УДК 004.912, 004.62 **DOI** 10.15622/ia.24.1.7

М.Д. Кузнецов, Е.С. Новикова

КОРПУС ПОЛИТИК КОНФИДЕНЦИАЛЬНОСТИ ВЕБ-СЕРВИСОВ И УСТРОЙСТВ ИНТЕРНЕТА ВЕЩЕЙ ДЛЯ АНАЛИЗА ИНФОРМИРОВАННОСТИ СУБЪЕКТОВ ПЕРСОНАЛЬНЫХ ДАННЫХ

Кузнецов М.Д., Новикова Е.С. **Корпус политик конфиденциальности веб-сервисов и устройств Интернета Вещей для анализа информированности субъектов персональных данных.**

Аннотация. Информация о том, какие персональные данные собираются и обрабатываются различными устройствами и цифровыми сервисами, представлена в политиках конфиденциальности, однако, как показывают исследования, пользователи крайне редко их читают и, как следствие, не осознают, какие риски информационной безопасности, связанные с обработкой персональных данных, возникают. Решение проблемы повышения информированности субъектов персональных данных связано с разработкой методов поддержки принятия решений, которые представляют политики конфиденциальности в виде, более простом для понимания, например, в виде количественных оценок рисков и пиктограмм и позволяют принимать осознанные решения. Их разработка требует наличия структурированного и размеченного корпуса документов. В настоящей работе систематизируются корпусы политик конфиденциальности, находящиеся в открытом доступе, показываются их отличительные характеристики, такие как год создания, объем и наличие аннотаций. Также представлено описание нового корпуса документов, написанных на русском языке, даются результаты структурного и семантического анализа собранных политик безопасности, и выполняется сравнение с корпусом политик конфиденциальности, написанных на английском языке. Показано, что описание сценариев хранения, сбора и обработки данных в документах на русском языке составляет всего 25% объема текста документа, что может говорить об отсутствии деталей о том, какие типы данных собираются, какие механизмы для сбора используются, и каковы сроки их хранения, что влияет на “прозрачность” использования персональных данных.

Ключевые слова: персональные данные, политики конфиденциальности, корпус документов, семантический анализ, латентное размещение Дирихле.

1. Введение. По мере роста уровня цифровизации современного общества увеличивается объем собираемых и обрабатываемых персональных данных, что в свою очередь приводит к росту уровня информационных угроз, связанных с их утечкой. Согласно аналитическому отчету компании InfoWatch [1] в мире наблюдается уверенное увеличение числа инцидентов, связанных с нарушением конфиденциальности данных с ограниченным доступом, в т.ч. персональных данных. Следует отметить, что причиной утечек конфиденциальной информации не всегда является действия внешнего или внутреннего злоумышленника. Так, например, в 2024 компания Avast, разрабатывающая решения в области информационной безопасности,

была оштрафована за передачу персональных данных своих пользователей третьим лицам без их согласия [2].

Риски утечки персональной информации также возникают в результате использования различных “умных” устройств и веб-сервисов. Согласно исследованиям, доля “умных” домохозяйств в мире [3] и в России [4] постоянно увеличивается, спросом пользуются как системы управления домом с голосовыми ассистентами, так и различные датчики безопасности и охранные системы, включая IP-камеры. Между тем, регулярно появляются отчеты о выявленных уязвимостях в программном коде устройств Интернета Вещей и управляющих сервисах [5], эксплуатация которых приводит к утечке разнородной информации, начиная с видео данных, заканчивая данными о потреблении электроэнергии домохозяйствами [6, 7]. Эти данные позволяют получить подробную информацию о распорядке дня, привычках и образе жизни домохозяйства, которая может быть использована третьими лицами в различных целях, например, для недобросовестного целевого маркетинга, вмешательства в личную жизнь и совершения преступных действий, связанных с финансами.

Для предотвращения неправомочного использования персональных данных в 2022 в РФ были приняты поправки к ФЗ “О Персональных данных” № 152-ФЗ, в которых были ужесточены требования к сбору, обработке, хранению и передаче персональных данных третьим лицам. В частности, в статье 5 главы 2 ФЗ № 152-ФЗ сформулированы требования к целям обработки таких данных, которые должны быть “конкретными, заранее определенными и законными”, а согласия на их обработку должны быть “информированными и сознательными”.

Информация о том, какие персональные данные собираются и обрабатываются устройствами, веб-сервисами и приложениями, обычно представлена в соглашениях на обработку персональных данных и/или в политиках конфиденциальности компаний разработчиков устройств. Однако, в большинстве случаев эти документы написаны на сложном юридическом языке, что не всегда позволяет пользователю их понять [8, 9]. Как следствие, они дают свое согласие на сбор, обработку и хранение персональных данных без четкого понимания того, как организован этот процесс и какие риски, связанные с обработкой данных, возникают. Таким образом, проблема обеспечения прозрачности политик безопасности, включающая в себя задачу повышения осведомленности владельцев данных о том, как используются их данные, является важной и тесно связана с разработкой методик, способов, инструментов и систем поддержки принятия решений (СППР), позволяющих пользователю

цифровых сервисов и услуг оценить целесообразность использования того или иного сервиса в контексте сбора и использования его персональных данных.

Проблема повышения уровня информированности пользователей устройств Интернета Вещей активно исследуется во всем мире [10–12], и в настоящее время предложены различные решения по анализу политик конфиденциальности, написанных на естественном языке [11–16]. Они позволяют выявить и охарактеризовать в структурированном виде различные сценарии использования персональных данных, такие как сбор, хранение, передача персональных данных и т.д. Однако, разработанные методы анализа документов предложены для политик конфиденциальности, написанных на английском языке, проблема анализа и исследования политик конфиденциальности на русском языке не проработана.

В данной статье авторы представляют структурный и семантический анализ нового корпуса русскоязычных политик конфиденциальности веб-сервисов, которые доступны на территории России, и выполняют его сравнительный анализ с другим корпусом политик конфиденциальности, написанных на английском языке. Таким образом, научно-практический вклад авторов статьи заключается в:

1. сравнительном анализе существующих корпусов политик конфиденциальности, которые могут быть использованы для разработки методов их анализа на основе глубокого машинного обучения;
2. создании нового корпуса политик конфиденциальности, написанных на русском языке;
3. выполнении структурного и семантического анализа нового корпуса политик конфиденциальности и сравнения его характеристик с корпусом политик конфиденциальности на английском языке.

Полученный корпус политик конфиденциальности послужит основой для разработки автоматизированных методов обработки и анализа пользовательских соглашений, которые могут применяться пользователями при принятии решений, касающихся управления персональными данными при выборе цифровых сервисов и устройств, которые выполняют обработку персональных данных.

Статья организована следующим образом. В разделе 2 обсуждаются основные направления исследований в области анализа текстов политик конфиденциальности и выполняется сравнительный анализ корпусов политик конфиденциальности, находящихся в открытом доступе. В разделе 3 представлена методика сбора политик конфиденциальности, которая была использована для создания нового корпуса. В разделах 4

и 5 представлены результаты структурного и семантического анализа собранного корпуса политик конфиденциальности, а также выполняется его сравнительный анализ с корпусом политик конфиденциальности для устройств Интернета Вещей, написанных на английском языке [17]. В разделе 6 обсуждаются полученные результаты и формулируются дальнейшие направления исследований.

2. Методы анализа политик конфиденциальности. Согласно ФЗ "О Персональных данных" № 152-ФЗ персональными данными являются данные, относящиеся "прямо или косвенно определенному или определяемому физическому лицу (субъекту персональных данных)". К ним относятся как общедоступные данные, такие как фамилия и отчество субъекта, его возраст, образование, электронная почта и телефон, так и биометрические, специальные и иные персональные данные. Следует отметить, что в Российском законодательстве нет четкого определения, какие персональные данные следует относить к категории "иные", однако в мировой практике [10, 18] к персональным данным относятся в том числе данные, которые позволяют уникально идентифицировать устройства пользователя, например, IP- и MAC-адреса устройств, цифровой отпечаток браузера и т.д., поскольку если указать свои персональные данные, например, ФИО или телефон, цифровой отпечаток устройства позволяет уникально идентифицировать пользователя и отслеживать его поведение в сети.

Анализ существующих подходов к анализу политик конфиденциальности позволил авторам выделить два основных подхода к решению данной проблемы. В рамках первого подхода решается задача построения формализованного представления различных сценариев использования персональных данных [12, 19–21]. Под сценариями использования персональных данных понимается деятельность, связанная с обработкой персональных данных, включая их сбор, хранение и передачу третьим лицам. Формальное представление таких сценариев использования может служить основой как для определения правил обработки персональных данных [22], так и для оценки рисков нарушения их конфиденциальности [12].

В рамках второго подхода выполняется анализ политик конфиденциальности, написанных на естественном языке, целью разрабатываемых методов и моделей является повышение "прозрачности" и понятности документов для пользователей. В [15, 23, 24] обсуждается проблема сбора политик конфиденциальности и предлагается схема аннотирования, которая отражает основные характеристики различных сценариев использования персональных данных. В [15] представлен

подход к автоматическому определению различных типов персональных данных, таких как электронная почта, контактный телефон, адрес, геолокация, которые упоминаются в тексте политик конфиденциальности, разработанных для Android-приложений. В [25] авторы разработали подход к определению вариантов отказа от использования персональных данных, представленных в тексте политики конфиденциальности. В [26, 27] авторы исследуют проблему неоднозначности и общности политики безопасности и предлагают основанный на онтологии подход к уменьшению нечеткости терминов политики безопасности путем установления семантических отношений между ними.

