DOI 10.36622/VSTU.2023.4.26.001
УДК 004.056

ЦЕЛЕПОЛАГАНИЕ ПРОЕКТНОЙ ДЕЯТЕЛЬНОСТИ ПО СОЗДАНИЮ ИНСТРУМЕНТАРИЯ АВТОМАТИЗИРОВАННОГО ВЫЯВЛЕНИЯ И РИСК-АНАЛИЗА ДЕСТРУКТИВНЫХ КОНТЕНТОВ, АФФИЛИРОВАННЫХ С ПЕРСОНАЛОМ КОРПОРАЦИЙ

А.Г. Остапенко, И.А. Боков, С.В. Лихобабин, Д.С. Ясенко, Е.Ю. Чапурин

Рассматриваются видео, аудио и графические контенты социальных сетей как фактор обеспечения информационной безопасности корпораций. Осуществляется целеполагание проектной деятельности по созданию автоматизированного инструментария выявления и риск-анализа вышеуказанных контентов, включая парсинг ресурсов соцсетей, селекцию собранных контентов по признакам деструктивности и их риск-анализ для выработки рекомендаций по разграничению доступа к корпоративной информации. Оценивается актуальность проектной деятельности, исследуются аналоги, предлагаются архитектура и алгоритмы создаваемого инструментария. Формулируются имеющиеся противоречия, вытекающие из них задачи исследования и ожидаемые результаты с соответствующей им новизной, практической ценностью и теоретической значимостью. Обсуждаются перспективы организации риск-анализа исследуемых контентов и использование его результатов для выработки рекомендаций по разграничению корпоративного доступа к информации.

Ключевые слова: контент, соцсеть, риск, парсинг, персонал, корпорация, доступ.

Введение

Видео [1-10], аудио [11-20] и графический [21-29] контенты, циркулирующие в социальных сетях, являются весьма информативной базой для оценки состояний той или иной аудитории их пользователей. Разумеется, это относится и к персоналу корпораций, аффилированному с подобными видами контента. Отсюда вытекает реальный практический интерес в осуществлении проектной деятельности, ориентированной на риск-анализ множества корпоративных контентов для обеспечения информационной безопасности корпорации. Традиционным инструментом обеспечения защищенности выступают здесь механизмы разграничения доступа к информации, которые в нашем контексте должны опираться на данные о деструктивной активности персонала в социальных сетях (рис. 1). Отсюда объектом проектной деятельности выступает контент, циркулирующий в ресурсах социальных сетей и аффилированный с персоналом рассматриваемой корпорации. В свою очередь предметом исследования следует считать выявление вышеуказанного контента, его селекция по признакам деструктивности и персонифицированный риск-анализ деструктивных контентов для выработки рекомендаций по разграничению доступа к корпоративной информации. С учетом вышеизложенного целью проектной деятельности является создание инструментария, программно-технически обеспечивающего совокупность перечисленных аналитических операций с указанными видами контента социальных сетей.

Программно-технические модули риск анализа персонала корпорации в части публикационной активности его представителей в социальных сетях на основе исследования видеоконтента

Объект исследования: Видеоконтент, сгенерированный с использованием технологии deepfake, циркулирующий в ресурсах социальных сетей «YouTube»,

«ВКонтакте» и аффилированный с персоналом рассматриваемой корпорации.

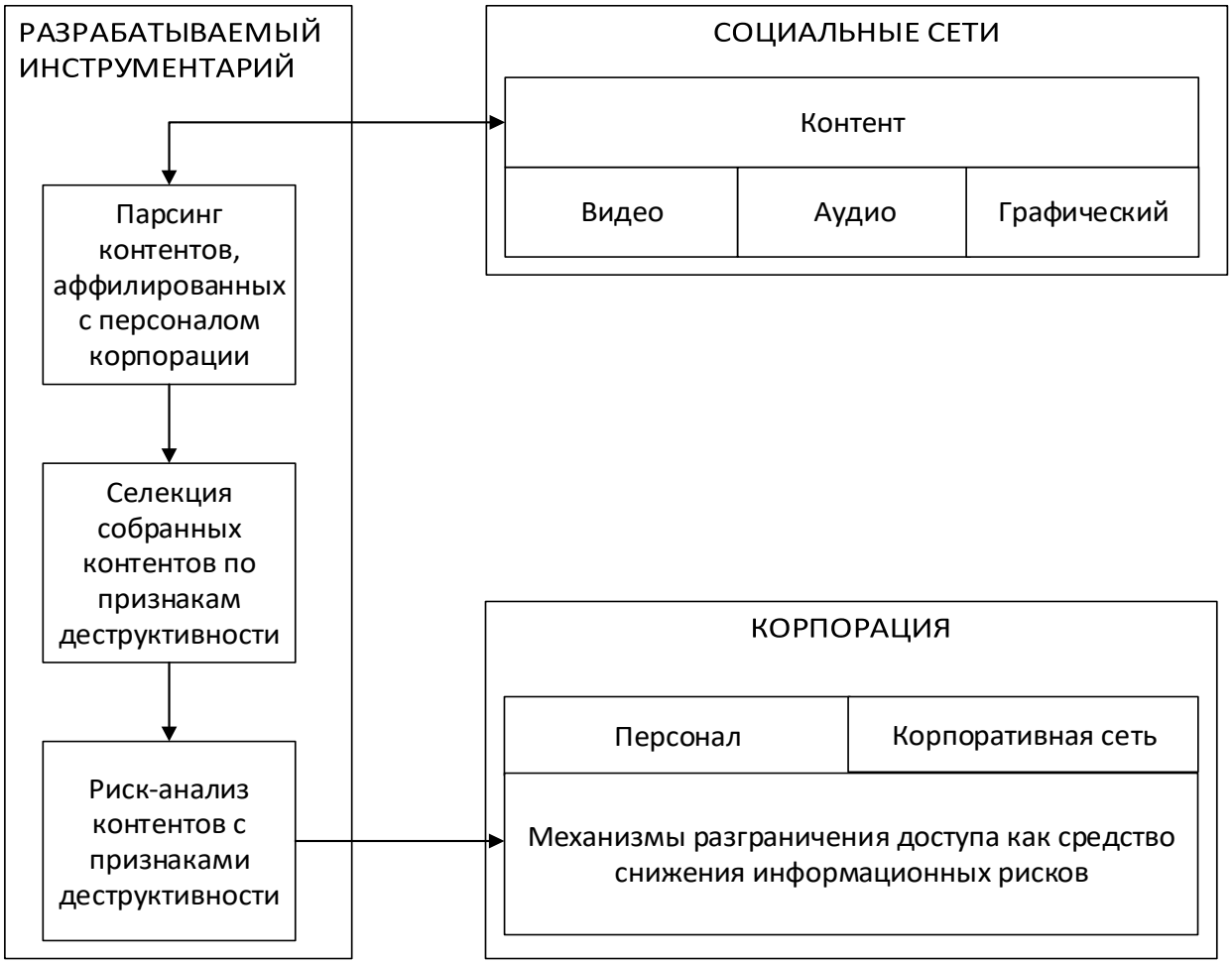

Рис. 1. Пространство проектной деятельности и ее целеполагания

Предмет исследования: выявление видеоконтента, сгенерированного с использованием технологии deepfake, и персонифицированный риск-анализ деструктивных видеоконтентов для выработки рекомендаций по разграничению доступа к корпоративной инфраструктуре.

Цель исследования: Разработка алгоритма и создание программного инструментария риск-анализа социосетевой активности персонала корпорации для регулирования риска возможных атак со стороны сотрудников-инсайдеров на основе метрик deepfake-видеоконтента, созданного, опубликованного и распространяемого ими в социальных сетях.

Актуальность

В последние годы стремительное развитие технологий создания deepfake значительно упростило и расширило возможности манипуляции аудиовизуальным контентом. Синтетические изображения и видеоматериалы стали настолько реалистичными, что их всё сложнее различить от настоящих фактов. Кроме того, известны случаи, вернее попытки, использовать синтетические аудио- и видеозаписи в качестве вещественных доказательств в суде [1].

Общее число deepfake по всему миру в течение первых месяцев 2023 года увеличилось в несколько раз по сравнению с аналогичным периодом предшествующего. Об этом говорит исследование компании DeepMedia [2].

Отмечается, что взрывной рост количества контент-фальшивок в глобальном масштабе объясняется резко снизившимися затратами на создание таких аудио- и видеоматериалов. Если раньше на точную имитацию голоса с учётом работы серверного оборудования и использования алгоритмов искусственного интеллекта (ИИ) требовалось около $10 тыс., то к началу мая 2023-го расходы снизились всего до нескольких долларов. Связано это с появлением генеративных ИИ-моделей нового поколения и более мощных аппаратных платформ, спроектированных специально с прицелом на нейросети и машинное обучение [2].

В наши дни deepfake используются как средство шантажа, для проведения информационных и политических диверсий, а также для создания ложных событий и новостей, диффамации и искусственного разрушения политической обстановки в отдельных странах или обществах [1-4]. Однако риски и угрозы бесконтрольного применения deepfake-технологий выглядят не менее внушительно в разрезе корпоративной безопасности:

1. Аутентификация и подлинность данных: deepfake-контент способен создавать фальшивые видеозаписи, которые могут быть восприняты как подлинные. Это серьезно подрывает способность корпораций аутентифицировать и подтверждать подлинность данных и информации [4].

2. Социальная инженерия и фишинг: deepfake-контент может быть использован для создания манипулятивных сценариев, в которых сотрудникам корпорации злоумышленники могут представиться легитимными лицами с целью получения доступа к конфиденциальным данным или системам.

3. Обман машинного обучения: deepfake-материалы способны вводить в заблуждение системы машинного обучения и искусственного интеллекта, используемые для обнаружения угроз и анализа данных. Это усложняет обнаружение вредоносных действий в корпоративных сетях [1,2].

4. Имперсонация и маскировка: deepfake может быть использован для имитации ключевых лиц в корпорации, включая руководителей и сотрудников, что дает возможность провести различные виды атак, включая целевые нападения внутри организации.

5. Репутационный риск: если deepfake-контент становится общедоступным и признается фальшивым, это может серьезно повредить репутацию корпорации и вызвать доверительный кризис среди клиентов и партнеров.

6. Инсайдерские угрозы: Сотрудники действующие или бывшие корпорации могут применять deepfake-технологии для создания поддельных видео с участием коллег или руководителей, что может быть использовано в целях вымогательства, шантажа или внутренней дестабилизации организации [5].

7. Законодательство и его соблюдение: Угрозы, связанные с deepfake-технологией, могут создавать дополнительные требования к корпоративной политике безопасности и потребовать обязательного обучения сотрудников в соответствии с законодательными актами и корпоративными стандартами. Однако на данный момент в российском законодательстве не предусмотрена ответственность за мошенничество с применением технологии deepfake [1].

С учетом вышеуказанных угроз очевидно, что технологии deepfake представляют серьезную угрозу не только для обычных граждан и компаний, но и для международной информационной безопасности в целом. В связи с этим становится актуальной разработкой мер регулирования deepfake с целью минимизации потенциального ущерба от их использования. Снижение влияния перечисленных факторов будет продемонстрировано при оценке эффективности предлагаемого инструментария.

Аналоги

1. Deepware – сервис с открытым исходным кодом, позволяющий анализировать видео на предмет наличия deepfake [6]. Наибольшая эффективность достигается при анализе deepfake с участием знаменитостей. К недостаткам можно отнести:

- отсутствие методики оценки риска влияния deepfake-контента на пользователей;
- ориентированность на социальные сети «YouTube», «Facebook»*, «Twitter». Для анализа иных источников требуется загружать видеозапись самостоятельно.

2. KaiCatch – корейское коммерческое приложение, дающее возможность распознавать поддельный фото и видеоконтент. Как заявляет владелец, сервис позволяет определить аномальные искажение лиц с 90% вероятностью [6]. Однако приложение реализовано исключительно на платформе Android, интерфейс поддерживает только корейский язык. Отсутствует методика оценки риска влияния deepfake-контента на пользователей.

3. Microsoft Video Autentificator [7] – разработано Microsoft Research, командой Microsoft Responsible AI и Комитетом Microsoft AI, этики и воздействия в инженерии и исследованиях (AETHER). Инструмент был создан с использованием общедоступного набора данных таких как Face Forensic++. Microsoft протестировала этот инструмент на наборе данных DeepFake Detection Challenge [8]. На данный момент сервис находится на стадии закрытого тестирования. Доступ к приложению планируется открыть к выборам президента США в 2024 году.

4. FakeBuster – помогает выявлять deepfake во время онлайн-конференций и в соцсетях [6]. Разработчики тестировали программу во время звонков в Zoom и Skype. В основу FakeBuster легла 3D-свёрточная нейросеть, которую обучили на комбинации наборов данных, таких как Deeperforensics, DFDC [7], VoxCeleb (что говорит о заточенности нейросети под анализ deepfake с участием знаменитостей), и неназванных видеороликов. В открытый доступ программа ещё не выпущена и на данный момент нет информации о возможности анализировать видеозаписи пользователей.

