES

1. Shabrova A.I., Terenin A.A., Babak N.G. Methodology for risk assessment from confidential information disclosure in data sources using data mining. *Sovremennye informacionnye tehnologii i IT-obrazovanie = Modern Information Technologies and IT-Education*. 2022;18(3):666–679 (in Russ.). https://doi.org/10.25559/SITITO.18.202203.666-679
2. Stolbov A.P. De-identification of personal data in health care. *Vrach i informacionnye tekhnologii = Medical Doctor and Information Technologies*. 2017;3:76–91 (in Russ.). Available from URL: https://elibrary.ru/zgyvot
3. Spevakov A.G., Kalutskiy I.V., Nikulin D.A., Shumailova V.A. Depersonalization of personal data during processing of information in automated systems. *Telekommunikatsii = Telecommunications*. 2016;10:16–20 (in Russ.). Available from URL: https://www.elibrary.ru/wwvxmt
4. Oleksy M., Ropiak N., Walkowiak T. Automated anonymization of text documents in Polish. *Procedia Computer Science*. 2021;192(1):1323–1333. https://doi.org/10.1016/j.procs.2021.08.136
5. Saluja B., Kumar G., Sedoc J., Callison-Burch C. Anonymization of Sensitive Information in Medical Health Records. In: *CEUR Workshop Proceedings*. 2019;2421:647–653. Available from URL: https://ceur-ws.org/Vol-2421/MEDDOCAN_paper_2.pdf
6. Roy A. *Recent Trends in Named Entity Recognition* (*NER*). arXiv. 2021. https://doi.org/10.48550/arxiv.2101.11420
7. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I. Attention is all you need. In: *Advances in Neural Information Processing Systems*. 2017. https://doi.org/10.48550/arXiv.1706.03762
8. Ratinov L., Roth D. Design Challenges and Misconceptions in Named Entity Recognition. In: *Proceedings of the Thirteenth Conference on Computational Natural Language Learning* (*CoNLL 2009*). 2009. P. 147–155. Available from URL: https://aclanthology.org/W09-1119.pdf
9. Fisher J., Vlachos A. *Merge and label: A novel neural network architecture for nested NER*. arXiv. 2019. https://doi.org/10.48550/arXiv.1907.00464
10. Fu Y., Tan C., Chen M., Huang S., Huang F. Nested named entity recognition with partially-observed TreeCRFs. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2021;35(14):12839–12847. https://doi.org/10.1609/aaai.v35i14.17519
11. Dai X., Karimi S., Hachey B., Paris C. *An effective transition-based model for discontinuous NER*. arXiv. 2020. https://doi.org/10.48550/arXiv.2004.13454
12. Lothritz C., Allix K., Veiber L., Klein J., Bissyande T.F.D.A. Evaluating pretrained transformer-based models on the task of fine-grained named entity recognition. In: *Proceedings of the 28th International Conference on Computational Linguistics*. 2020. P. 3750–3760. http://doi.org/10.18653/v1/2020.coling-main.334
13. Kuratov Y., Arkhipov M. *Adaptation of deep bidirectional multilingual transformers for Russian language*. arXiv. 2019. https://doi.org/10.48550/arXiv.1905.07213