Наиболее часто используемым корпусом политик конфиденциальности является набор данных OPP-115 [23], который был разработан в рамках проекта Usable Privacy Policy [28] (UPP). Он включает в себя 115 политик конфиденциальности веб-сайтов, которые были собраны с помощью сервиса Amazon Alexa [29], который отражает актуальность и популярность веб-сайтов и публичной веб-директории DMOZ.org, содержащей ссылки на веб-сайты различных категорий, внесенных туда реальными пользователями, что может привести к попаданию в директорию нерелевантных веб-сайтов. Несомненным преимуществом этого корпуса документов является наличие аннотаций и схемы аннотирования, разработанной ее авторами. Она включает в себя различные сценарии использования персональных данных и информацию об экспертах, выполняющих аннотирование текстов. Каждая политика аннотировалась несколькими экспертами, что позволило получить более 20 000 аннотаций, отражающих различные аспекты использования персональных данных. В [30] была продемонстрирована связь между разработанной схемой аннотаций и принципами Общего регламента ЕС о защите данных (GDPR).

Другим аннотированным корпусом политик конфиденциальности является набор данных APP-350 [24], состоящий из политик конфиденциальности приложений, размещенных на площадке Google Play [31]. Авторы не предоставляют подробного описания того, как он был собран, однако можно предположить, что для его создания использован программный интерфейс сервиса Google Play, который предоставляет широкие возможности для сбора необходимых данных. Набор данных MAPS [24] представляет собой расширение корпуса политик APP-350. Он также сформирован на основе политик конфиденциальности приложений, представленных на платформе Google Play, и состоит из более 1 миллиона документов. Однако, в отличие от APP-350, он не содержит аннотаций.

В [32] представлен корпус политик конфиденциальности, отличительной чертой которого является период его формирования: сбор документов осуществлялся на протяжении более 20 лет. Таким образом, он состоит из более чем миллиона документов, которые обновлялись и изменялись в течение этого периода времени. Его авторы также разработали программный инструмент, который извлекает различные фрагменты, такие как n-граммы, именованные сущности, URL-адреса, чтобы оценить, как содержание политики безопасности меняется со временем, и показали, что с течением времени политики конфиденциальности становятся все более сложными для понимания. Основным источником данных для сбора данного набора послужил сервис Amazon Alexa.

Для тестирования методов анализа политик конфиденциальности на больших объемах данных Р. Заим и К. Барбер [33] собрали набор политик конфиденциальности, написанных на разных языках и собранных для более чем 1,5 млн. веб-сайтов. В качестве исходной точки сбора документов они использовали DMOZ. Следует отметить, что в ходе исследования собранного набора данных, авторы также показали, какие категории сайтов чаще всего не имеют политики конфиденциальности. В настоящее время ресурс DMOZ недействителен и заменен аналогичным проектом Curlie [34], который построен на базе проектов Open Directory Project (ODP) и DMOZ.

В [35] представлен набор данных PrivaSeer, состоящий из более чем миллиона политик безопасности, написанных на английском языке. Его авторы оценили уровень сходства между документами, провели тесты на их читаемость, проанализировали наличие различных сценариев использования персональных данных с помощью поиска ключевых фраз и слов, также был выполнен его семантический анализ с помощью методов тематического моделирования.

В корпусе политик IoTDataset [17] представлены политики конфиденциальности, разработанные специально для устройств Интернета Вещей. Сбор политик осуществлялся путем анализа продуктов на площадках интернет-торговли Amazon [36] и Walmart [37]. Авторы рассматривали следующие типы умных устройств: “умные весы”, “умные часы”, “умный браслет” и пр. Были проанализированы результаты поисковых запросов для первых 30-ти страниц. Было выявлено, что только 23% производителей умных устройств имеют свой официальный сайт, и чуть более половины из них имеют собственную политику конфиденциальности. Всего было получено 798 документов, после исключения политик, длины которых в символах не превышали

1 000, осталось 592 документа. Следует отметить, что ручной анализ “коротких” документов показал, что они появляются либо из-за того, что у некоторых производителей на сайте пустая страница с политикой конфиденциальности, либо она отсутствует. Позже предложенный авторами подход к сбору политик конфиденциальности был использован в [38] для создания набора политик в рамках проекта PrivacyLens, в котором уже содержится более 1200 документов.

Перечисленные выше корпусы политик конфиденциальности активно используются в исследовательских проектах, посвященных анализу политик конфиденциальности, написанных на естественном языке, однако ни один из них не содержит документы, написанные на русском языке. В таблице 1 приведен сравнительный анализ корпусов политик конфиденциальности, рассмотренных выше. Они активно используются в исследовательских проектах, посвященных анализу политик конфиденциальности, в частности в рамках проекта Polisis [39] разработан сервис, который позволяет визуализировать сценарии использования персональных данных, извлеченные из политик конфиденциальности, написанных на английском языке, а в проекте Pribot [39] решена задача создания чат бота, который отвечает на вопросы по политикам. Аналогичных сервисов, позволяющих анализировать политики конфиденциальности на русском языке, нет, поэтому решаемая в данной работе задача по созданию корпуса политик конфиденциальности на русском языке, может послужить основой для разработки подобных решений.

3. Методика сбора политик конфиденциальности. Для сбора данного корпуса документов была адаптирована методика, предложенная в [17]. Ее выбор обусловлен следующими факторами. Во-первых, в большинстве работ несмотря на то, что приводится источник документов, информации, касающейся аспектов практической реализации, недостаточно для разработки программного инструмента. Во-вторых, многие источники данных, например проект Curlie [34], не позволяют собрать достаточного числа необходимых документов, поскольку русскоязычный сегмент Интернета в них представлен скудно. Данные, хранимые в веб-директориях, довольно редко обновляются, это делается волонтерами, поэтому гарантий получения актуальных данных нет. Кроме того, необходимость собрать пользовательские соглашения требует выполнения дополнительных действий с найденными страницами сайтов.

Таблица 1. Сравнительный анализ наборов данных политик безопасности

#	Название	Количество элементов	Источник данных	Аннотирование	Особенности
1	OPP-115 [23] (2016)	115	Amazon Alexa	Да	Исследование политики в рамках проекта “Usable Privacy Policy”. Аннотировано квалифицированными юристами, собственный метод аннотирования.
2	MAPS [24] (2019)	> 1 млн.	Google Play	Нет	
3	APP-350 [24] (2019)	350	MAPS	Да	
4	Princeton-Leuven Longitudinal Privacy Policy Dataset [32] (2021)	> 1 млн.	Amazon Alexa	Нет	Предназначен для оценки изменений в политике конфиденциальности с течением времени. Содержит политики за последние 20 лет, авторы также представили краулер.
5	A Large Publicly Available Corpus of Website Privacy Policies Based [33] (2020)	> 1.5 млн.	DMOZ	Нет	Формирование набора данных для дальнейших исследований. Для генерации использовался DMOZ, крупнейший сетевой каталог.
6	PrivaSeer: Corpus of Web Privacy Policies [35] (2020)	> 1 млн.	Данные Common Crawl, собранные с 2008 года	Нет	Часть проекта PrivaSeer, поисковой системы по политикам конфиденциальности. Разработана собственная методика создания набора данных, включающая сборщик документов, механизмы фильтрации и методы классификации и дедупликации.
7	PolicyQA: A Reading Comprehension Dataset for Privacy Policies [14] (2020)	25 017	OPP-115	Да	Часть проекта PrivacyCheck. Состоит из 25 017 примеров объяснения языка политики безопасности, дает ответы на 714 вопросов о политиках конфиденциальности.
8	PrivacyQA [40] (2019)	1 750	Google Play	Да	Состоит из 1 750 вопросов для 35 политик конфиденциальности на английском языке. Вопросы представлены категориями из схемы аннотирования к набору OPP-115.
9	IoTDataset [17] (2021)	592	Amazon, Walmart, Google, и производители IoT устройств	Нет	Предназначен для анализа политик конфиденциальности устройств Интернета Вещей. Создан на основе политик безопасности производителей IoT-устройств.
10	PrivacyLens (2023)	> 1 200	Amazon, Walmart, и производители IoT устройств	Нет	Собран в рамках проекта PrivacyLens. Предназначен для анализа политик конфиденциальности устройств Интернета Вещей. Создан на основе политик безопасности производителей IoT-устройств, позволяет сравнивать версии политик конфиденциальности.
11	PPinRussian dataset (данная работа)	9 051	mail.ru top, rambler top	Нет	Предназначен для анализа политик конфиденциальности на русском языке. Создан на основе политик конфиденциальности веб-сервисов.

Методика, разработанная авторами данной работы и детально описанная в [17], позволяет решить две задачи:

1. определение источников данных, удовлетворяющих заданным условиям отбора;

2. очистка документов от излишней HTML-разметки.

Она была успешно использована при создании корпуса политик конфиденциальности на английском языке IoTDataset, разработанных специально для IoT устройств, и применена в проекте PrivacyLens других авторов [38]. Однако, в отличие от оригинальной методики, исходной точкой для сбора данных послужили ссылки на веб-сервисы, полученные с платформ интернет-аналитики от компаний Mail.ru [41] и Rambler [42]. Ссылки были собраны по следующим категориям: “World”, “State”, “Business”, “House”, “Cars”, “Internet”, “Job”, “Computers”, “Rest”, “Culture”, “Science”, “WapSites”, “Sport”, “Mysterious”, “Industry”, “References”, “MassMedia”, “Humor”.