5. FakeCatcher – способен отличать реального человека от deepfake в том числе за счёт цветовых изменений в кровотоке – эту информацию считывают с множества точек на лице. Затем при помощи машинного обучения технология обрабатывает полученную информацию. На данный момент Intel не планирует публичный релиз инструмента.

Противоречия

1. Между необходимостью выявления видеоконтента, сгенерированного с применением технологии deepfake и опубликованного в социальных сетях «Вконтакте», «YouTube», и отсутствием открытых, отечественных инструментов мониторинга публикационной активности в сообществах и каналах, доступных для внешних исследований.

2. Между необходимостью количественной оценки рисков успешности инсайдерских атак персонала корпорации видеоконтентами с признаками деструктивности, и неспособностью аналогов к такой оценке при распространении видеозаписей, учитывающей особенности публикуемого deepfake-контента.

3. Между жесткими требованиями по контролю за дискреционным и мандатным доступом пользователей к критическим активам корпорации и отсутствием инструментов интеллектуальных подсказок для лиц, принимающих решения, в условиях информационного противоборства.

Задачи

1. Разработка алгоритма и программная реализация модуля, предназначенного для сканирования открытых аккаунтов и каналов сотрудников корпорации на социальных медиа-платформах «YouTube» и «ВКонтакте» с последующим автоматизированным выявлением deepfake видеоконтента на базе инструментария искусственных нейронных сетей (без подготовки тренировочного датасета,

содержащего материалы с «жертвой» deepfake).

2. Разработка методики количественной оценки и регулирования информационных рисков инсайдерских атак видеоконтентом, размещаемым сотрудниками корпорации в социальных сетях.

3. Реализация модуля интеллектуальных подсказок лицу, принимающему решения, по контролю за дискреционным и мандатным доступом к критичным информационным активам организации на основе уровня личного риска пользователя.

Ожидаемые результаты

1. Программная реализация, на основе аппарата искусственных нейронных сетей, алгоритма выявления deepfake-видеоконтента опубликованного сотрудниками организации в социальных сетях «Вконтакте» и «YouTube». Свидетельство о государственной регистрации соответствующего программного комплекса в реестре программ для ЭВМ.

2. Методика оценки риска инсайдерских атак в организации, на базе анализа, публикуемого сотрудниками видеоконтента.

3. Программная реализация модуля интеллектуальных подсказок лицу, принимающему решения по предоставлению доступа к критическим информационным активам предприятия. Соответствующая база подсказок и сценариев реагирования в зависимости от уровня персонального риска пользователя.

Новизна

1. Разрабатываемое программное обеспечение, в отличие от аналогов, ориентировано на выявление следов использования deepfake-технологий в видеозаписях, опубликованных в открытых сообществах и каналах социальных сетей «Вконтакте» и «YouTube», при полном отсутствии необходимости в заранее размеченных наборах тренировочных данных с участием ключевых лиц, сотрудников корпорации, государственных и медийных личностей, что в свою очередь, позволяет использовать программное обеспечение без привязки к конкретной отрасли бизнеса или государственной деятельности.

2. В отличие от аналогов, разрабатываемая методика, позволит оценить уровень персонального риска сотрудников корпорации без привязки к автоматическим инструментам поведенческого анализа, зачастую учитывающих активность сотрудника лишь в процессе трудовой деятельности в рамках выполнения рабочих задач.

3. Впервые будет собрана база сценариев превентивного реагирования на подрывную деятельность сотрудников корпорации, учитывающих угрозу влияния видеоконтента на персонал.

Практическая ценность

1. Программная реализация алгоритма позволит повысить эффективность выявления deepfake-видеоконтента, в котором используются изображения не только медийных личностей и руководителей мировых корпораций, но и любых персон при условии частичного или полного отсутствия массивов видеоданных, содержащих параметры их внешности, особенностей мимики, жестикуляции, для любой корпорации вне зависимости от отрасли ее деятельности, организационного устройства и локации.

2. Разрабатываемая методика может быть применена при комплексной оценке информационной безопасности корпорации без необходимости устанавливать стороннее программное обеспечение анализа журналов действий пользователя, что значительно снижает уровень недоверия к руководителю подразделения. Методика позволит проводить количественную оценку рисков и дополняет собой качественные оценки, полученные на базе опросов, собеседований, что поможет сформировать наиболее полную картину ландшафта угроз со стороны инсайдеров.

3. Модуль интеллектуальных подсказок, как и база сценариев реагирования, позволит частично автоматизировать работу лица, принимающего решения, по управлению

доступом к критическим информационным активам, тем самым значительно снизит влияние человеческого фактора при управлении подобными доступами.

Теоретическая значимость

1. Создаваемый алгоритм упрощает процесс селекции deepfake-видеоконтента, что стимулирует аналитическую разработку внутренних правил социальной сети, государственных нормативных актов, направленных на регулирование распространения подобных контентов.

2. Разрабатываемая методика развивает теорию риск-анализа видеоконтента в направлении учета особенностей восприятия контента, сгенерированного с помощью deepfake-технологий, и его влияния на информационно-психологическое состояние сотрудника корпорации.

3. В создаваемой методике впервые будет уделено внимание корреляции просмотров видеоконтента с трудовой деятельностью пользователя, в том числе деструктивной, что в дальнейшем позволит разработать методическое обеспечение для формирования в корпорации политики и регламентов безопасности, направленных на снижение ожидаемого ущерба от инсайдерских атак с применением методов социальной инженерии.

Архитектура и алгоритмы создаваемого инструментария

Очевидно, что подходящей средой для распространения deepfake-контента являются социальные сети для обмена медиаконтентом. В российском сегменте наиболее популярными сетями являются: «Вконтакте», «TikTok», «Telegram», «Одноклассники», «Дзен», «YouTube» [9]. При анализе объекта исследования, было выявлено, что движение deepfake-контента в некоторых сетях не представляет угрозы в контексте корпоративной информационной безопасности:

- «TikTok», «Instagram»*, «Facebook»*, «Twitter» – были заблокированы на территории Российской Федерации. Данное решение значительно снизило публикационную активность аудитории из РФ;

- «Telegram», «Дзен» – циркуляция видеоконтента затрудняется структурой связей узлов в сети. Основной тип контента – текстовый, графический;

- «Одноклассники» – популярна у отдельных слоев населения (средний возраст аудитории 44 года) [10]. Аудитория в значительно меньшей степени задействована в генерации deepfake-контента.

В связи с вышесказанным социальные сети «Вконтакте» и «YouTube» представляют наибольший интерес для данного исследования, являясь площадками для активного распространения видео-контента, созданного с применением технологии deepfake.

Разрабатываемое программное обеспечение представляет собой совокупность набора связанных программных модулей и базы данных, содержащей результаты работы комплекса (рис. 2).

В качестве исходных данных, для сканирования социальных сетей необходимы уникальные идентификаторы. В случае «Вконтакте» требуются идентификаторы личных страниц пользователя или администрируемых им сообществ. Для «YouTube» необходимы идентификаторы каналов пользователей, которыми могут служить URL-ссылки.

На первом шаге работы программного комплекса, с помощью парсеров, происходит сбор всех когда-либо опубликованных видеозаписей, размещенных на страницах(каналах) пользователей, при этом для «YouTube» учитываются не только полнометражные видео, но и короткие видеозаписи, называемые shorts (рис. 3-4).

Другой режим работы парсеров, позволяет осуществлять мониторинг исследуемых ресурсов в реальном времени, однако использование данного режима влечет за собой высокую нагрузку на вычислительные мощности. В то же время, частота публикации видеоконтента заметно ниже, чем текстового или графического контента, что связано с необходимостью в

предобработке публикуемого видео как автором, так и модерацией со стороны социальной сети. В связи с вышеперечисленным, режим сканирования в реальном времени должен применяться только в отдельных, наиболее интересных для пользователя, случаях.

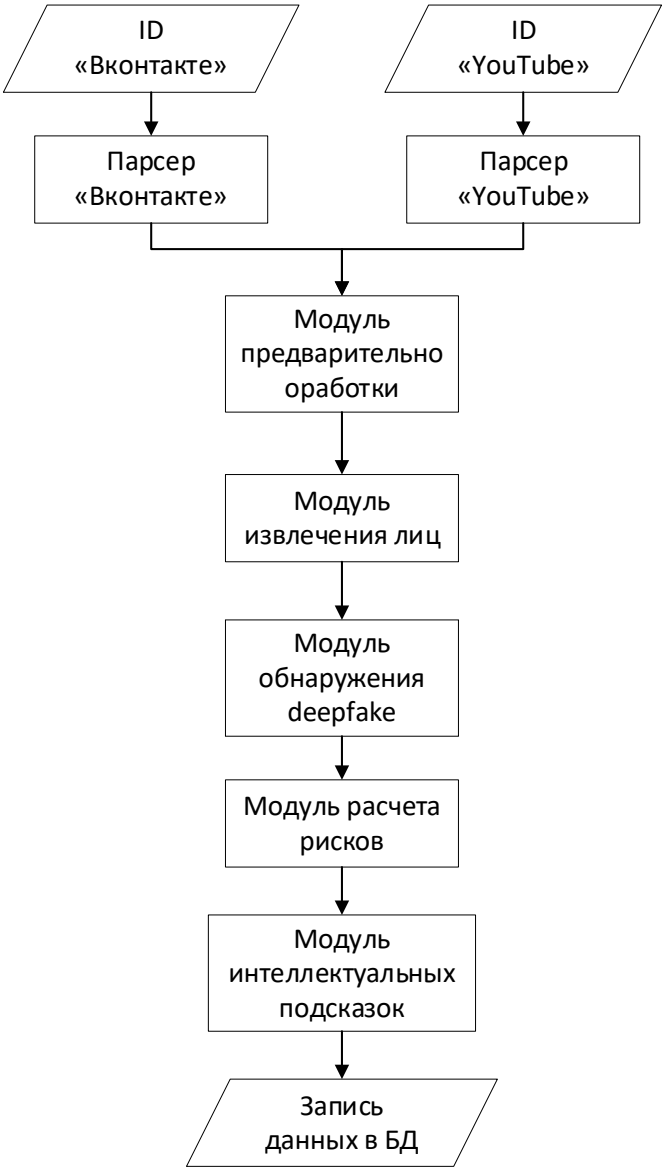

Рис. 2. Архитектура программного комплекса

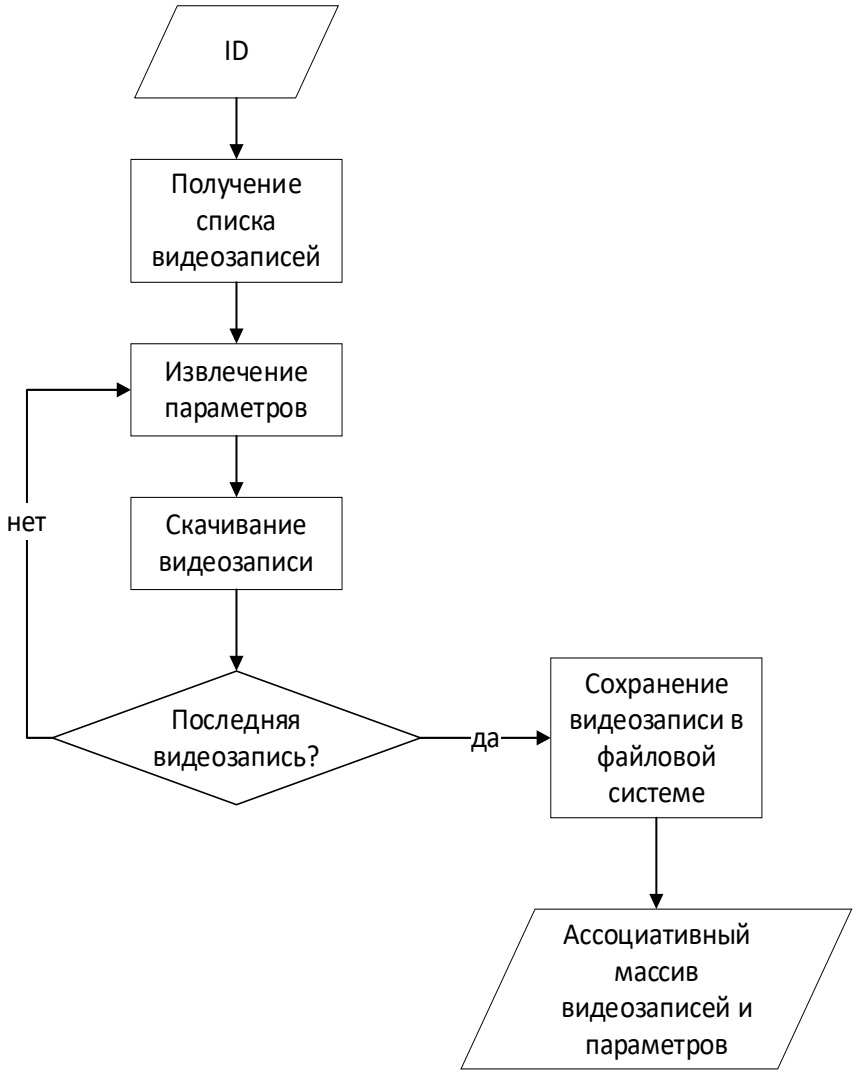

Рис. 3. Парсер сети «Вконтакте»

Для взаимодействия с сетью используются официальные интерфейсы разработчиков социальной сети, что в свою очередь накладывает ряд ограничений на объект исследования:

1. Количественное и временное ограничение для отправки запросов к VK API. Для «YouTube» данная проблема решается использованием библиотеки, позволяющей эмулировать действия пользователя, которая обращается не к API, а к драйверу браузера. К сожалению, данное ограничение приводит к замедлению сканирования из-за задержек, добавленных искусственно.