Processing in the Enterprise. O'Reilly Media, Inc.; 2021. 336 p. ISBN 978-1-4920-6257-8. URL: https://spacy.io/universe/project/applied-nlp-in-enterprise/
16. Singco V.Z., Trillo J., Abalorio C., Bustillo J.C., Bojocan J., Elape M. OCR-based Hybrid Image Text Summarizer using Luhn Algorithm with Finetune Transformer Models for Long Document. *Int. J. Emerging Technol. Adv. Eng*. 2023;13(02):47–56. http://doi.org/10.46338/ijetae0223_07
17. Soltau H., Shafran I., Wang M., Shafey L.E. *RNN Transducers for Nested Named Entity Recognition with constraints on alignment for long sequences*. arXiv. 2022. https://doi.org/10.48550/arXiv.2203.03543
18. Абирхаев Е.А., Ерохин А.Ф., Пушкин П.Ю. Методы обезличивальных данных: обзор и анализ. *Наукосфера*. 2021;6(2):57–31. URL: https://www.elibrary.ru/item.asp?id=46561812
19. Кротов А.Д., Серышев А.С., Ефанова Н.В. Разработка приложения для обезличивания персональных данных. В сб.: *Цифровизация экономики: направления, методы, инструменты*: *сб. материалов III всероссийской научно-практической конференции*. Краснодар: Кубанский государственный аграрный университет; 2021. С. 294–297. ISBN 978-5-9074-3005-1. URL: https://www.elibrary.ru/item.asp?id=44891383
20. Фот Ю.Д., Коробова Е.О. Обезличивание персональных данных в системе управления персоналом предприятий нефтегазового сектора. В сб.: *Роль нефтегазового сектора в технико-экономическом развитии Оренбуржья: сб. трудов научно-практической конференции*. Саратов: ООО «Амирит»; 2021. С. 161–168. ISBN 978-5-0014-0888-8. URL: https://www.elibrary.ru/item.asp?id=48392659
21. 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
22. Du Y., Li C., Guo R., Yin X., Liu W., Zhou J., Bai Y., Yu Z., Yang Y., Dang Q., Wang H. *PP-OCR: A practical ultra lightweight OCR system*. arXiv. 2020. https://doi.org/10.48550/arXiv.2009.09941
23. Pan J., Shapiro J., Wohlwend J., Han K.J., Lei T., Ma T. *ASAPP-ASR: Multistream CNN and self-attentive SRU for SOTA speech recognition*. arXiv. 2020. https://doi.org/10.48550/arXiv.2005.10469
24. Ryffel T., Trask A., Dahl M., Wagner B., Mancuso J., Rueckert D., Passerat-Palmbach J. *A generic framework for privacy preserving deep learning*. arXiv. 2018. https://doi.org/10.48550/arXiv.1811.04017

14. Conneau A., Khandelwal K., Goyal N., Chaudhary V., Wenzek G., Guzman F., Grave E., Ott M., Zettlemoyer L., Stoyanov V. *Unsupervised cross-lingual representation learning at scale*. arXiv. 2020. https://doi.org/10.48550/arXiv.1911.02116
15. Patel A.A., Arasanipalai A.U. *Applied Natural Language Processing in the Enterprise*. O'Reilly Media, Inc.; 2021. 336 p. ISBN 978-1-4920-6257-8. Available from URL: https://spacy.io/universe/project/applied-nlp-in-enterprise/
16. Singco V.Z., Trillo J., Abalorio C., Bustillo J.C., Bojocan J., Elape M. OCR-based Hybrid Image Text Summarizer using Luhn Algorithm with Finetune Transformer Models for Long Document. *Int. J. Emerging Technol. Adv. Eng*. 2023;13(02):47–56. http://doi.org/10.46338/ijetae0223_07
17. Soltau H., Shafran I., Wang M., Shafey L.E. *RNN Transducers for Nested Named Entity Recognition with constraints on alignment for long sequences*. arXiv. 2022. https://doi.org/10.48550/arXiv.2203.03543
18. Abirkhaev E.A., Erokhin A.F., Pushkin P.Yu. Methods of depersonalizing data: overview and analysis. *Naukosfera*. 2021;6(2):57–31 (in Russ.). Available from URL: https://www.elibrary.ru/item.asp?id=46561812
19. Seryshev A.S., Krotov A.D., Efanova N.V. Development of an application for personal data depersonalization. In: *Digitalization of the Economy: Directions, Methods, Tools: Proceedings of the 3rd All-Russian Scientific and Practical Conference*. Krasnodar: Kuban State Agrarian University; 2021. P. 294–297 (in Russ.). ISBN 978-5-9074-3005-1. Available from URL: https://www.elibrary.ru/item.asp?id=44891383
20. Fot U.D., Korobova E.O. Depersonalization of personal data in the personnel management system of oil and gas sector enterprises. In: *The Role of the Oil and Gas Sector in the Technical and Economic Development of the Orenburg Region: Proceedings of the scientific-practical conference. Saratov: Amirit;* 2021. P. 161–168 (in Russ.). ISBN 978-5-0014-0888-8. Available from URL: https://www.elibrary.ru/item.asp?id=48392659
21. 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
22. Du Y., Li C., Guo R., Yin X., Liu W., Zhou J., Bai Y., Yu Z., Yang Y., Dang Q., Wang H. *PP-OCR: A practical ultra lightweight OCR system*. arXiv. 2020. https://doi.org/10.48550/arXiv.2009.09941
23. Pan J., Shapiro J., Wohlwend J., Han K.J., Lei T., Ma T. *ASAPP-ASR: Multistream CNN and self-attentive SRU for SOTA speech recognition*. arXiv. 2020. https://doi.org/10.48550/arXiv.2005.10469
24. Ryffel T., Trask A., Dahl M., Wagner B., Mancuso J., Rueckert D., Passerat-Palmbach J. *A generic framework for privacy preserving deep learning*. arXiv. 2018. https://doi.org/10.48550/arXiv.1811.04017