Таким образом, использованная методика состоит из следующих шагов:

– сбор гиперссылок на веб-страницы на платформе интернет-аналитики;

– поиск политик конфиденциальности на сайте веб-страниц;

– загрузка политик конфиденциальности в HTML-формате;

– очистка и подготовка политик конфиденциальности к дальнейшему анализу.

Под очисткой и подготовкой политик конфиденциальности понимается удаление HTML-разметки из текста документа и добавление Markdown-разметки для сохранения структуры текста.

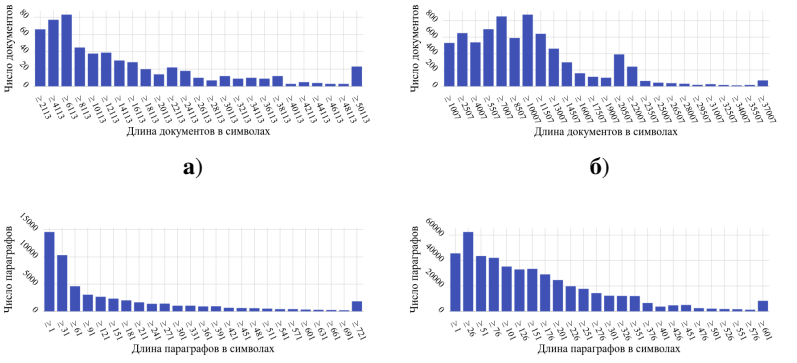
В ходе формирования набора данных было проанализировано 25 568 веб-сайтов, получено 21 585 (84%) необработанных политик конфиденциальности. После очистки текстов политик от HTML-разметки из набора данных были исключены документы длиной менее 1 000 символов. В результате было собрано 9 051 документов, полученный набор документов был назван PPinRussian. Следует отметить, что в ходе сбора политик конфиденциальности были выявлены случаи, когда вместо текста политики был размещен массивный рекламный блок или появлялось сообщение об отсутствии такого домена с предложением о его покупке/аренде.

4. Структурные особенности текстов политик конфиденциальности. Часто, чтобы определить методы и алгоритмы для дальнейшего анализа текстов, необходимо выполнить структурный анализ собранного корпуса документов. Под структурным анализом документов авторы понимают исследование значений длин документов, параграфов в символах, анализ наличия таких структурных элементов

форматирования, как списки, таблицы, разделы, заголовки. Понимание того, как организован текст в политиках конфиденциальности, позволяет более точно извлекать и связывать фрагменты текстов, описывающих различные аспекты сценариев использования персональных данных.

В данном разделе представлены результаты структурного анализа собранного корпуса политик конфиденциальности и выполнено его сравнение с набором политик конфиденциальности IoTDataset, написанных на английском языке.

На рисунке 1 представлены распределения длин политик и параграфов в символах.



а) **б**)

в) **г**)

Рис. 1. Гистограммы распределения длин: а, б) документов; в, г) параграфов для англоязычного и русскоязычного датасетов соответственно

Можно заметить, что наиболее распространенная длина документа в англоязычном корпусе составляет 7 000-8 000 символов. Это соответствует 3.9-4.5 стандартным страницам по 1 800 символов. Наиболее распространенными являются короткие параграфы длиной до 200 символов, состоящие из 2-3 предложений. В русскоязычном корпусе распределение длин документов отличается, наиболее часто встречаемыми являются документы длиной 8 000-9 000 символов, что соответствует 4.5-5 страницам, а также политики длиной 3 000-5 000 символов. Длины параграфов в корпусах также существенно различаются, причиной этому служат лингвистические особенности языков – в английском языке слова короче, предложения состоят из меньшего числа слов. Для политик конфиденциальности на русском языке характерны более длинные формулировки, особенно это верно для

вступительной части документа, где перечисляются различные названия нормативных документов, на основании которых производится обработка персональных данных, а также приводятся используемые в политике термины и определения.

Также в корпусах политик безопасности было проанализировано распределение таких структурных элементов текста, как таблицы, упорядоченные и неупорядоченные списки, параграфы и заголовки. На рисунке 2 показано распределение этих элементов в текстах политик конфиденциальности на русском и английском языках. Представленные данные позволяют сделать вывод о структуре политик безопасности. Так, обычно политики безопасности состоят из заголовков и параграфов. Наиболее часто используемыми элементами структуры помимо заголовков и параграфов являются нумерованные или ненумерованные списки и их элементы, в то время как таблицы используются редко. Русскоязычный документ в среднем состоит из 34.1 параграфов (85.1% документа), 3.8 заголовков (9.5% документа), 0.9 нумерованных списков (2.2% документа), 1.1 ненумерованных списков (2.7% документа) и 0.2 таблиц (0.5% документа). Документ на английском языке будет иметь похожую структуру: чуть меньше параграфов (31.5 параграфов или 44.9% документа), значительно больше заголовков – 33 заголовка (47.2% документа), 0.7 нумерованных списков (1% документа), 4.4 ненумерованных списков (6.3% документа) и 0.5 таблиц (0.7% документа).

Однако следует отметить, что такое большое число заголовков в английских документах может быть связано со сложностью подсчета данного элемента структурирования текста из-за большой вариативности HTML-разметки, используемой для их создания. Почти все сайты используют свой способ задания разделов, свои собственные правила нумерации разделов, заголовков и списков. На некоторых сайтах списки и заголовки нумеруются с помощью HTML-разметки, на других – нумерация задается вручную. В частности, компания Huawei предоставляет более 50 политик конфиденциальности для своих сервисов [43] и, хотя визуально все они имеют одинаковую структуру, для создания заголовков используется до 16 различных вариантов разметки HTML. Таким образом, даже в рамках одной компании не существует единой конвенции для оформления политик безопасности и их структурной компоновки, поэтому авторы посчитали заголовками строки длиной менее 100 символов и не содержащие маркеров “list item” (маркер элемента списка).

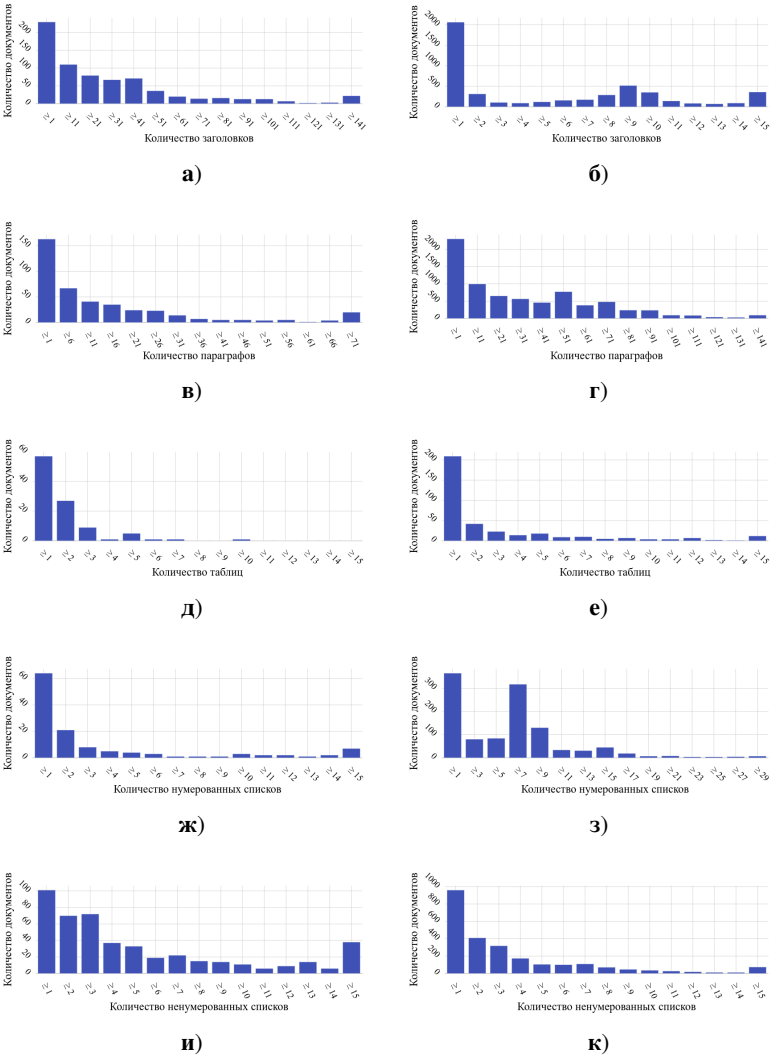


Рис. 2. Гистограммы распределения элементов структурирования данных в тексте политики безопасности: а, б) заголовки; в, г) параграфы; д, е) таблицы; ж, з) нумерованные списки; и, к) ненумерованные списки для англоязычного и русскоязычного датасетов соответственно

Очевидно, что такой подход не дает точных результатов, поскольку короткие параграфы, состоящие из одной строки, такие как контактная информация производителей, также относятся к заголовкам.

5. Семантический анализ корпусов политик конфиденциальности. При семантическом анализе политик конфиденциальности наибольший интерес представляет получение информации о том, какие характеристики сценариев сбора и обработки персональных данных описаны в документах, например, какие типы персональных данных собираются, в каких целях они собираются, на каких законных основаниях выполняется их обработка, каковы сроки их хранения, каким образом организована их защита и т.д.