2. Невозможность мониторинга приватных личных страниц, закрытых сообществ и видеоканалов. Данное ограничение может быть решено внедрением в такие сообщества пользователей, однако подобный подход не может быть применим в контексте анализа сотрудников одной корпорации.

3. Лимиты, накладываемые поставщиком телекоммуникационных услуг. Данное ограничение препятствует скачиванию видеозаписей большого объема, что затрудняет последующий анализ. Решением проблемы является скачивание видео пакетами с задержкой. Кроме того, генеративные искусственные нейронные сети не предоставляют возможность создания

длительных deepfake-записей, поэтому данным ограничением можно пренебречь.

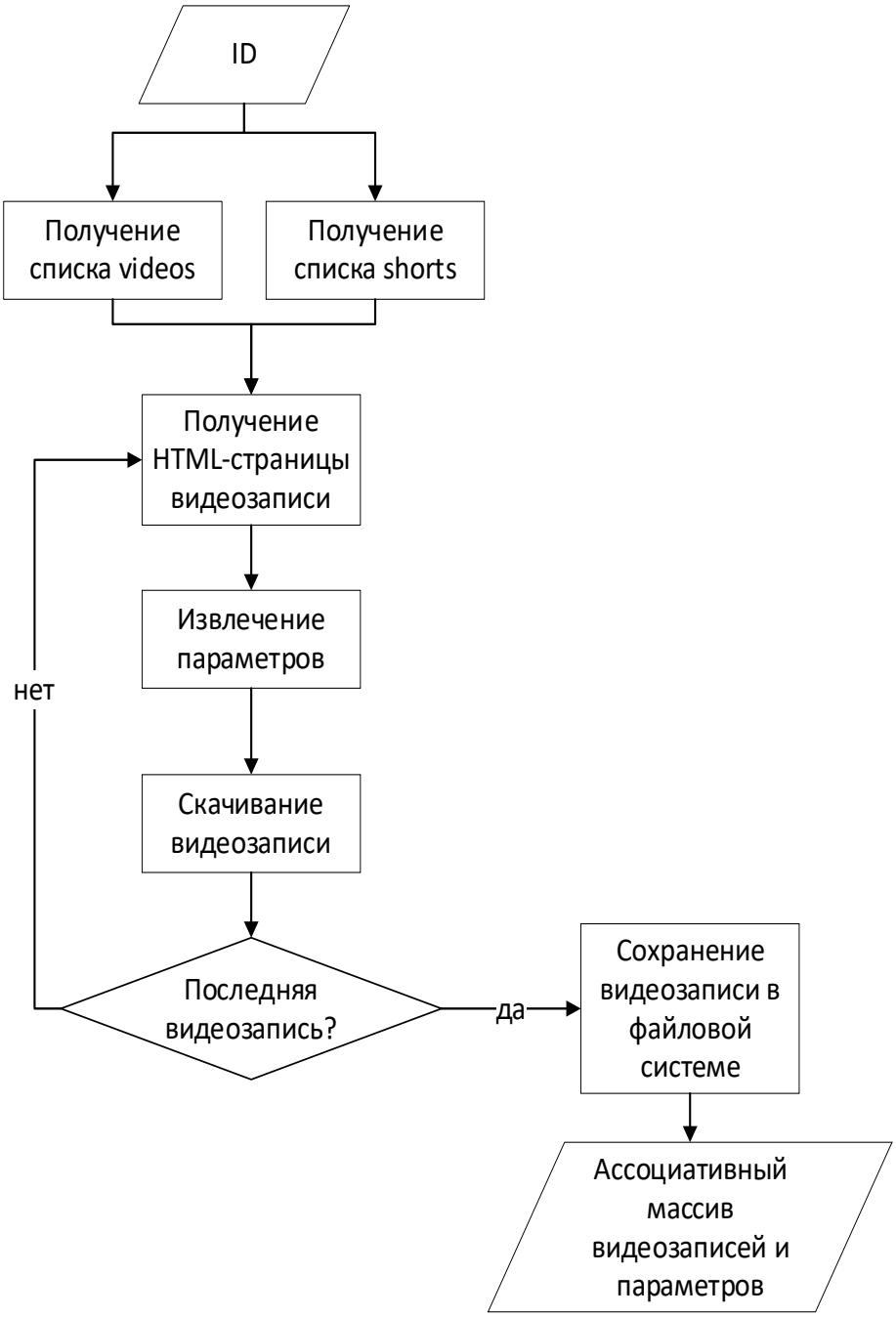

Рис. 4. Парсер сети «YouTube»

В общем случае, видеозапись представляет собой совокупность двух объектов, связанных одним временным интервалом: видеоряд, аудиодорожка. В данной работе наибольший интерес представляет видеоряд, поэтому для его извлечения разрабатывается модуль предварительной обработки (рис. 5). На вход модуля предварительной обработки из модуля парсинга передается видеофайл в формате .mp4, а на выходе пользователь получает набор отдельных кадров видео, подготовленных к последующему анализу с использованием нейронных сетей.

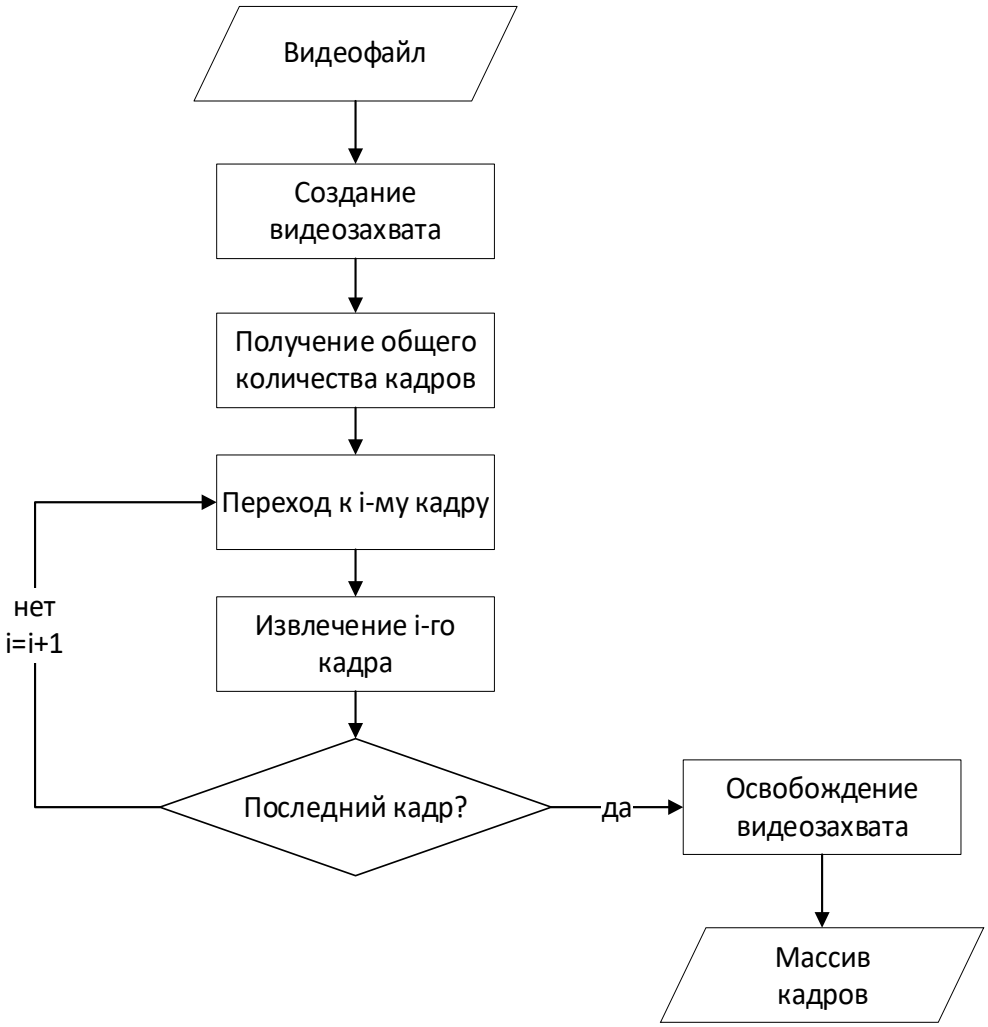

Рис. 5. Модуль предварительной обработки

Этап извлечения лиц и обнаружения фейков является наиболее важным в данном программном комплексе. Видеозапись, представленная на данном этапе в виде массива кадров, передается на вход последовательности нейронных сетей (рис. 6-7). Первая сверточная нейронная сеть – MTCNN позволяет детектировать на каждом кадре лица людей (рис. 6). Сеть FaceNet по координатам детектированного лица выделяет его характеристики и передает в сеть DBSCAN, которая проводит кластеризацию лиц (рис. 7).

Модуль обнаружения deepfake содержит в себе еще одну нейронную сеть, которая по полученным кадрам анализирует различия одних и тех же лиц, таких как тени, нарушение границ лица, появление сторонних объектов, перемещение глаз, улыбки.

На выходе модуля обнаружения deepfake пользователь получает информацию о вероятности применения технологий генеративных нейронных сетей при создании исследуемой видеозаписи. Пользователь самостоятельно задает порог вероятности, выше которого видеозапись считается deepfake, однако рекомендованным значением является вероятность равная 0,8.