Об авторах

Бабак Никита Григорьевич, аспирант, кафедра вычислительных машин, систем и сетей Института информационных и вычислительных технологий ФГБОУ ВО «Национальный исследовательский университет «МЭИ» (111250, Россия, Москва, Красноказарменная ул., д. 14, стр. 1); главный эксперт по защите данных, Департамент кибербезопасности ПАО «Сбербанк России» (117312, Россия, Москва, ул. Вавилова, д. 19). E-mail: nikita.enrollee@gmail.com. ResearcherID HHY-9372-2022, SPIN-код РИНЦ 3687-6548, https://orcid.org/0000-0001-7129-1018

Белорыбкин Леонид Юрьевич, директор проектов по защите данных, Департамент кибербезопасности ПАО «Сбербанк России» (117312, Россия, Москва, ул. Вавилова, д. 19). E-mail: lbelorybkin@gmail.com. https://orcid.org/0000-0002-8575-5773

Оцоков Шамиль Алиевич, д.т.н., профессор, кафедра КБ-4 «Интеллектуальные системы информационной безопасности» Института кибербезопасности и цифровых технологий ФГБОУ ВО «МИРЭА – Российский технологический университет» (119454, Россия, Москва, пр-т Вернадского, д. 78). E-mail: shamil24@mail.ru. Scopus Author ID 57212622267, https://orcid.org/0000-0001-7451-5443

Теренин Алексей Алексеевич, к.т.н., управляющий директор, Департамент кибербезопасности ПАО «Сбербанк России» (117312, Россия, Москва, ул. Вавилова, д. 19). E-mail: aaterenin@yandex.ru. http://orcid.org/0000-0002-6242-6117

Шаброва Анастасия Игоревна, архитектор по защите данных, Департамент кибербезопасности ПАО «Сбербанк России» (117312, Россия, Москва, ул. Вавилова, д. 19). E-mail: shabrova1113@gmail.com. https://orcid.org/0000-0002-4315-3061

About the authors

Nikita G. Babak, Postgraduate Student, Department of Computing Machines, Systems and Networks, Institute of Information Technologies and Computer Science, National Research University MPEI (14/1, Krasnokazarmennaya ul., Moscow, 111250 Russia); Chief Data Protection Officer, Cybersecurity Department, Sberbank of Russia (19, Vavilova ul., Moscow, 117312 Russia). E-mail: nikita.enrollee@gmail.com. ResearcherID HHY-9372-2022, RSCI SPIN-code 3687-6548, https://orcid.org/0000-0001-7129-1018

Leonid Yu. Belorybkin, Director of Data Protection Projects, Cybersecurity Department, Sberbank of Russia (19, Vavilova ul., Moscow, 117312 Russia). E-mail: lbelorybkin@gmail.com. https://orcid.org/0000-0002-8575-5773

Shamil A. Otsokov, Dr. Sci. (Eng.), Professor, Department of Intelligent Information Security Systems, Institute of Cybersecurity and Digital Technologies, MIREA – Russian Technological University (78, Vernadskogo pr., Moscow, 119454 Russia). E-mail: shamil24@mail.ru. Scopus Author ID 57212622267, https://orcid.org/0000-0001-7451-5443

Alexey A. Terenin, Cand. Sci. (Eng.), Managing Director, Cybersecurity Department, Sberbank of Russia (19, Vavilova ul., Moscow, 117312 Russia). E-mail: aaterenin@yandex.ru. http://orcid.org/0000-0002-6242-6117

Anastasia I. Shabrova, Data Protection Architect, Cybersecurity Department, Sberbank of Russia (19, Vavilova ul., Moscow, 117312 Russia). E-mail: shabrova1113@gmail.com. https://orcid.org/0000-0002-4315-3061