В настоящей работе целью семантического анализа является извлечение тем, представленных в документе, и определение ключевых слов для них. Такие слова могут быть использованы для определения того, какие сценарии обработки персональных данных представлены в политике конфиденциальности.

Извлечение тем из корпуса политик конфиденциальности осуществлялось с помощью латентного размещения Дирихле (Latent Dirichlet Allocation, LDA), которое позволяет представить документ в виде множества тем, описанных комбинацией ключевых слов [44]. Для применения этого метода каждая политика была разбита на множество параграфов. Кроме того, было сделано предположение, что каждый параграф может содержать описание одного сценария использования персональных данных, т.е. выбиралась тема, аффилиация текста с которой была более некоторого заданного порога θ. Таким образом, выполненный анализ включал следующие шаги:

1. извлечение текстов параграфов из документов,
2. генерирование тематических моделей на основе анализа всего набора параграфов,
3. определение возможных характеристик сценариев использования персональных данных в соответствии с полученными семантическими моделями.

Перед применением LDA каждый параграф был преобразован в вектор слов. Затем были удалены стоп-слова, после чего была выполнена лемматизация текста с помощью библиотеки Python NLTK [45]. Исследования в [11] показали, что метрика TF-IDF (Term Frequency – Inverse Document Frequency) позволяет извлечь более подробную информацию о сценариях использования данных, так как эта модель векторизации текста присваивает большие веса словам, которые

встречаются реже, соответственно она позволяет более точно определить особенности содержания текста.

Число тем задается вручную, и данный параметр определяет качество тематического моделирования. Качество порождаемых моделей может быть оценено с помощью показателя когерентности порожденных тематических моделей. Когерентность показывает, насколько неслучайно слова, являющиеся значимыми для темы, появляются в тексте. В исследовании была использована метрика c_v, рассчитывающаяся как среднее от косинусных сходств s_{cos} векторов слов ($w_{n,k}$) в тексте и векторов, отражающих темы (w_k^*) (формула 1):

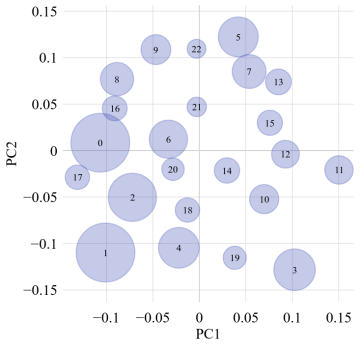
$$c_v = \frac{\sum_{k=1}^{K} \sum_{n=1}^{N} s_{cos}(w_{n,k}, w_k^*)}{N \cdot K}, \quad (1)$$

где N – количество слов, а K – количество тем. Результаты оценки приведены на рисунках 3(а) и 3(б). Оптимальное число семантических тем, на котором значение когерентности достигает максимума, для русскоязычного корпуса документов и англоязычного корпуса документов оказалось разным. Для русскоязычного корпуса PPinRussian оно равно 44, а для англоязычного IoTDataset – 23. На рисунках 3(в) и 3(г) представлена визуализация тематических кластеров параграфов, полученных в результате применения LDA и метода главных компонент (Principal Component Analysis, PCA). Радиус круга отражает число параграфов в кластере, т.е. чем больше параграфов в кластере, тем больше радиус. Кроме того, на рисунках 3(д) и 3(е) можно наблюдать кластеризацию с помощью алгоритма стохастического вложения соседей с t-распределением (t-distributed Stochastic Neighbor Embedding, t-SNE) по каждому документу с точки зрения их семантики, при этом можно заметить схожесть многих из них, что может быть связано с использованием шаблонов или составом аспектов в целом.

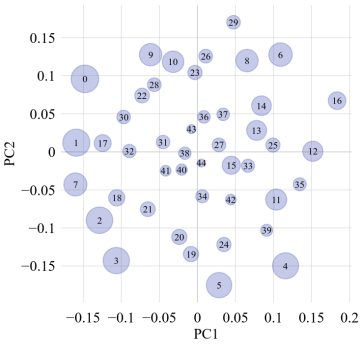
Очевидно, что и для корпуса русскоязычных документов и англоязычных документов есть кластеры, которые отличаются большим числом элементов в них, кроме того, они хорошо отделимы от других. Анализ таких кластеров показал, что к ним отнесены параграфы, которые содержат единообразную информацию, представленную одним и тем же набором слов, на основе которых может быть определена тема.

а)

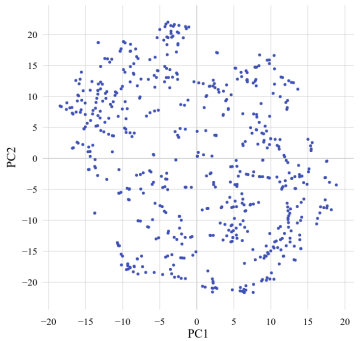
б)



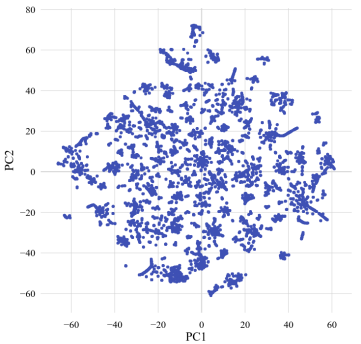
в)



г)



д)



е)

Рис. 3. Зависимость когерентности модели от количества выявляемых тем: а) датасет англоязычных политик; б) датасет русскоязычных политик; в) визуализация тематических кластеров для англоязычных документов IoTDataset; г) русскоязычных документов PPinRussian; д) визуализация семантики по каждому документу для IoTDataset; е) PPinRussian

Выделение тем, описанных характерными словами и их весами, в соответствии с показателем когерентности позволяет получить наибольшее количество таких тем, при этом среди них не наблюдается очевидных повторений. При получении тем с их количеством, заданным ниже, чем рекомендуемое в соответствии с когерентностью, произойдет утрата менее явных тем, отражающих специфические аспекты сценариев использования персональных данных. Кроме того, некоторые обобщенные темы имеют законченный смысл именно по совокупности включенных в них тем, но не обязательно пересекающихся по составу лексем, образующих их. Еще одной причиной формирования обобщенных тем является попытка получить осмысленное и описывающее некоторый аспект политики значение, которое в результате работы алгоритма LDA представляется в виде списка не связанных в единую языковую конструкцию лексем.

В таблице 2 в качестве примера представлены ключевые слова для тем 0, 1 и 2, определенные для русскоязычного корпуса документов.

Таблица 2. Ключевые слова и их веса для тем 0, 1 и 2

№	Тема 0		Тема 1		Тема 2	
	Вес	**Лексема**	**Вес**	**Лексема**	**Вес**	**Лексема**
1	0.031	сбор	0.037	электронный	0.011	ip-адрес
2	0.026	использование	0.033	адрес	0.010	выявление
3	0.021	хранение	0.027	почта	0.010	данный
4	0.021	персональный	0.016	телефон	0.008	персональный
5	0.018	предоставление	0.015	e-mail	0.008	статистика
6	0.017	операция	0.014	доставка	0.008	информация
7	0.016	данный	0.012	номер	0.007	сайт
8	0.016	средство	0.011	товар	0.007	законность
9	0.015	обновление	0.010	пользователь	0.007	пользователь
10	0.015	изменение	0.008	сайт	0.007	оператор
11	0.015	передача	0.007	информация	0.007	проблема
12	0.015	накопление	0.007	заказ	0.007	свой
13	0.014	уточнение	0.006	письмо	0.007	посетитель
14	0.014	систематизация	0.006	отправка	0.007	решение
15	0.013	удаление	0.006	информирование	0.007	цель
16	0.013	действие	0.005	посредством	0.006	использоваться
17	0.013	уничтожение	0.005	уведомление	0.006	технический
18	0.013	извлечение	0.005	услуга	0.006	осуществлять
19	0.012	обезличивание	0.005	мы	0.006	проводить
20	0.012	пользователь	0.005	связываться	0.005	актуализация

Из него следует, что тема 0 посвящена общим вопросам по сбору, обработке и использованию персональных данных, в теме 1 детализируются типы собираемых данных – электронный адрес, почта, телефон, а также кратко представлены возможные цели использования – доставка и информирование. В теме 2 также представлены типы

собираемых данных – IP-адрес, который в первую очередь используется для сбора статистики. Данные выводы получены путем ручного анализа параграфов, отнесенных к этим темам. Примерами других тем, представленных в документе являются особенности реализации обратной связи и уведомлений (тема 4), разрешение споров (тема 6), распространение персональных данных, в т.ч. их трансграничная передача (темы 17 и 39), цели использования персональных данных (тема 38), маркетинговые и новостные рассылки (тема 20), защита персональных данных (тема 42). Наиболее нетривиальной задачей оказалось определение семантических тем для кластеров небольшого объема, расположенных в центре графика рассеивания на рисунке 3(г). Выявленные ключевые слова достаточно общие, поэтому определить особенность той или иной темы не представлялось возможным, кроме того, параграфы, отнесенные к данным кластерам не имели четко выраженных общих концепций. В таблице 3 представлены примеры таких кластеров.