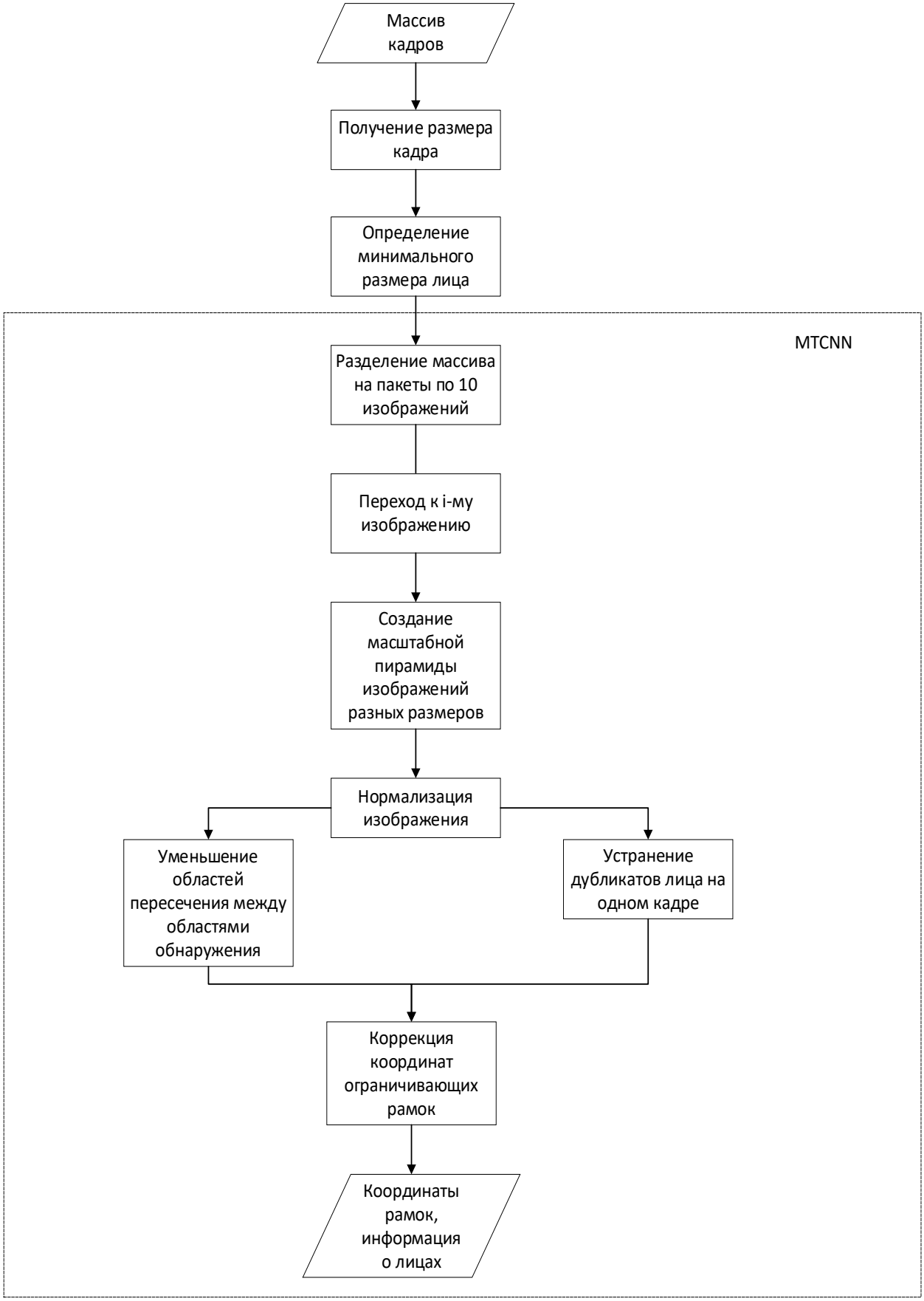

Рис. 6. Определение координат рамок лиц

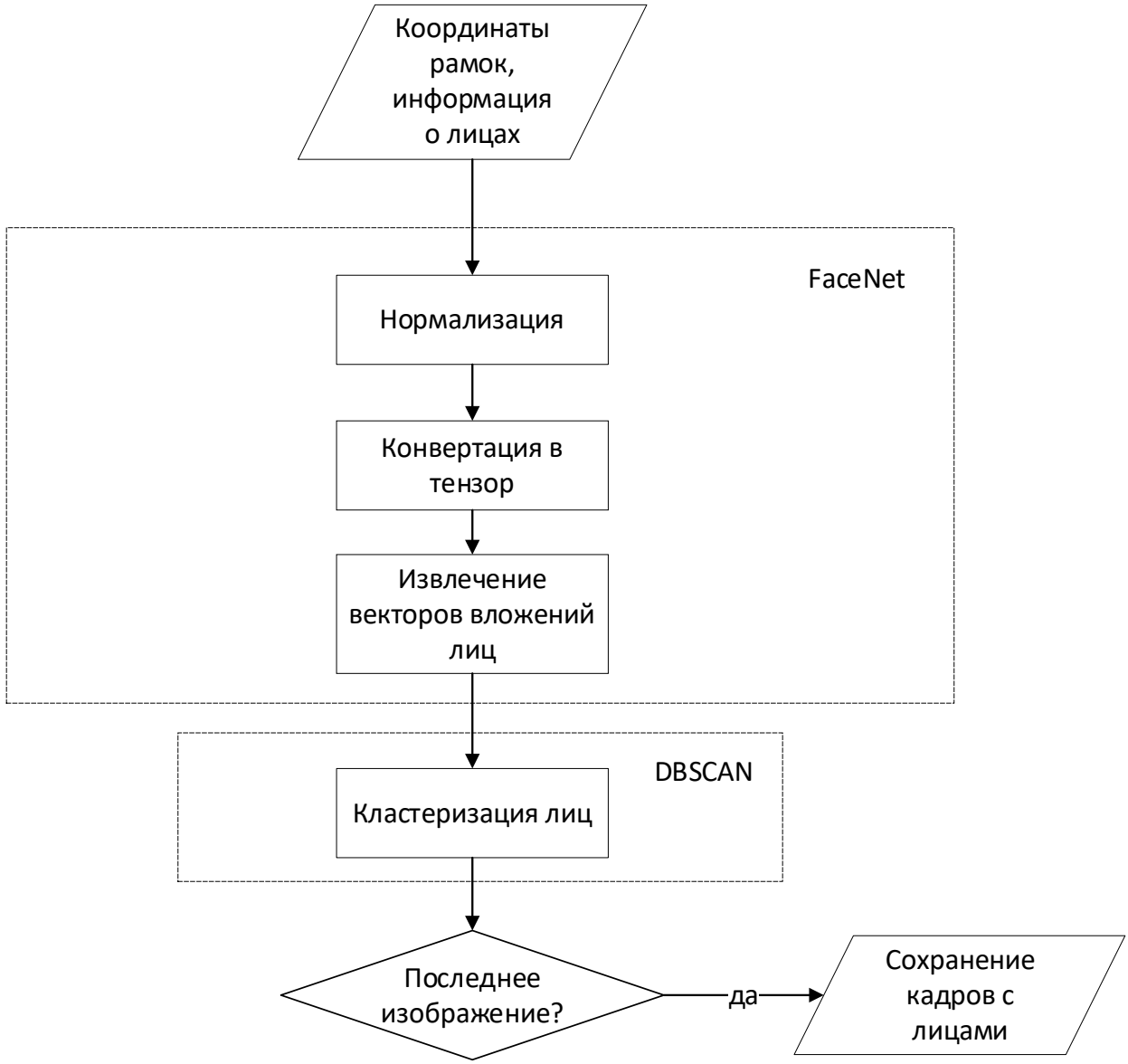

Рис. 7. Выделение характеристик и кластеризация лиц

Модуль интеллектуальных подсказок на основе рассчитанного на предыдущем шаге уровня персонального риска формирует рекомендации по реагированию с целью снижения риска инсайдерских атак. Наибольшую ценность представляют рекомендации по контролю доступа исследуемого сотрудника к критичным информационным активам.

Основной целью управления доступом является ограничение операций, которые разрешено выполнять субъекту (пользователю) над объектом. Эти операции могут включать, например, чтение и запись.

Введем некоторые обозначения:

S — множество субъектов,

O — множество объектов,

R = (r1 , r2 , . . . , rn)— множество прав доступа.

Для реализации этих прав используется матрица доступов M, где строки соответствуют субъектам, а столбцы — объектам (рис. 8). В каждой ячейке указаны права доступа, которыми обладает данный субъект по отношению к данному объекту.

Таким образом, текущее состояние системы Q может быть однозначно представлено в следующем виде:

$$Q= (S, O, M).$$

	obj_1	obj_2	...	obj_i	...	...	obj_m
sub_1							
sub_2							
...							
sub_k				$r_y, r_x, ..., r_w$			
...							
...							
sub_n							

Рис. 8. Матрица доступа

На выходе из модуля лицо, принимающее решение получает список из элементарных операций, переводящих систему из состояния

$$Q = (S, O, M) \text{ в } Q' = (S', O', M'),$$

вида:

1. Добавление права r в ячейку M [s; o]:

$$op = \text{enter r into M [s, o]},$$

правило соответствует ситуации низкого уровня персонального риска, при котором права сотрудника на отдельный объект могут быть расширены.

2. Удаление права r из ячейки M [s; o]:

$$op = \text{delete r from M [s, o]},$$

правило соответствует высокому значению персонального риска, права на объект должны быть урезаны.

3. Удаление субъекта s:

$$op = \text{destroy subject s},$$

уровень персонального риска критичен, рекомендуется поставить на регулярный контроль деятельность сотрудника, в крайних случаях лишить сотрудника возможности доступа к критическим активам и провести внутреннее расследование.

Однако кроме операций по смене доступа, модуль интеллектуальных подсказок предлагает лицу, принимающему решения, набор возможных сценариев с целью снижения уровня риска. Сценарии позволят своевременно провести мероприятия с сотрудником, тем самым предотвратить как потенциальные инсайдерские атаки с его стороны, так и повысить лояльность персонала к корпорации.

Программно-технические модули риск-анализа персонала корпорации в части публикационной активности его представителей в социальных сетях на основе исследования аудиоконтента

Объект исследования: Аудиоконтент, публикуемый, прослушиваемый и распространяемый персоналом корпорации на личных страницах сотрудников и сообществ, в социальной сети «ВКонтакте».

Предмет исследования: Выявление и измерение параметров аудиоконтента, позволяющих оценить наличие деструктивности в его содержимом и степень влияния на персонал корпорации.

Цель исследования: Создание инструментария для автоматизированной идентификации и риск-анализа деструктивного аудиоконтета в социальной сети «ВКонтакте» для выработки

рекомендаций по регулированию рисков его воздействия на персонал корпорации.

Актуальность

В современном мире из-за стремительного развития компьютерных технологий и всеобщей цифровизации коммуникация социальные сети стали неотъемлемой частью жизни людей [11]. В последние годы пользователи социальных сетей стали все чаще использовать голосовые сообщения вместо текстовых. Несмотря на то, что они были доступны на многих платформах около 10 лет, только в последние годы голосовые сообщения начали входить в повседневную жизнь пользователей, в настоящее время более 60% россиян используют их [12-13].

Совместно с сообщениями во многих социальных сетях присутствует раздел с музыкой, в которые ежедневно загружаются более 60 000 аудиозаписей [14]. Общее количество пользователей, слушающих музыку каждый день, увеличивается ежегодно, по результатам опросов более 60% респондентов ежедневно слушают её [15].

Однако увеличение поступающего пользователям аудиоконтента влечет за собой и его всестороннюю направленность. Например, музыка и политика – две разные сферы, но на самом деле они имеют точки соприкосновения. Однако, при детальном рассмотрении можно выяснить, что с давних времен на формирование общественных суждений и политических взглядов имела огромное влияние именно музыка. Патриотические песни, гимны, воздействуя на массы, вызывают различные эмоции у слушателей [16]. В настоящее время политическая угроза пропаганды активно используется по всему миру, в том числе на концертах и фестивалях.

Различную пропаганду и оскорбительный контент многие платформы не цензурируют, а проверяют только лишь на авторские права. В связи с этим, влияние на слушателей и общество возрастает.

Очевидно, что одним из видов платформ, для распространения аудиоконтента, являются социальные сети. В СНГ регионе наиболее популярными являются: «ВКонтакте», «Telegram», «Одноклассники», «TikTok», «Instagram»* [18]. При анализе объекта исследования, было выявлено, что аудиоконтент не представляет угрозы в некоторых социальных сетях:

- «TikTok» и «Instagram»* – с недавнего времени были заблокированы на территории Российской Федерации, доступ к данным ресурсам стал затруднительным для большей части аудитории, что снизило популярность данных сетей;

- «YouTube» – видеохостинг, предоставляющий пользователям доступ к огромному числу видеоконтента;

- «Telegram» – в данной социальной сети преимущественно преобладает текстовый формат сообщений с минимальным количеством аудиофайлов;

- «Одноклассники» – платформа с основной аудиторией людей 40-64 лет, основная направленность текстовый и графический контент [19];

- «ВКонтакте» – имеет поддержку текстового, графического, видео и аудио контента. Является самой популярной социальной сетью в СНГ.

Из вышесказанного следует, что социальная сеть «ВКонтакте» представляет наибольший интерес для данного исследования, являясь стриминговой площадкой и мессенджером, где наиболее активно распространяется и публикуется аудиоконтент.

При рассмотрении социальной сети «ВКонтакте» можно выделить несколько сущностей для мониторинга:

- личные страницы пользователей – аудиоконтент может быть добавлен в «понравившиеся», опубликован на странице;

- сообщества – в основном тематические сообщества, где аудио публикуются в виде постов с подборками;

- раздел музыки – отсутствует связь с пользователями и сообществами, возможно отследить текущий тренд;

- личные сообщения – основная среда голосовых сообщений, доступ невозможно получить третьим лицам.

При этом мониторингу и анализу подвергается только общедоступная информация. В случае, если пользователь или сообщество ограничило доступ к странице, получить эти данные невозможно. Таким образом основной исследуемый контент находится: в открытых личных страницах пользователей; в сообществах.

Анализ аудиоконтента в социальных сетях, распространяемого и прослушиваемого сотрудниками корпорации остается актуальным и важным вопросом в контексте обеспечения корпоративной информационной безопасности. Сотрудники многих компаний используют социальные сети, в которых публикуется аудиоконтент, что обуславливает следующие угрозы:

1. Распространение низкокачественного аудиоконтента: Социальные сети дали возможность многим людям стать музыкантами и опубликовывать свои работы без профессиональной проверки. Это приводит к большому количеству низкокачественного аудиоконтента с неприемлемыми текстами. Анализ которого позволяет идентифицировать и предупредить распространение подобных произведений.

2. Негативное влияние на слушателей: Аудиоконтент может иметь влияние на эмоциональное состояние и поведение слушателей. Например, некоторые тексты могут поощрять насилие, употребление наркотиков или другие негативные действия [20].

3. Психологическое воздействие: Различный аудиоконтент имеет сильное психологическое воздействие на людей, он может вызывать депрессию, агрессию, чувство одиночества и другие негативные эмоции, особенно у молодежи, которая является основной аудиторией социальных сетей [20].