Таблица 3. Ключевые слова и их веса для тем 38, 41 и 42

№	Тема 38		Тема 41		Тема 42	
	Вес	**Лексема**	**Вес**	**Лексема**	**Вес**	**Лексема**
1	0.022	персональный	0.012	персональный	0.024	неправомерный
2	0.022	обработка	0.010	данный	0.018	защита
3	0.021	данный	0.010	пользователь	0.014	мера
4	0.016	цель	0.009	состояние	0.013	сайт
5	0.015	автоматизированный	0.009	сайт	0.013	случайный
6	0.013	техника	0.008	информация	0.013	копирование
7	0.013	вычислительный	0.008	категория	0.013	организационный
8	0.012	средство	0.008	оператор	0.013	персональный
9	0.010	помощь	0.007	данные	0.013	необходимый
10	0.010	несовместимый	0.007	получать	0.012	технический
11	0.009	информация	0.007	заказ	0.012	данный
12	0.007	данные	0.007	уведомление	0.011	информация
13	0.007	передача	0.007	обработка	0.011	администрация
14	0.007	допускаться	0.006	случай	0.011	принимать
15	0.006	заранее	0.006	субъект	0.011	пользователь
16	0.006	пользователь	0.005	лицо	0.010	блокирование
17	0.006	использование	0.005	услуга	0.010	доступ
18	0.006	сбор	0.005	товар	0.010	уничтожение
19	0.006	сайт	0.005	специальный	0.010	иной
20	0.006	получать	0.005	возврат	0.010	распространение

Поскольку многие темы оказались схожи между собой, было принято решение объединить их в группы, которые явно характеризуют особенности использование персональных данных. Были сформулированы следующие обобщенные темы: (1) термины и определения политики, (2) сбор трекинговых персональных данных, (3) сбор, обработка и хранение

персональных данных, (4) распространение и уничтожение персональных данных, (5) изменение персональных данных, (6) разрешение споров, (7) уведомление, маркетинг и персонализация, (8) цели обработки персональных данных, (9) защита персональных данных, (10) правовые основания обработки, (11) обновление политики безопасности, (12) согласие пользователя на обработку персональных. Объединение кластеров выполнялось с учетом извлеченных ключевых слов тем и их близости в пространстве проекций, после чего было проанализировано распределение обобщенных тем в корпусе документов. Для этого каждый параграф был отнесен к одному кластеру, если вероятность принадлежности была более или равна $\theta = 0.5$, если вероятность была ниже, то такой параграф исключался из дальнейшего анализа. Далее для параграфа определялась обобщенная группа. Результаты семантического анализа корпусов документов представлены в таблице 4.

Таблица 4. Распределение обобщенных тем в PPinRussian корпусе политик конфиденциальности на русском языке

№	**Обобщенная тема**	**Процент, %**
1	Термины и определения политики	12.1
2	Сбор трекинговых персональных данных	9.8
3	Сбор обработка и хранение персональных данных	20.4
4	Передача третьим лицам и уничтожение персональных данных	11.7
5	Изменение персональных данных	3.5
6	Разрешение споров	1.8
7	Уведомление, маркетинг, персонализация	4.4
8	Цели обработки персональных данных	9.4
9	Защита персональных данных	6.2
10	Правовые основания обработки персональных данных	12
11	Обновление политики безопасности	3.1
12	Согласие пользователя на обработку персональных данных	5.6

На рисунке 4 подробно показано распределение семантических тем в политиках конфиденциальности. Как и в предыдущем случае для каждого параграфа определялся тематический кластер, если порог принадлежности для этого кластера был $\theta \geqslant 0.5$, и обобщенная тема. Далее каждая политика конфиденциальности представлялась в виде вектора, содержащего число параграфов заданной темы, после чего все документы были разделены на кластеры в соответствии с распределением тем в текстах с помощью алгоритма кластеризации k-means. Для построения графика на рисунке 4 все политики были упорядочены сначала по вычисленным кластерам, а затем по количеству параграфов в документе. Таким образом, ось x соответствует номеру документа в упорядоченном списке всех документов, и, следовательно,

ширина столбца диаграммы пропорциональна количеству документов в соответствующем кластере. Ось y показывает число параграфов, отнесенных к каждой обобщенной теме. Таким образом, построенная диаграмма показывает среднее количество параграфов в кластере документов с заданной темой, а цветные участки каждого столбца отражают количество и соотношение тем в каждом конкретном кластере.

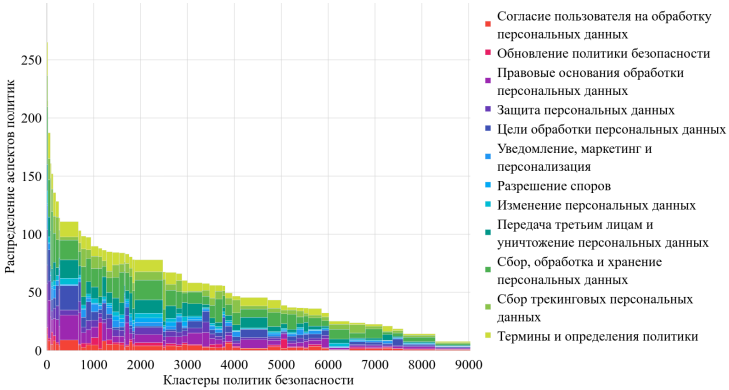


Рис. 4. Результат кластеризации русскоязычных политик безопасности с учетом распределения параграфов по обобщенным темам

Аналогичным образом были проанализированы темы англоязычного корпуса IoTDataset. Число семантических кластеров в этом корпусе почти в 2 раза меньше по сравнению с русскоязычным, что, возможно, объясняется как особенностями языка, так и тем, что в состав данного корпуса входят только политики для устройств Интернета Вещей. В большинстве случаев определить тему по ключевым словам было достаточно просто, исключение составили лишь некоторые кластеры, например кластеры 5 и 17. Примеры ключевых слов для некоторых тематических кластеров представлены в таблице 5. Например, в теме 3 обсуждаются права особой аудитории – граждан США в Калифорнии, которые защищены дополнительными законодательными актами в области обработки персональных данных, в теме 4 представлены меры безопасности по защите персональных данных, тема 6 описывает особенности обработки персональных данных специальных категорий пользователей – несовершеннолетних. В отличие от этих тем, в темах 5 и 17, приведенных в таблице 6, нет ключевых слов с ярко выраженным вкладом.

Таблица 5. Ключевые слова и их веса для тем 3, 4 и 5]

№	Тема3		Тема4		Тема5	
	Вес	Лексема	Вес	Лексема	Вес	Лексема
1	0.023	california	0.031	security	0.013	information
2	0.013	privacy	0.012	term	0.011	home
3	0.010	resident	0.011	data	0.008	may
4	0.009	right	0.009	information	0.008	access
5	0.009	notice	0.008	use	0.007	use
6	0.008	information	0.007	personal	0.007	personal
7	0.007	service	0.007	privacy	0.007	way
8	0.007	state	0.007	right	0.007	cooky
9	0.006	policy	0.006	condition	0.006	data
10	0.006	use	0.006	service	0.006	u
11	0.006	may	0.006	legal	0.005	service
12	0.005	united	0.006	sale	0.005	collect
13	0.005	cooky	0.005	b	0.005	send
14	0.005	personal	0.005	described	0.005	product
15	0.005	product	0.005	policy	0.005	provide
16	0.005	purpose	0.004	technical	0.005	email
17	0.004	website	0.004	notice	0.005	following
18	0.004	change	0.004	de	0.005	purchase
19	0.004	law	0.003	may	0.005	message
20	0.004	consumer	0.003	product	0.004	order

Таблица 6. Ключевые слова и их веса для тем 6 и 17

№	Тема 6		Тема 17	
	Вес	Лексема	Вес	Лексема
1	0.014	address	0.008	information
2	0.014	child	0.007	contract
3	0.009	information	0.007	data
4	0.008	childrens	0.006	personal
5	0.008	privacy	0.006	last
6	0.007	service	0.006	order
7	0.007	policy	0.006	service
8	0.006	website	0.005	payment
9	0.006	personal	0.005	privacy
10	0.006	site	0.005	performance
11	0.006	age	0.005	use
12	0.006	data	0.005	collect
13	0.005	may	0.005	may
14	0.005	u	0.004	policy
15	0.005	collect	0.004	u
16	0.005	name	0.004	site
17	0.005	use	0.004	transaction
18	0.005	ip	0.004	access
19	0.005	e-mail	0.004	name
20	0.005	term	0.004	website

Множества ключевых слов этих тем достаточно похожи, отличие заключается в том, что в теме 5 упоминаются данные, позволяющие

отслеживать действия пользователей в сети (cookies), а в теме 17 включены слова “платежи” (payment) и “транзакции” (transaction), что позволяет предположить, что речь идет о финансовых данных.

Для англоязычного корпуса также были сформулированы обобщенные темы: (1) вопросы пользователя по политике безопасности, (2) сторонние веб-сайты, (3) особая аудитория, (4) защита персональных данных, (5) сбор персональных данных, (6) сбор данных, позволяющих отслеживать поведения пользователя на веб-сайте, (7) распространение персональных (передача третьим лицам), (8) хранение персональных данных, (9) персонализация и маркетинг, (10) изменение политики безопасности. На рисунке 5 представлено распределение обобщенных тем в корпусе IoTDataset, а в таблице 7 показан результат кластеризации политик безопасности с учетом распределения параграфов по обобщенным темам.

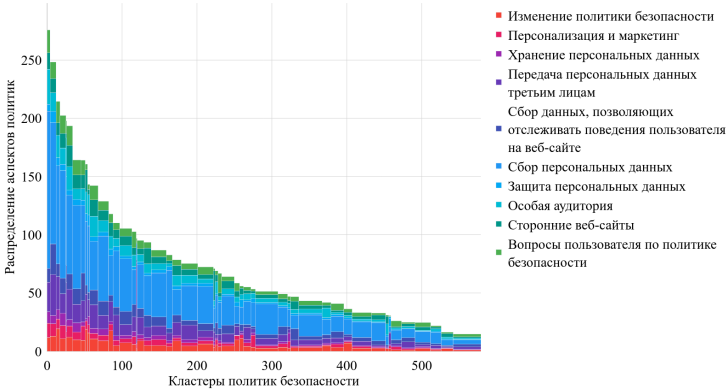


Рис. 5. Результат кластеризации англоязычных политик безопасности с учетом распределения параграфов по обобщенным темам

Очевидно, что сформированные обобщенные темы для корпуса политик на русском языке достаточно сильно отличаются от обобщенных тем корпуса политик на английском языке. Отлично также их количественное присутствие в документах. Некоторые темы присутствуют в обоих корпусах, в частности, это темы, характеризующие сбор персональных данных, данных по отслеживанию действий пользователя на веб-сайте, а также передаче третьим лицам. В обоих корпусах присутствует информация о том, какие данные собираются в

целях персонализации или маркетинга, какие меры осуществляются для защиты персональных данных.

Таблица 7. Распределение обобщенных тем в IoTDataset корпусе политик конфиденциальности на английском языке

№	Обобщенная тема	Процент, %
1	Вопросы пользователя по политике безопасности	7.5
2	Сторонние веб-сайты	7.1
3	Особая аудитория	7.7
4	Защита персональных данных	3.4
5	Сбор персональных данных	39.4
6	Сбор данных, позволяющих отслеживать поведение пользователя на веб-сайте	8.4
7	Передача персональных данных третьим лицам	11.6
8	Хранение персональных данных	3.3
9	Персонализация и маркетинг	4.6
10	Изменение политики безопасности	7

Ключевые отличия заключаются в том, что в политиках на русском языке отдельно прописываются положения по урегулированию споров и механизмов получения согласия на использование персональных данных. Особое внимание уделяется вопросам уничтожения персональных данных (до 10% от общего числа параграфов), а также прописывается ответственность пользователя за изменение персональных данных. Следует также отметить наличие достаточно объемной преамбулы в документах, где разъясняются общие термины и определения (более 10% от общего числа параграфов). В англоязычных политиках отдельно указываются особенности обработки данных, которые принадлежат субъектам, имеющим специальную категорию, например, несовершеннолетние или граждане стран, на территории которых действуют дополнительные законодательные акты в области защиты персональных данных. Прописываются особенности хранения персональных данных, часто с указанием конкретной длительности, а также детально описываются механизмы уведомления субъектов персональных данных в случае изменения политик безопасности (7% от общего числа параграфов). В политиках на английском языке значительную часть документа (39% от общего числа параграфов) так же занимает описание, каким образом осуществляется сбор данных, детализируются их типы. Такие отличия в политиках, в первую очередь, связаны с законодательными актами, действующими в Российской Федерации и/или на территориях других стран. Например, последние изменения в ФЗ “О персональных данных” № 152-ФЗ в части передачи информации третьим лицам и наличия письменного согласия субъекта персональных данных нашли отражение в политиках безопасности,

данные темы представлены достаточно объемно, занимая суммарно до 15% от объема текста (в параграфах), при этом описание сценариев хранения, сбора и обработки данных занимает всего 25% объема текста документа, что может говорить об отсутствии деталей о том, какие типы данных собираются, какие механизмы для сбора используются, и каковы сроки их хранения, что влияет на “прозрачность” и понятность самих текстов документов.

6. Заключение. Политика конфиденциальности – это официальный способ информирования пользователей о том, какие персональные данные собираются и как эти данные обрабатываются. Обычно, эти политики вызывают затруднения при чтении и понимании. В результате пользователи пропускают их и не понимают, кто и как использует их персональные данные и каковы риски нарушения их конфиденциальности. Таким образом, задача автоматизированного анализа политик безопасности, написанных на естественном языке, и их представления в прозрачной форме является весьма актуальной. Это особенно важно в настоящее время в связи с требованиями правовых документов к прозрачности обработки персональных данных с одной стороны, и стремительным развитием Интернета вещей и веб-сервисов с другой стороны. Каждый день люди используют множество “умных” устройств и сервисов, которые собирают большое количество разнообразных персональных данных, включая такие чувствительные из них, как данные о здоровье и биометрические данные, и не учитывают связанные с этим риски нарушения их конфиденциальности.

В настоящее время исследователи предложили различные подходы на основе машинного обучения для анализа политик безопасности, написанных на естественном языке, и представления их в прозрачной форме. Применение таких подходов требует использования аннотированных наборов данных политик безопасности для обучения модели анализа с целью анализа особенностей сценариев сбора и обработки данных. Авторы данной статьи провели сравнительный анализ существующих наборов данных и выделили их отличительные особенности, такие как год создания, объем и наличие аннотаций.

В статье представлены основные характеристики сформированного корпуса документов, включая результаты семантического моделирования. Семантический анализ выявил различия в темах, представленных в созданном наборе данных, по сравнению с набором данных IoTDataset. Например, были выявлены темы, связанные с правовыми основаниями сбора, терминами и определениями и разрешением споров между сторонами. Последнее означает, что при использовании для обучения

моделей анализа с соответствующими аннотациями этот набор данных может повысить точность обнаружения и рассуждений о различных аспектах сценариев использования персональных данных, включая аспекты, связанные с обязательствами обработчиков данных.

Будущие исследования будут включать в себя разработку автоматизированной проверки собранных документов, дальнейшее автоматизированное обнаружение различных аспектов использования персональных данных и расчет рисков конфиденциальности, связанных с использованием устройств или веб-сайтов.

Литература

1. Исследование утечек информации в отраслях за три года. URL: https://www.infowatch.ru/analytics/analitika/issledovaniye-utechek-informatsii-v-otraslyakh-za-tri-goda (дата обращения 20.05.2024).
2. Американские власти оштрафовали Avast за распространение персональных данных пользователей. URL: https://xakep.ru/2024/02/26/avast-ftc (дата обращения 20.05.2024).
3. Number of Internet of Things (IoT) connections worldwide from 2022 to 2023, with forecasts from 2024 to 2033. URL: https://www.statista.com/statistics/1183457/iot-connected-devices-worldwide (дата обращения 20.05.2024).
4. Самодолов А.П., Самодолова О.А., Николаенко Е.В. Особенности развития “умных домов” в России // Вестник ЮУрГУ. Серия: Строительство и архитектура. 2021. Т. 21. № 2. С. 78–85.
5. Отчет об уязвимостях в устройствах Интернета Вещей. URL: https://www.cnet.com/home/security/your-home-security-camera-could-be-hacked-so-treat-it-that-way (дата обращения 20.05.2024).
6. Mitigating Smart Meter Security Risk: A Privacy-preserving Approach. URL: https://eepower.com/technical-articles/mitigating-smart-meter-security-risk-a-privacy-preserving-approach/ (дата обращения 20.05.2024).
7. Alanazi F., Kim J., Cotilla-Sanchez E. Load Oscillating Attacks of Smart Grids: Vulnerability Analysis // IEEE Access. 2023. vol. 11. pp. 36538–36549. DOI: 10.1109/access.2023.3266249.
8. Steinfeld N. “I agree to the terms and conditions”: (How) do users read privacy policies online? An eye-tracking experiment // Computers in Human Behavior. 2016. vol. 55. part B. pp. 992–1000. DOI: 10.1016/j.chb.2015.09.038.
9. Karegar F., Pettersson J.S., Fischer-Hubner S. The Dilemma of User Engagement in Privacy Notices: Effects of Interaction Modes and Habituation on User Attention // ACM Transactions on Privacy and Security (TOPS). 2020. vol. 23. no. 1. pp. 1–38. DOI: 10.1145/3372296.
10. Регламент Европейского регулирования персональных данных. URL: http://data.europa.eu/eli/reg/2016/679/oj (дата обращения 20.05.2024).
11. Harkous H., Fawaz K, Lebret R, Schaub F, Shin KG, Aberer K. Polisis: automated analysis and presentation of privacy policies using deep learning // Proceedings of the 27th USENIX Security Symposium. 2018. pp. 531–548.
12. Novikova E., Doynikova E., Kotenko I. P2Onto: Making Privacy Policies Transparent // Computer Security, CyberICPS SECPRE ADIoT 2020, Proceedings of the International Workshop on Attacks and Defenses for Internet-of-Things. 2020. pp. 235–252.