4. Репутационный риск: Корпорации стремятся поддерживать положительную репутацию, как среди сотрудников, так и среди партнеров и клиентов. Позволение сотрудникам прослушивать экстремистский аудиоконтент на рабочем месте или использовать его в различных мероприятиях может нанести ущерб репутации компании.

5. Нарушение законодательства: В большинстве стран существует законодательство, направленное на борьбу с публичными призывами к насилию, терроризму и ненависти. Аудиофайлы, содержащие такие призывы, могут быть признаны незаконными и влечь за собой административную или уголовную ответственность.

Все вышеперечисленное говорит о том, что анализ аудиоконтента является серьезной угрозой для корпоративной информационной безопасности. В связи с этим, возникает необходимость регулирования рисков перечисленных деструктивных факторов. Однако существующий инструментарий не позволяет в полной мере провести обработку и анализ контента. Это будет продемонстрировано при оценке эффективности предлагаемого инструментария.

Аналоги

1. «LF-сервис» – сервис обнаружения ненормативной лексики и терроризма, обрабатывает только текстовый формат. К его недостаткам можно отнести:

- возможность использования только с применением плагина или телеграмм-бота;

- сервис полностью коммерческий, с побуквенной оплатой и ограничением в 20000 символов.

2. «Окулус» – система, разработанная РКН России, создавалась только для закрытого использования. Работает с 2022 года и анализирует фото и видео файлы. Из недостатков – открытый исходный код и API отсутствуют [17].

3. «Detecting Propaganda Online» – зарубежный сервис обнаружения пропаганды в режиме реального времени. Включает в себя различные классификационные модели, которые используют нейросети для анализа фото и видео контента. Поддерживает только английский язык и не работает с аудиофайлами и текстом.

Противоречия

1. Между необходимостью выявления деструктивного аудиоконтента, публикуемого и прослушиваемого сотрудниками корпорации в социальных сетях, и отсутствием в аналогах инструментов мониторинга подобной активности.

2. Между отсутствием у аналогов методического обеспечения для идентификации деструктивного аудиоконтента и необходимостью создания метрик аудиоконтента, позволяющих оценить наличие деструктивности в содержимом и ее степень.

3. Между отсутствием у аналогов функций обработки аудиоконтента и необходимостью создания методики и программного обеспечения для автоматизированного риск-анализа аудиофайлов, с последующей выработкой рекомендаций по регулированию информационных корпоративных рисков.

Задачи

1. Создание программного модуля, предназначенного для сканирования аккаунтов сотрудников корпорации и сообществ в социальной сети «ВКонтакте» с целью автоматического выявления аудиоматериалов, аффилированных с персоналом корпорации.

2. Разработка алгоритма, использующего нейросети транскрибации и эмоционального анализа, и его программная реализация, позволяющая на основе аудио-метрик, обнаружить деструктивность в содержимом аудиофайлов.

3. Разработка методик оценки и регулирования рисков воздействия деструктивного аудиоконтента на персонал корпорации.

Ожидаемые результаты

1. Программная реализация модуля автоматизированного мониторинга аудиоконтента, публикуемого, прослушиваемого и распространяемого персоналом корпорации на страницах сотрудников и сообществ, в социальной сети «ВКонтакте».

2. Алгоритм автоматизированного анализа аудиоконтента, выявляющий деструктивность в его содержании и ее степень, на основе искусственных нейронных сетей и параметров аудиофайла. Свидетельство о государственной регистрации соответствующего программного комплекса в реестре программ для ЭВМ.

3. Методическое обеспечение и его программная реализация в виде модуля риск-анализа аудиоконтента и формирования интеллектуальных подсказок лицу, принимающему решения по предоставлению доступа к информационным активам корпорации, включая сценарии реагирования для разных уровней риска.

Новизна

1. В отличие от аналогов, разрабатываемое программное обеспечение ориентированно на аудиоконтент, аффилированный с персоналом корпорации.

2. Разрабатываемая методика выявления аудиоконтента с признаками деструктивности предоставляет новые метрики для анализа аудиоконтента, в отличие от аналогов, учитывающие технические параметры аудиозаписи и эмоциональное состояние автора.

3. Разрабатываемая методика риск-анализа, в отличие от аналогов, позволяет также учитывать параметры аудиоконтента при регулировании рисков его воздействия на персонал корпорации.

Практическая ценность

1. Программная реализация алгоритма автоматического мониторинга аудиоконтента сотрудников корпорации в социальной сети «ВКонтакте» может быть использована наравне с инструментами поведенческого анализа.

2. Разрабатываемые аудио-метрики расширят возможности анализа аудиоконтента, что позволит различным стриминговым платформам и другим

компаниям, использующим аудио, проводить модерацию контента в автоматизированном режиме.

3. Разрабатываемая методика позволит корпорации проводить комплексную оценку рисков нарушения её информационной безопасности.

Теоретическая значимость

1. Разрабатываемые алгоритмы развивают способы мониторинга контентов в социальной сети «ВКонтакте», так как в настоящее время нет аналогов, учитывающих деструктивный аудиоконтент.

2. Создаваемое методическое обеспечение развивает теорию риск-анализа контентов, через введение новых аудио-метрик, позволяющих адекватно оценить его деструктивность в содержании и ее степень;

3. Разрабатываемая методика риск-анализа и выработки рекомендаций лицу, принимающему решения по предоставлению доступа к информационным активам корпорации, позволит не только идентифицировать потенциально опасных сотрудников на ранних стадиях, минимизируя возможный ущерб, но и предложить различные сценарии по его снижению.

Архитектура и алгоритмы создаваемого инструментария

Программный комплекс состоит из набора модулей (рис. 9).

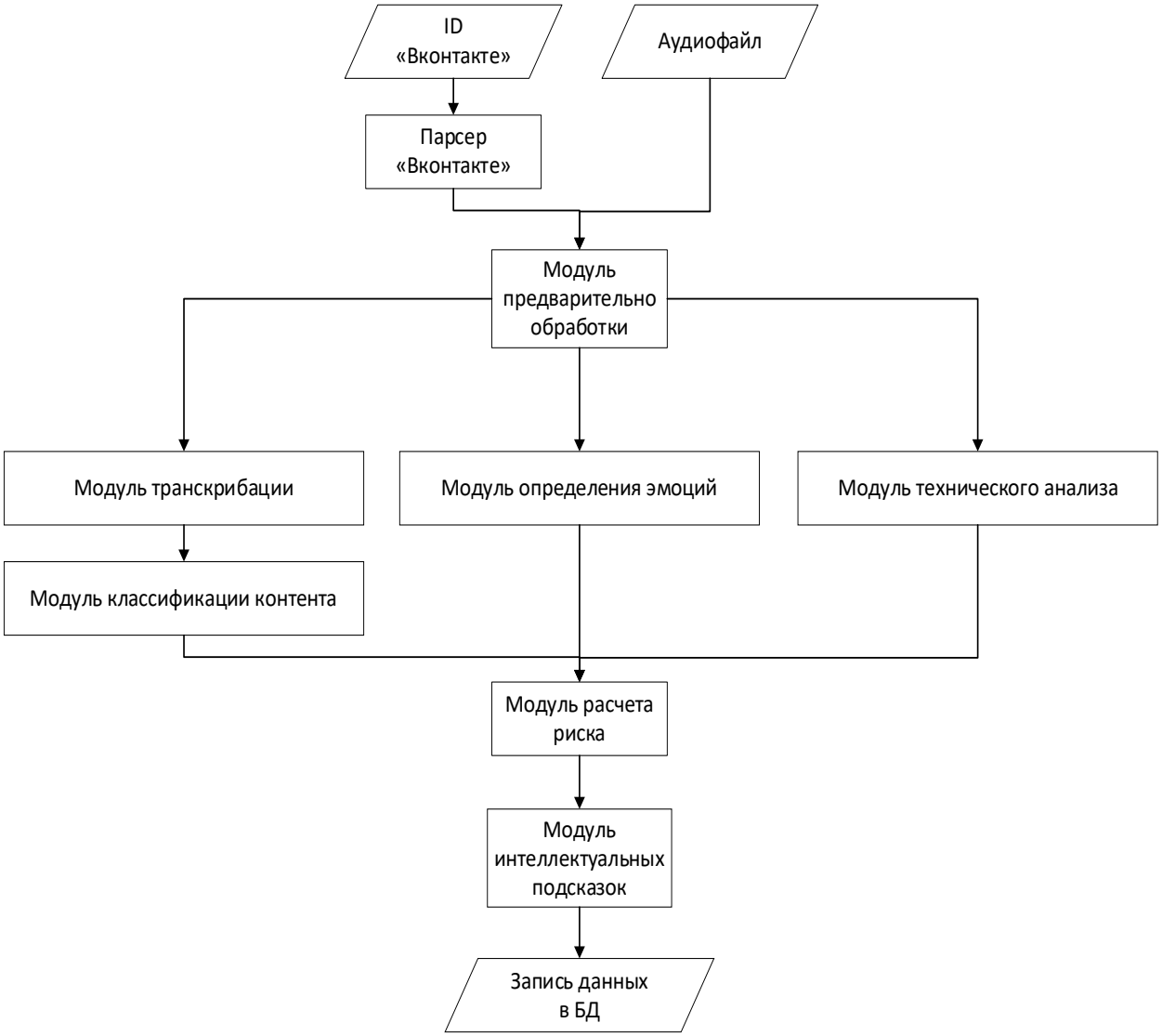

Рис. 9. Архитектура программного комплекса

Парсер сети «Вконтакте»

В настоящее время для автоматизации получения информации из социальных сетей и других задач, связанных с поиском информации на веб-странице, применяют парсинг (рис. 10). В данной работе реализовано два подхода к его разработке:

- получение информации о пользователе или сообществе, используя VK API;

- получение аудиозаписи используя данные из HTML страницы.

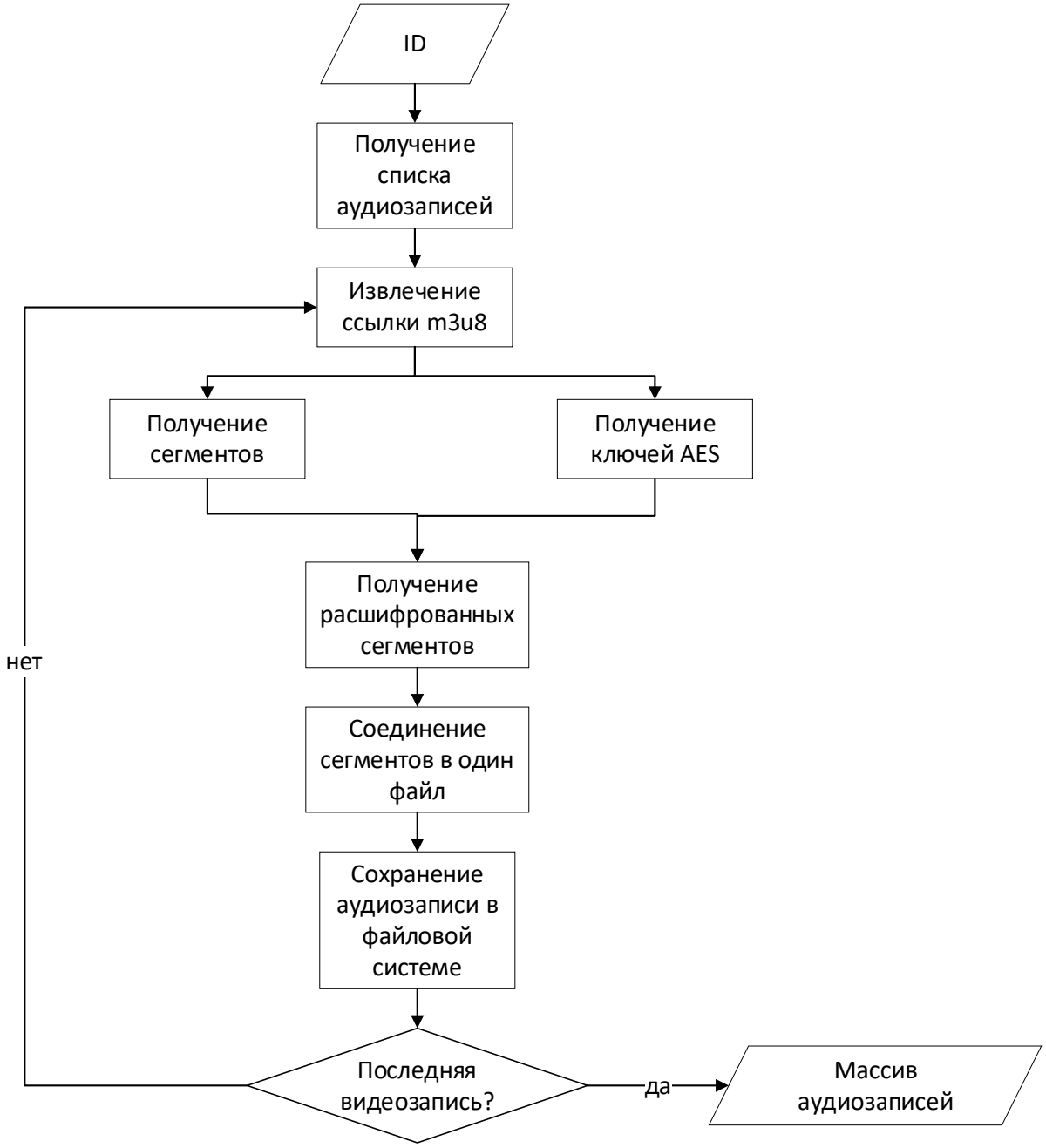

Рис. 10. Парсер сети «Вконтакте»

Для работы с различными веб-ресурсами существует несколько способов взаимодействия, однако самый простой и популярный метод — это интерфейс API,

представляющий собой набор готовых функций, структур. При этом использование API накладывает ограничения, так как это готовые методы и в них отсутствует гибкость.