13. Kuznetzov M., Novikova E. Towards application of text mining techniques to the analysis of the privacy policies // Proceedings of the 10th Mediterranean Conference on Embedded Computing. 2021. pp. 1–4. DOI: 10.1109/meco52532.2021.9460130.
14. Ahmad W., Chi J., Tian Y., Chang K.-W. PolicyQA: A Reading Comprehension Dataset for Privacy Policies // Proceedings of the Findings of the Association for Computational Linguistics (EMNLP). 2020. pp. 743–749.
15. Harkous H., et al. Polisis: automated analysis and presentation of privacy policies using deep learning // Proceedings of the 27th USENIX Conference on Security Symposium. 2018. pp. 531–548.
16. Zaeem R.N., German R.L., Barber K.S. PrivacyCheck: Automatic Summarization of Privacy Policies Using Data Mining // ACM Transactions on Internet Technology. 2018. vol. 18. vol. 4. DOI: 10.1145/3127519
17. Kuznetsov M., et al. Privacy Policies of IoT Devices: Collection and Analysis // Sensors. 2022. vol. 22. no. 5. DOI: 10.3390/s22051838.
18. Правила защиты конфиденциальности детей в Интернете. URL: https://www.ftc.gov/legal-library/browse/rules/childrens-online-privacy-protection-rule-coppa (дата обращения 20.05.2024).
19. Palmirani M., Martoni M., Rossi A., Bartolini C., Robaldo L. Legal ontology for modelling GDPR concepts and norms // Legal Knowledge and Information Systems. Amsterdam: IOS Press. 2018. vol. 313. pp. 91–100. DOI: 10.3233/978-1-61499-935-5-91.
20. Pandit H.J., O'Sullivan D., Lewis D. An Ontology Design Pattern for Describing Personal Data in Privacy Policies // 9th Workshop on Ontology Design and Patterns. 2018. vol. 2195. pp. 29–39.
21. Oltramari A., Piraviperumal D., Schaub F., Wilson S., Cherivirala S., Norton T.B., Russel N.C., Story P., Reidenberg, Sadeh N. PrivOnto: a semantic framework for the analysis of privacy policies // Semantic Web. 2018. vol. 9. no. 2. pp. 185–203.
22. Cano-Benito J., Cimmino A., Garcia-Castro R. Toward the ontological modeling of smart contracts: A solidity use case // IEEE Access. 2021. vol. 9. pp. 140156–140172. DOI: 10.1109/access.2021.3115577.
23. Wilson Ah., et al. The Creation and Analysis of a Website Privacy Policy Corpus // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016. pp. 1330–1340. DOI: 10.18653/v1/P16-1126.
24. Zimmeck S., et al. MAPS: scaling privacy compliance analysis to a million apps // In Proceedings on Privacy Enhancing Technologies 2019. vol. 3. pp. 66–86. DOI: 10.2478/popets-2019-0037.
25. Kumar V.H., Iyengar R., Nisal N., Feng Y., Habib H., Story P., Cherivirala S., Nagan M., Cranor L., Wilson S., Schaud F., Sadeh N. Finding a Choice in a Haystack: Automatic Extraction of Opt-Out Statements from Privacy Policy Text // Proceedings of The Web Conference. 2020. pp. 1943–1954. DOI: 10.1145/3366423.3380262.
26. Hosseini M.B., Heaps J., Slavin R., Niu J., Breaux T. Ambiguity and Generality in Natural Language Privacy Policies // IEEE 29th International Requirements Engineering Conference (RE). 2021. pp. 70–81. DOI: 10.1109/RE51729.2021.00014.
27. Hosseini M.B., Breaux T., Slavin R., Niu J., Wang X. Analyzing Privacy Policies through Syntax-Driven Semantic Analysis of Information Types // Information and Software Technology Journal. 2021. vol. 138. DOI: 10.1016/j.infsof.2021.106608.
28. Веб-страница проекта Usable Privacy Policy. URL: https://usableprivacy.org (дата обращения 21.05.2024).
29. Веб-сайт Amazon Alexa. URL: https://www.alexa.com (дата обращения 22.05.2024).

30. Poplavska E., Norton T.B., Wilson S., Sadeh N. From Prescription to Description: Mapping the GDPR to a Privacy Policy Corpus Annotation Scheme // Proceedings of the 33rd International Conference on Legal Knowledge and Information Systems. 2020. pp. 243–246.
31. Веб-сайт сервиса Google Play. URL: https://play.google.com/store (дата обращения 24.05.2024).
32. Amos R., Acar G., Kshirsagar M., Narayanan A., Mayer J. Privacy Policies over Time: Curation and Analysis of a Million-Document Dataset // Proceedings of the Web Conference. 2021. pp. 2165–2176. DOI: 10.1145/3442381.3450048.
33. Zaeem R.N., Barber K.S. A Large Publicly Available Corpus of Website Privacy Policies Based on DMOZ // In Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy. 2021. pp. 143–148. DOI: 10.1145/3422337.3447827.
34. Веб-директория Curlie. URL: https://curlie.org (дата обращения 26.05.2024).
35. Srinath M., Wilson S., Giles C. Privacy at Scale: Introducing the PrivaSeer Corpus of Web Privacy Policies // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. 2021. pp. 6829–6839. DOI: 10.18653/v1/2021.acl-long.532.
36. Веб-сайт Amazon. URL: https://www.amazon.com (дата обращения 26.05.2024).
37. Веб-сайт Walmart. URL: https://www.walmart.com/ (дата обращения 28.05.2024).
38. Hamid A., Samidi H.R., Finin T., Pappachan P., Yus R. PrivacyLens: A Framework to Collect and Analyze the Landscape of Past, Present, and Future Smart Device Privacy Policies // arXiv pradprint arXiv.2308.05890. 2023.
39. Polisis. URL: https://www.epfl.ch/labs/lsir/polisis/ (дата обращения 02.06.2024).
40. Ravichander A., Black A., Wilson S., Norton T., Sadeh N. Question Answering for Privacy Policies: Combining Computational and Legal Perspectives // Proceedings of 2019 Conference on Empirical Methods in Natural Language Processing. 2019. pp. 4947–4958. DOI: 10.18653/v1/D19-1500.
41. Веб-сайт аналитической площадки Mail.ru Top. URL: https://top.mail.ru (дата обращения 02.06.2024).
42. Веб-сайт аналитической площадки Rambler Top-100. URL: https://top100.rambler.ru (дата обращения 02.06.2024).
43. Политика безопасности компании Huawei. URL: https://www.huawei.com/eu/privacy-policy (дата обращения 02.06.2024).
44. Blei D., Ng A., Jordan M. Latent Dirichlet Allocation // Journal of Machine Learning Research. 2003. vol. 3. pp. 993–1022.
45. Веб-сайт библиотеки NLTK. URL: https://www.nltk.org (дата обращения 02.06.2024).

Кузнецов Михаил Дмитриевич — младший научный сотрудник, лаборатория проблем компьютерной безопасности, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук (СПб ФИЦ РАН). Область научных интересов: компьютерные технологии, методы онтологического моделирования и формализации текста. Число научных публикаций — 23. mkuznetsov7991@gmail.com; 14 линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(911)830-0669.

Новикова Евгения Сергеевна — канд. техн. наук, старший научный сотрудник, лаборатория проблем компьютерной безопасности, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук (СПб ФИЦ РАН). Область научных интересов: безопасность информационных систем, обнаружение аномалий методами машинного обучения, конфиденциальность данных. Число научных публикаций — 60. novikova@comsec.spb.ru; 14 линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(812)328-7181.

DOI 10.15622/ia.24.1.7

M. KUZNETSOV, E. NOVIKOVA

CORPUS OF PRIVACY POLICIES FOR WEB SERVICES AND INTERNET OF THINGS DEVICES FOR ANALYZING THE AWARENESS OF PERSONAL DATA SUBJECTS

Kuznetsov M.D., Novikova E.S. **Corpus of privacy policies for web services and Internet of Things devices for analyzing the awareness of personal data subjects.**

Abstract. Information about what personal data is collected and processed by various devices and digital services is presented in privacy policies, however, as studies show, users rarely read them and, as a result, do not realize which data security risks associated with the processing of personal data arise. The solution to the problem of increasing the awareness of personal data subjects is associated with the development of decision support methods that present privacy policies in a form that is easier to understand, for example, in the form of quantitative risk assessments and pictograms. Their development requires a structured and marked-up corpus of documents. This paper systematizes the corpora of privacy policies that are in the open access and shows their distinctive characteristics, such as the year of creation, volume and presence of annotations. A description of a new corpus of documents written in Russian is also presented, the results of a structural and semantic analysis of the collected security policies are given, and a comparison with the corpus of privacy policies written in English is made. It has been shown that the description of scenarios for storing, collecting and processing data in documents in Russian accounts for only 25% of the volume of the document text, which may indicate a lack of details about what types of data are collected, what mechanisms are used for collection, and what are the storage periods, which affects the "transparency" of the use of personal data.

Keywords: personal data, privacy policies, document corpus, semantic analysis, Latent Dirichlet allocation.

References

1. Issledovanie utechek informacii v otrasljah za tri goda [Study of information leaks in industries over three years]. Avaliable at: https://www.infowatch.ru/analytics/analitika/issledovaniye-utechek-informatsii-v-otraslyakh-za-tri-goda (accessed 20.05.2024). (In Russ.).
2. Amerikanskie vlasti oshtrafovali Avast za rasprostranenie personal'nyh dannyh pol'zovatelej [American authorities fined Avast for distributing user personal data]. Available at: https://xakep.ru/2024/02/26/avast-ftc (accessed 20.05.2024). (In Russ.).
3. Number of Internet of Things (IoT) connections worldwide from 2022 to 2023, with forecasts from 2024 to 2033. Available at: https://www.statista.com/statistics/1183457/iot-connected-devices-worldwide (accessed 20.05.2024).
4. Samodolov A.P., Samodolova O.A., Nikolaenko E.V. [Features of the development of "smart homes" in Russia]. Vestnik JuUrGU. Serija: Stroitel'stvo i arhitektura – Bulletin of SUSU. Series: Construction and architecture. 2021. vol. 21. no. 2. pp. 78–85. (In Russ.).
5. Internet of Things Device Vulnerability Report. Available at: https://www.cnet.com/home/security/your-home-security-camera-could-be-hacked-so-treat-it-that-way (accessed 20.05.2024).
6. Mitigating Smart Meter Security Risk: A Privacy-preserving Approach. Available at: https://eepower.com/technical-articles/mitigating-smart-meter-security-risk-a-privacy-preserving-approach/ (accessed 20.05.2024).