Достоинства использования API: простота интеграции в проекты; доступная документация.

Недостатки использования API: низкая гибкость – не всегда возможно получить все необходимые данные; необходимы различные права для использования методов.

VK API предоставляет методы для работы с пользователями и сообществами, однако с 16 декабря 2016 года был отключен публичный интерфейс для аудиозаписей. Поэтому существует другой способ взаимодействия – сканирование HTML страницы для получения необходимых данных. Однако этот подход имеет свои особенности.

Достоинства сканирования HTML страницы: возможность работы с любым сайтом; получение только необходимых данных из полей.

Недостатки сканирования HTML страницы: сложность реализации; изменение структуры сайта, требует изменения программы сканирования.

Таким образом, используя VK API, мы получаем данные о пользователях и сообществах. Затем при помощи сканирования HTML страницы, мы получаем ссылку на m3u8 файл, который является сегментированным аудиофайлом. Часть сегментов зашифровано методом AES-128.

Также, в файле m3u8 находится ссылки на ключи шифрования, при помощи которых зашифрованные сегменты расшифровываются и соединяются в один аудиофайл.

Модуль предварительной обработки

Данный модуль проверяет аудиофайлы, состоящие из сегментов. Затем создает массив аудиофайлов, конвертирует формат файлов, так как для работы с нейронными сетями транскрибации и определения эмоций необходимы различные форматы. После чего данные аудиофайлы становятся доступны для последующего анализа.

Модуль транскрибации текста

Нейросеть преобразования аудиофайлов в текстовый формат, использующая модель Conformer-Transducer. В качестве исходных данных поступает аудиофайл формата wav. После обработки в качестве выходных данных поступает текстовый документ, с транскрибированной речью в виде строки.

Модуль определения эмоций

Нейросеть для анализа эмоциональных оттенков речи человека, использует речевую модель WavLM. Модель поддерживает 7 различных эмоций: злость, отвращение, страх, счастье, интерес, грусть, нейтрально. При снижении количества эмоций значительно увеличивается вероятность определения эмоции.

На вход поступает аудиофайл формата ogg или mp3. На этапе обработки, в качестве выходных данных, предоставляется вектор эмоций и вероятность определения каждой из них.

Модуль интеллектуальных подсказок

Данный модуль позволяет информировать лицо, принимающее решения по предоставлению доступа к критическим информационным активам предприятия, различные сценарии реагирования на превышения установленных пределов уровней риска.

Программно-технические модули риск-анализа персонала организации в части публикационной активности его представителей в социальных сетях на основе исследования графического контента

Объект исследования: графический контент, созданный при помощи искусственного интеллекта (далее «синтетический графический контент»), подверженный чтению и распространению персоналом корпорации на открытых для внешнего мониторинга личных страницах социальных сетей «Вконтакте» и «Instagram»*.

Предмет исследования: выявление и измерение параметров синтетического графического деструктивного контента, аффилированного с персоналом корпорации.

Цель исследования: на основе метрик графического контента, генерируемого сотрудниками корпорации в социальных сетях, разработка методического и программного обеспечения,

автоматизированного выявления и риск-анализа синтетического графического контента для регулирования рисков атак со стороны инсайдеров.

Актуальность

Область генерирования синтетического контента при помощи ИИ стремительно развивается на сегодняшний день. Способность обнаруживать изображения, созданные ИИ, теперь становится важной для обеспечения информационной безопасности, как для организаций, так и для отдельных лиц. Ранее ИИ часто порождал изображения с явными визуальными дефектами, которые можно было легко выявить человеческим взором. Однако в настоящее время возможности моделей ИИ в создании высокоточных и фотореалистичных изображений значительно возросли. Сгенерированные ИИ изображения теперь находятся на таком высоком уровне качества, что их подлинность практически невозможно определить визуально [21, 22].

Так в декабре 2019 года «Facebook»* удалил 682 аккаунта, которые якобы использовали фотографии профилей, созданные искусственным интеллектом и выдаваемые за американцев, в целях создания информационной поддержки Дональду Трампу.

Очевидно, что социальные сети, охватывающие огромные аудитории пользователей, являются основной средой циркуляции разнообразного контента и главными площадками для распространения синтетически сгенерированного контента являются именно социальные сети. В российском сегменте интернета особенно популярны следующие платформы: «ВКонтакте», «Instagram»*, «TikTok», «Telegram», «Одноклассники», «YouTube» [23, 24]. В связи с особенностями работы «TikTok» и «YouTube» с видеоформатами [23, 24], более возрастной аудиторией «Одноклассников» [23], а также ориентацией «Telegram» на текстовые сообщения, они могут быть исключены из данного исследования [23]. Вместо этого, больше внимания стоит уделить социальным сетям, «ВКонтакте» и «Instagram»*, которые предоставляют разнообразные форматы контента и охватывают широкий диапазон пользовательской аудитории [23, 24].

В связи с вышеперечисленными характеристиками социальных сетей, социальные сети «ВКонтакте» и «Instagram*» представляют наибольший интерес для данного исследования, являясь популярными площадками для активного распространения графического контента.

На сегодняшний день синтетически сгенерированный графический контент может легко использоваться для нарушения конфиденциальности, мошенничества, манипуляций общественным мнением, шантажа, распространения дезинформации и ведения информационной войны. В связи с этим следует рассмотреть риски и угрозы распространения подобного контента в контексте обеспечения корпоративной информационной безопасности:

1. Нарушения конфиденциальности: ИИ может быть использован для создания фальшивых фотографий людей, включая тех, кто не согласился с использованием своего изображения. Это может нарушать частную жизнь и конфиденциальность личных данных [25, 26, 27].

2. Нарушение законодательства: ИИ может быть использован для создания изображений, пропагандирующих экстремистские, террористические и националистические идеи, а также для распространения и употребления запрещенных веществ. Это может привести к нарушению законодательства [25, 26, 27].

3. Репутационные и имиджевые риски: Оценка графического контента позволяет определить, каким образом персонал организации и публикуемый им контент ассоциируются в онлайн-среде с брендом компании и как это может влиять на ее репутацию. Например, негативные изображения могут повредить репутации [26, 27].

4. Социальное манипулирование: искусственно сгенерированный графический контент может использоваться для манипуляций общественным мнением, шантажа коллег или для получения

несанкционированного доступа путем маскировки под легитимное лицо [25, 27].

5. Риски атак инсайдеров: Сотрудники или бывшие сотрудники организации могут использовать искусственно сгенерированные изображения в целях вымогательства, шантажа или внутренней дестабилизации организации [25, 26, 27].

Аналоги

1. Hive Moderation - нейросеть-модератор. Она использует возможности машинного обучения для того, чтобы идентифицировать определенный контент: находит в чатах нецензурную ругань, обнаруживает эротику или порнографию, в том числе визуальную, а потом помечает нарушителей, чтобы впоследствии с ними разобрались модераторы-люди. И она тоже может обнаруживать сгенерированные нейросетью изображения. Имеет доступный демонстрационный инструмент для анализа изображений на следы ИИ. [28]

2. Облачные сервисы: Azure AI Content Moderator, Google Cloud Vision, Clarifai, Amazon Rekognition. Предоставляют свой функционал как API для разработчиков [29].

Данные сервисы очень похожи и предназначены для автоматической фильтрации текста, изображений и видео. Включают в себя службы модерации контента на основе ИИ, которые позволяют тщательно классифицировать контент, проверять изображения на предмет порнографического содержимого, насилия, распознавать текст и лица, а также автоматически определять автора контента (человек/ИИ) [29].

3. Бесплатные онлайн сервисы для детектирования ИИ-изображений: AI or Not, Illuminarty.ai, Maybe's AI Art Detector, contentatscale.ai. Распознают изображение сгенерированные таким сетями как MidJourney, DALL-E, Stable Diffusion [28].

Недостатки аналогов

1. Отсутствие методик оценки риска воздействия фейкового графического контента на аудиторию социальной сети.

2. Отсутствие российской локализации, во всех аналогах.

3. Малый охват возможностей распознавания изображений, сгенерированных нейронными инструментами, у бесплатных аналогов (Хорошо распознают только изображения сгенерированные популярными нейросетями).

4. Hive Moderation и другие сервисы модерации – предоставляют свой расширенный функционал только на платной основе.

5. Сервисы модерации такие, как: Azure AI Content Moderator, Google Cloud Vision, Clarifai, Amazon Rekognition, предоставляют только API для разработчиков и не являются отдельными, доступными для использования приложениями.

Противоречия

1. Между необходимостью обнаружения деструктивного графического контента, аффилированного с персоналом корпорации, и отсутствием в аналогах инструментов мониторинга личных страниц сотрудников в социальных сетях «Вконтакте» и «Instagram»*.

2. Между необходимостью селекции синтетического графического контента, сгенерированного нейронной сетью с произвольной архитектурой, и направленностью аналогов на конкретные, наиболее популярные, модели ИИ.

3. Между отсутствием в аналогах методики риск-анализа, учитывающей особенности синтетического графического контента, и необходимостью в оценке и регулировании информационных корпоративных рисков.

Задачи

1. Создание программных модулей автоматизированного сбора графического контента из открытых личных страниц сотрудников корпорации в социальных сетях «Instagram»* и «ВКонтакте».

2. Разработка алгоритма и его программная реализация, использующая аппарат искусственных нейронных сетей, позволяющая выделить фейковые изображения, сгенерированные произвольной моделью ИИ.

3. Разработка методики оценки и регулирования информационных корпоративных рисков воздействия деструктивного синтетического графического контента на персонал корпорации.

Ожидаемые результаты

1. Программные модули, реализующие алгоритмы автоматизированного сбора графического контента, среди информации, публикуемой сотрудниками корпорации на личных страницах в социальных сетях «Instagram»* и «ВКонтакте».

2. Алгоритм и его программная реализация, основанная на аппарате искусственных нейронных сетей, предназначенная для анализа графического контента на наличие следов ИИ-генерации.

3. Методика оценки и регулирования информационных корпоративных рисков, учитывающая особенности воздействия деструктивного графического контента на персонал корпорации.

Новизна

1. В отличие от аналогов, разрабатываемое программное обеспечение ориентировано на обнаружение в социальных сетях «Instagram»* и «ВКонтакте» графического контента, аффилированного с персоналом корпорации.

2. Алгоритм анализа графического контента, в отличие от аналогов, ориентирован на всевозможные архитектуры нейронных сетей, применяемых для генерации фейковых изображений, что позволит применить его при появлении новых моделей ИИ.

3. Разрабатываемая методика риск-анализа впервые позволит учесть параметры синтетического графического контента при регулировании рисков его влияния на персонал организации.

Практическая ценность

1. Программная реализация алгоритмов мониторинга графического контента, генерируемого сотрудниками корпорации, в социальных сетях «Instagram»* и «ВКонтакте», может быть использована совместно с инструментарием предотвращения утечек информации, в том числе фильтрами нежелательного контента.

2. Разрабатываемый алгоритм выявления синтетического графического контента, использующий аппарат искусственных нейронный сетей, позволит автоматизировать процесс обнаружения поддельных изображений в компаниях, чья деятельность связана с подтверждением подлинности цифровых объектов искусства.

3. Разрабатываемая методика позволит учесть влияние синтетического графического контента на персонал корпорации при проведении комплексной оценки рисков нарушения ее информационной безопасности.

Теоретическая значимость

1. Создаваемое программное обеспечение автоматизированного мониторинга «Instagram»* и «ВКонтакте» стимулирует разработку внутренних политик безопасности и регламентов социальных сетей, регулирующих распространение синтетического графического контента.

2. Алгоритм выявления следов ИИ-генерации в графическом контенте позволит в дальнейшем сформировать набор универсальных признаков для обнаружения подобных контентов вне зависимости от характеристик инструментов их создания.