7. Alanazi F., Kim J., Cotilla-Sanchez E. Load Oscillating Attacks of Smart Grids: Vulnerability Analysis. IEEE Access. 2023. vol. 11. pp. 36538–36549. DOI: 10.1109/access.2023.3266249.
8. Steinfeld N. “I agree to the terms and conditions”: (How) do users read privacy policies online? An eye-tracking experiment. Computers in Human Behavior. 2016. vol. 55. part B. pp. 992–1000. DOI: 10.1016/j.chb.2015.09.038.
9. Karegar F., Pettersson J.S., Fischer-Hubner S. The Dilemma of User Engagement in Privacy Notices: Effects of Interaction Modes and Habituation on User Attention. ACM Transactions on Privacy and Security (TOPS). 2020. vol. 23. no. 1. pp. 1–38. DOI: 10.1145/3372296.
10. General data protection regulation. Available at: http://data.europa.eu/eli/reg/2016/679/oj (accessed 20.05.2024).
11. Harkous H., Fawaz K, Lebret R, Schaub F, Shin KG, Aberer K. Polisis: automated analysis and presentation of privacy policies using deep learning. Proceedings of the 27th USENIX Security Symposium. 2018. pp. 531–548.
12. Novikova E., Doynikova E., Kotenko I. P2Onto: Making Privacy Policies Transparent. Computer Security, CyberICPS SECPRE ADIoT 2020, Proceedings of the International Workshop on Attacks and Defenses for Internet-of-Things. 2020. pp. 235–252.
13. Kuznetzov M., Novikova E. Towards application of text mining techniques to the analysis of the privacy policies. Proceedings of the 10th Mediterranean Conference on Embedded Computing. 2021. pp. 1–4. DOI: 10.1109/meco52532.2021.9460130.
14. Ahmad W., Chi J., Tian Y., Chang K.-W. PolicyQA: A Reading Comprehension Dataset for Privacy Policies. Proceedings of the Findings of the Association for Computational Linguistics (EMNLP). 2020. pp. 743–749.
15. Harkous H., et al. Polisis: automated analysis and presentation of privacy policies using deep learning. Proceedings of the 27th USENIX Conference on Security Symposium. 2018. pp. 531–548.
16. Zaeem R.N., German R.L., Barber K.S. PrivacyCheck: Automatic Summarization of Privacy Policies Using Data Mining. ACM Transactions on Internet Technology. 2018. vol. 18. vol. 4. DOI: 10.1145/3127519.
17. Kuznetsov M., et al. Privacy Policies of IoT Devices: Collection and Analysis. Sensors. 2022. vol. 22. no. 5. DOI: 10.3390/s22051838.
18. Children’s Online Privacy Protection Rule. Available at: https://www.ftc.gov/legal-library/browse/rules/childrens-online-privacy-protection-rule-coppa (accessed 20.05.2024).
19. Palmirani M., Martoni M., Rossi A., Bartolini C., Robaldo L. Legal ontology for modelling GDPR concepts and norms. Legal Knowledge and Information Systems. Amsterdam: IOS Press. 2018. vol. 313. pp. 91–100. DOI: 10.3233/978-1-61499-935-5-91.
20. Pandit H.J., O’Sullivan D., Lewis D. An Ontology Design Pattern for Describing Personal Data in Privacy Policies. 9th Workshop on Ontology Design and Patterns. 2018. vol. 2195. pp. 29–39.
21. Oltramari A., Piraviperumal D., Schaub F., Wilson S., Cherivirala S., Norton T.B., Russel N.C., Story P., Reidenberg, Sadeh N. PrivOnto: a semantic framework for the analysis of privacy policies. Semantic Web. 2018. vol. 9. no. 2. pp. 185–203.
22. Cano-Benito J., Cimmino A., Garcia-Castro R. Toward the ontological modeling of smart contracts: A solidity use case. IEEE Access. 2021. vol. 9. pp. 140156–140172. DOI: 10.1109/access.2021.3115577.
23. Wilson Ah., et al. The Creation and Analysis of a Website Privacy Policy Corpus. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016. pp. 1330–1340. DOI: 10.18653/v1/P16-1126.

24. Zimmeck S., et al. MAPS: scaling privacy compliance analysis to a million apps. In Proceedings on Privacy Enhancing Technologies 2019. vol. 3. pp. 66–86. DOI: 10.2478/popets-2019-0037.
25. Kumar V.H., Iyengar R., Nisal N., Feng Y., Habib H., Story P., Cherivirala S., Nagan M., Cranor L., Wilson S., Schaud F., Sadeh N. Finding a Choice in a Haystack: Automatic Extraction of Opt-Out Statements from Privacy Policy Text. Proceedings of The Web Conference. 2020. pp. 1943–1954. DOI: 10.1145/3366423.3380262.
26. Hosseini M.B., Heaps J., Slavin R., Niu J., Breaux T. Ambiguity and Generality in Natural Language Privacy Policies. IEEE 29th International Requirements Engineering Conference (RE). 2021. pp. 70–81. DOI: 10.1109/RE51729.2021.00014.
27. Hosseini M.B., Breaux T., Slavin R., Niu J., Wang X. Analyzing Privacy Policies through Syntax-Driven Semantic Analysis of Information Types. Information and Software Technology Journal. 2021. vol. 138. DOI: 10.1016/j.infsof.2021.106608.
28. Usable Privacy Policy website. Available at: https://usableprivacy.org (accessed 21.05.2024).
29. Amazon Alexa website. Available at: https://www.alexa.com (accessed 22.05.2024).
30. Poplavska E., Norton T.B., Wilson S., Sadeh N. From Prescription to Description: Mapping the GDPR to a Privacy Policy Corpus Annotation Scheme. Proceedings of the 33rd International Conference on Legal Knowledge and Information Systems. 2020. pp. 243–246.
31. Google Play website. Available at: https://play.google.com/store (accessed 24.05.2024).
32. Amos R., Acar G., Kshirsagar M., Narayanan A., Mayer J. Privacy Policies over Time: Curation and Analysis of a Million-Document Dataset. Proceedings of the Web Conference. 2021. pp. 2165–2176. DOI: 10.1145/3442381.3450048.
33. Zaeem R.N., Barber K.S. A Large Publicly Available Corpus of Website Privacy Policies Based on DMOZ. In Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy. 2021. pp. 143–148. DOI: 10.1145/3422337.3447827.
34. Curlie web directory. Available at: https://curlie.org (accessed 26.05.2024).
35. Srinath M., Wilson S., Giles C. Privacy at Scale: Introducing the PrivaSeer Corpus of Web Privacy Policies. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. 2021. pp. 6829–6839. DOI: 10.18653/v1/2021.acl-long.532.
36. Amazon website. Available at: https://www.amazon.com (accessed 26.05.2024).
37. Walmart website. Available at: https://www.walmart.com/ (accessed 28.05.2024).
38. Hamid A., Samidi H.R., Finin T., Pappachan P., Yus R. PrivacyLens: A Framework to Collect and Analyze the Landscape of Past, Present, and Future Smart Device Privacy Policies. arXiv pradprint arXiv.2308.05890. 2023.
39. Polisis. Available at: https://www.epfl.ch/labs/lsir/polisis/ (accessed 02.06.2024).
40. Ravichander A., Black A., Wilson S., Norton T., Sadeh N. Question Answering for Privacy Policies: Combining Computational and Legal Perspectives. Proceedings of 2019 Conference on Empirical Methods in Natural Language Processing. 2019. pp. 4947–4958. DOI: 10.18653/v1/D19-1500.
41. Web analytics platform Mail.ru Top. Available at: https://top.mail.ru (accessed 02.06.2024).
42. Web analytics platform Rambler Top-100. Available at: https://top100.rambler.ru (accessed 02.06.2024).
43. Huawei privacy policy. Available at: https://www.huawei.com/eu/privacy-policy (accessed 02.06.2024).

44. Blei D., Ng A., Jordan M. Latent Dirichlet Allocation. Journal of Machine Learning Research. 2003. vol. 3. pp. 993–1022.

45. NLTK library website. Available at: https://www.nltk.org (accessed 02.06.2024).

Kuznetsov Mikhail — Junior researcher, Laboratory of computer security problems, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS). Research interests: computer technologies, methods of ontological modeling and text formalization. The number of publications — 23. mkuznetsov7991@gmail.com; 39, 14-th Line V.O., 199178, St. Petersburg, Russia; office phone: +7(911)830-0669.

Novikova Evgenia — Ph.D., Senior researcher, Laboratory of computer security problems, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS). Research interests: privacy and personal data security, privacy-preserving computations, and machine learning-based anomaly and intrusion detection. The number of publications — 60. novikova@comsec.spb.ru; 39, 14-th Line V.O., 199178, St. Petersburg, Russia; office phone: +7(812)328-7181.