3. Разрабатываемая методика риск-анализа, учитывающая влияние синтетического графического контента на персонал корпорации, позволит принимать превентивные меры по отношению к конкретным сотрудникам, для снижения их потенциальной опасности и минимизации

ущерба, в случае осуществления ими инсайдерских атак.

Архитектура и алгоритмы создаваемого инструментария

Разрабатываемое в рамках данного исследования программное обеспечение для автоматизации сбора и анализа синтетического графического контента из профилей сотрудников организации в популярных социальных сетях «ВКонтакте» и «Instagram»* имеет следующую архитектуру (рис. 11):

1. Парсеры «ВКонтакте» и «Instagram»*: Модули, отвечающие за автоматизированный сбор изображений из профилей сотрудников в социальных сетях. Сбор осуществляется на основе предоставленных параметров, таких как идентификаторы пользователей «ВКонтакте» (рис. 12) и уникальные имена пользователей «Instagram»* (рис. 13). Аутентификация в API «ВКонтакте» осуществляется посредством использования access токена (рис. 12), а в «Instagram»* через вход в специально созданные технические аккаунты (рис. 13). Существует несколько проблем в работе с данными «Instagram»*: это блокировка в России, что требует использования прокси-серверов, а также активные меры «Instagram»* по борьбе с автоматизированным сбором данных. При этом получение официального токена для доступа к его API является сложным процессом, подразумевающим строгое соблюдение правил и процедур регистрации приложения, установленных Meta*.

2. Модуль обработки данных: Этот модуль выполняет проверку наличия информации, поступающей от парсеров объединяет ее в общий массив, и осуществляет загрузку изображений в файловую систему компьютера. При обнаружении новых данных от парсеров, модуль принимает решение о сохранении изображений, обеспечивая их доступность для последующего анализа и использования в рамках разрабатываемого программного обеспечения (рис. 14).

3. Модуль извлечения метаданных: Этот модуль отвечает за извлечение метаданных изображений с целью получения более детальной информации о конкретных изображениях. С точки зрения обеспечения информационной безопасности, модуль может предоставлять следующие атрибуты метаданных: информация о правообладателе, дата и время съемки, географические координаты и адрес места съемки. Эти данные могут быть полезны для дополнительного анализа и контроля, а также для обеспечения соответствия собранных изображений установленным стандартам и нормам безопасности.

4. Нейросетевой модуль детекции синтетических изображений (рис. 15): С использованием предварительно обученной сверточной нейронной сети (CNN) этот модуль определяет вероятность того, является ли изображение синтетическим или настоящим. Предназначен для автоматизированной фильтрации и классификации графического контента, собранного из профилей сотрудников в социальных сетях, и предоставляет информацию для последующего анализа и принятия решений. Перед тем как подать изображение на анализ в нейросетевую модель, модуль изменяет его размер до необходимого для работы нейросети, осуществляет перекодирование изображения в числовой формат, и проводит нормализацию цветов (рис. 15).

5. Модуль расчета риска: Этот модуль осуществляет анализ и оценку риска деструктивности синтетических изображений на основе разрабатываемой методологии и характеристик самих изображений. Используя установленные критерии и параметры, модуль определяет степень потенциальной угрозы, связанной с синтетическим контентом, что обеспечивает более эффективное принятие решений и реагирование на обнаруженные изображения.

6. Информирование оператора о превышениях установленных пределов риска и визуализация работы программы. Этот компонент отвечает за уведомление оператора о превышениях заданных порогов риска, предусмотренных в системе. Визуализация работы программы включает в себя создание и вывод на экран отчетов об

инцидентах в формате PDF. Кроме того, предусмотрено предоставление рекомендаций по управлению риском, которые также могут быть включены в отчеты. Дополнительно, информация об инцидентах записывается в базу данных, обеспечивая долгосрочное хранение данных для последующего анализа и мониторинга.

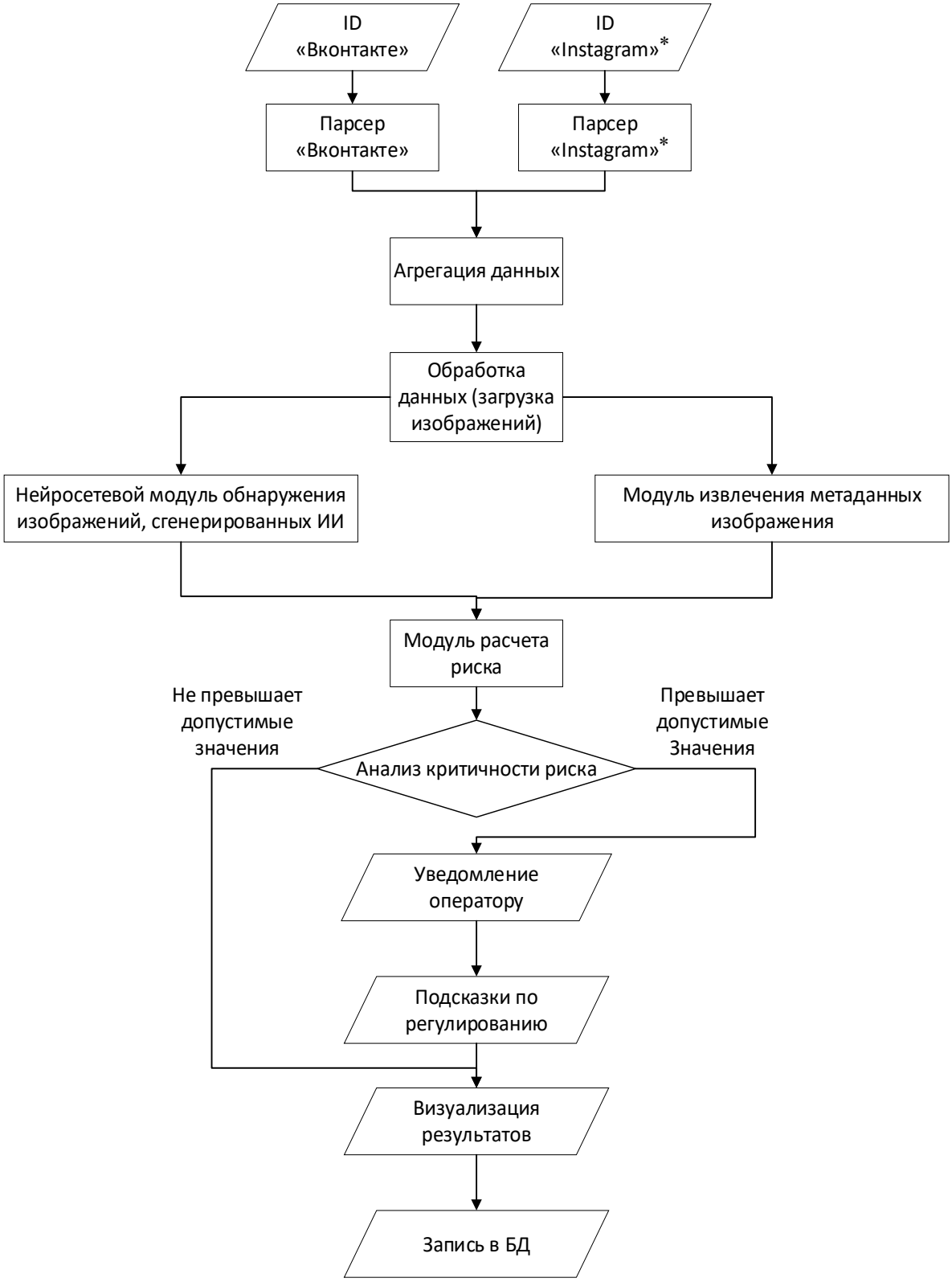

Рис.11. Архитектура программного комплекса

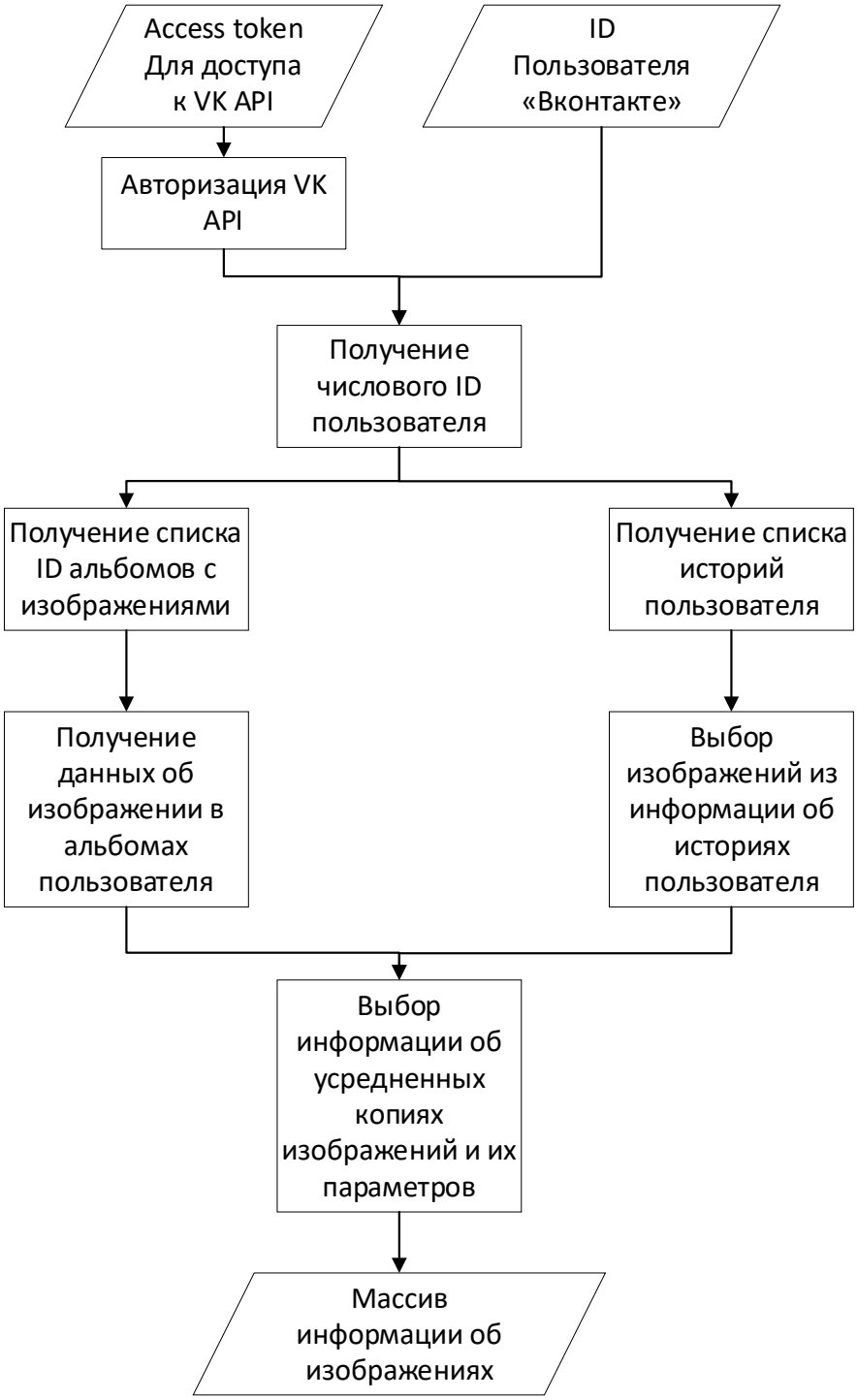

Рис. 12. Парсер сети «Вконтакте»

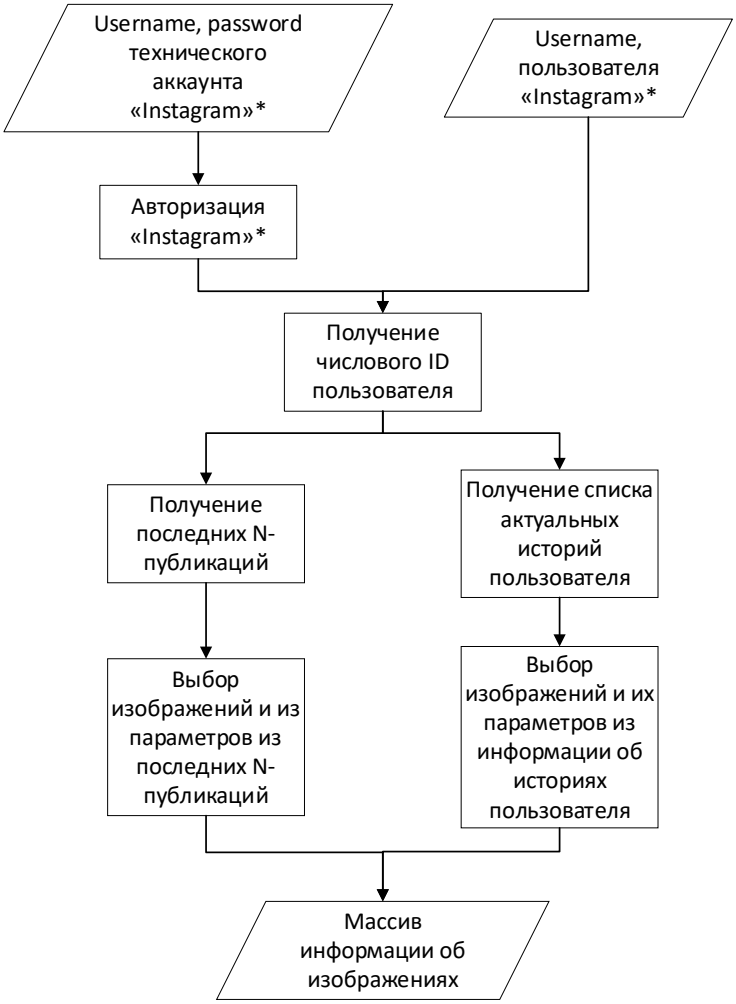

Рис. 13. Парсер сети «Instagram»*

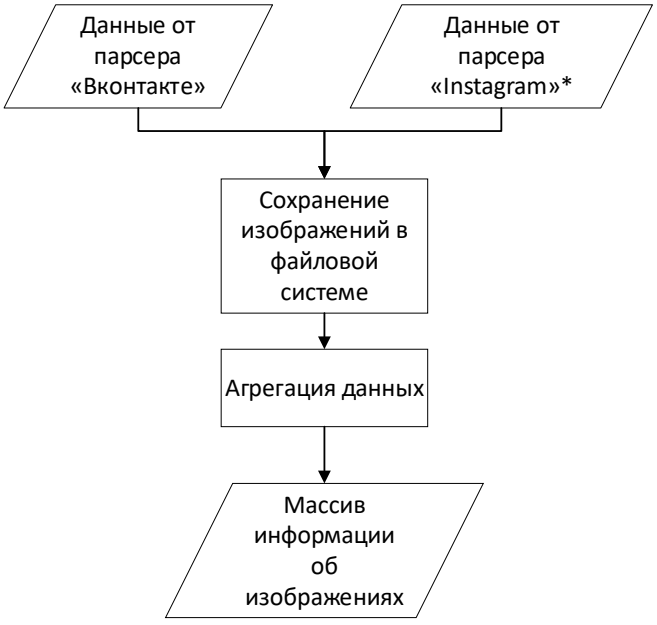

Рис. 14. Модуль обработки данных, полученных от парсера

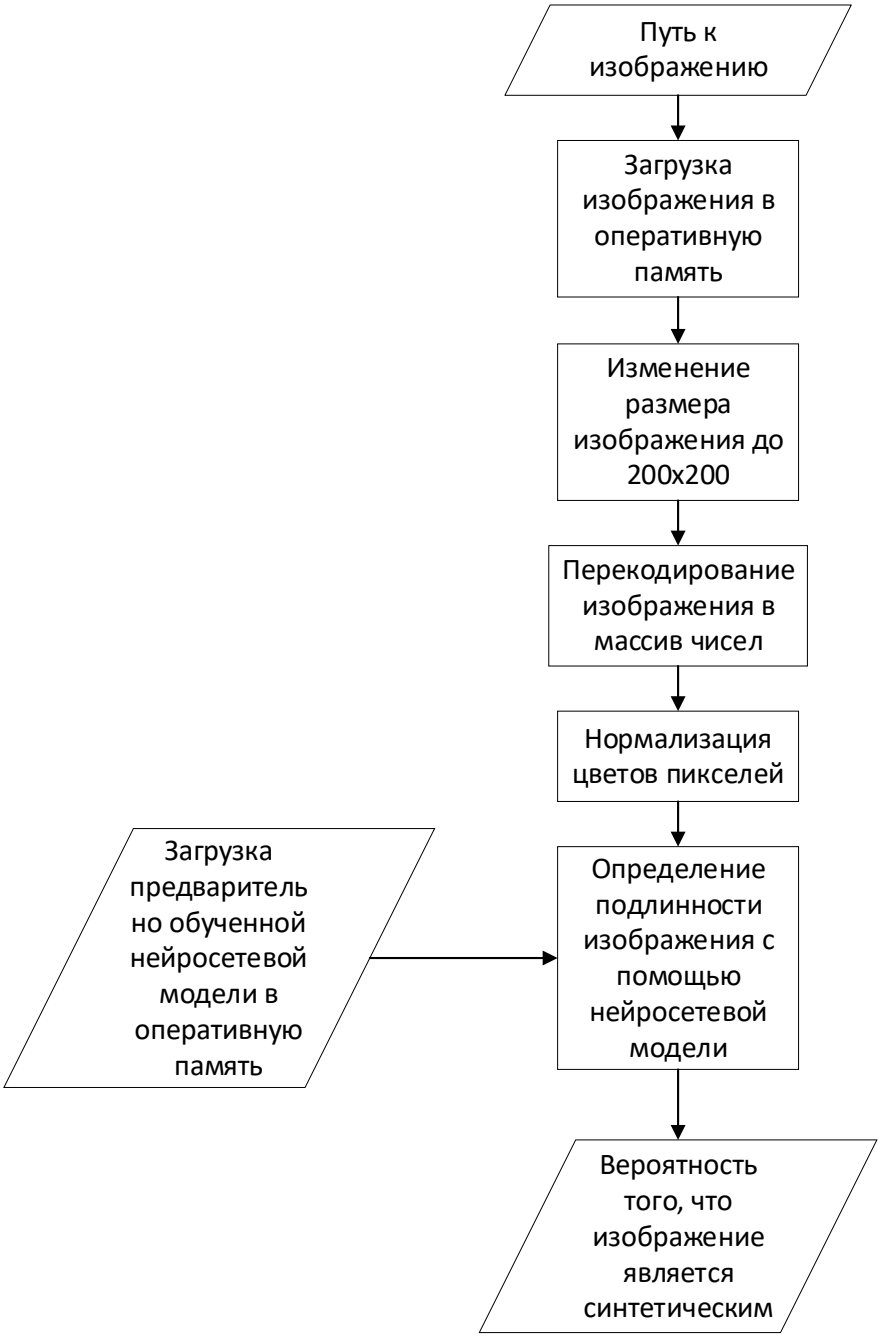

Рис. 15. Нейросетевой модуль распознания синтетических изображений

Заключение

За пределами настоящей публикации пока осталась конкретика осуществления риск-анализа контентов. Мышление не на уровне абстрактных программных модулей, а на базе реальных сущностей является ключом к успеху исследования. В нашей работе такой сущностью является риск, как мера безопасности, и здесь необходимо определиться:

- оставить в основе количественные оценки восприимчивости (лайки, репосты и т.п.) контента, постаравшись отсечь искусственно создаваемую популярность (боты и т.п.);

- предложить принципиально новые риск-оценки, вытекающие из существа построения, распространения и восприятия рассматриваемого класса контентов.

Существенным инструментальным подспорьем здесь могут выступить искусственные нейронные сети, обучаемые массивами деструктивных контентов, если, конечно, таковые удастся создать и

получится понять по каким метрикам следует искать аналогии, позволяющие оценить степень опасности распространяемого контента.

* Деятельность Meta (соцсети Facebook и Instagram) запрещена в России как экстремистская.

Список литературы

1. Дипфейк: невинная технология для развлечения или угроза современному обществу? URL: https://russiancouncil.ru/analytics-and-comments/analytics/dipfeyk-nevinnaya-tekhnologiya-dlya-razvlecheniya-ili-ugroza-sovremennomu-obshchestvu/ (дата обращения: 19.10.2023).

2. Courts and lawyers struggle with growing prevalence of deepfakes. URL: https://www.abajournal.com/web/article/courts-and-lawyers-struggle-with-growing-prevalence-of-deepfakes (дата обращения: 19.10.2023).

3. Experts Worry Deep Fakes Could Threaten 2024 US Elections. URL: https://themorningnews.com/news/2023/05/30/experts-worry-deep-fakes-could-threaten-2024-us-elections/ (дата обращения: 19.10.2023).

4. Elections in Africa: AI generated deepfakes could be the greatest digital threat in 2020. URL: https://paradigmhq.org/deepfakes/ (дата обращения: 19.10.2023).

5. Эксперт «лаборатории касперского» Тушканов: мошенники начали использовать нейросети и дипфейки. URL: https://tsargrad.tv/news/jekspert-laboratorii-kasperskogo-tushkanov-moshenniki-nachali-ispolzovat-nejroseti-i-dipfejki_782902 (дата обращения: 19.10.2023).

6. Проверяй, не доверяй: как отличить видео от дипфейка. URL: https://vc.ru/services/494595-proveryay-ne-doveryay-kak-otlichit-video-ot-dipfeyka (дата обращения: 19.10.2023).

7. New Steps to Combat Disinformation. URL: https://blogs.microsoft.com/on-the-issues/2020/09/01/disinformation-deepfakes-newsguard-video-authenticator/ (дата обращения: 19.10.2023).

8. Deepfake Detection Challenge. URL: https://www.kaggle.com/competitions/deepfake-detection-challenge/data (дата обращения: 19.10.2023).

9. Аудитория восьми крупнейших соцсетей в России в 2023 году: исследования и цифры. URL: https://ppc.world/articles/auditoriya-vosmi-krupneyshih-socsetey-v-rossii-issledovaniya-i-cifry/ (дата обращения: 19.10.2023).

10. Как продавать в Одноклассниках в 2023 году. URL: https://webim.ru/blog/kak-prodavat-v-odnoklassnikah-v-2023-godu/ (дата обращения: 19.10.2023).

11. Почему социальные сети стали неотъемлемой частью жизни? URL: https://www.tutorialspoint.com/why-has-social-media-become-an-integral-part-of-life (дата обращения: 23.10.2023).

12. ГОЛОСОВЫЕ СООБЩЕНИЯ URL: https://3snet.info/news/a-new-promotional-channel-voice-messages/ (дата обращения: 23.10.2023).

13. ВЦИОМ: более 60% опрошенных россиян пользуются голосовыми сообщениями URL: https://tass.ru/obschestvo/17546909 (дата обращения: 23.10.2023).

14. Ежедневно на Spotify загружается более 60 000 треков URL: https://telegra.ph/Ezhednevno-na-Spotify-zagruzhaetsya-bolee-60-000-trekov-02-26 (дата обращения: 23.10.2023).

15. Музыкальные предпочтения россиян URL: https://iom.anketolog.ru/2017/07/07/muzykal-nye-predpochteniya-rossiyan#:~:text=Как%20показывают%20результаты%20опроса%2C%2063%2C8%25,дороге%20%2F%20в%20транспорте%20(63%2C6%25) (дата обращения: 23.10.2023).

16. Музыка и политика: как музыка формирует политический дискурс URL: https://beatsphere.ru/muzyka-i-politika-kak-muzyka-formiruet-politicheskij-diskurs/ (дата обращения: 23.10.2023).

17. Роскомнадзор запустил интеллектуальную систему отслеживания незаконного контента в интернете «Окулус» URL: https://habr.com/ru/news/716464/ (дата обращения: 23.10.2023).

18. Популярные соцсети: что изменилось в 2023 году URL:

https://gb.ru/blog/populyarnye-sotsseti/ (дата обращения: 23.10.2023).

19. Аудитория восьми крупнейших соцсетей в России в 2023 году: исследования и цифры URL: https://ppc.world/articles/auditoriya-vosmi-krupneyshih-socsetey-v-rossii-issledovaniya-i-cifry/#ok (дата обращения: 23.10.2023).

20. Рэп-музыка и ментальное здоровье: как тексты могут влиять на нашу психику и эмоциональное состояние URL: https://dzen.ru/a/ZCVvRMWD9l4bwFcN (дата обращения: 23.10.2023).

21. Retaj Matroud Jasim, Tayseer Salman Atia. An evolutionary-convolutional neural network for fake image detection // Indonesian Journal of Electrical Engineering and Computer Science vol.29 №3 2023. С. 1657-1667 URL: https://www.researchgate.net/publication/368885302_An_evolutionary-_convolutional_neural_network_for_fake_image_detection (дата обращения: 23.10.2023).

22. Jordan J. Bird, Ahmad Lotfi. CIFAKE: Image classification and explainable identification of ai-generated synthetic images // Nottingham Trent University, Nottingham UK, 2023. 12 с. URL: https://www.researchgate.net/publication/ 369541020_CIFAKE_Image_Classification_and_Explainable_Identification_of_AI-Generated_Synthetic_Images (дата обращения: 23.10.2023).

23. Васильев Н. А. Новый социальный мир: Описание основных социальных сетей. // Гуманитарии юга России 2021. № 5(51). С. 31-52 URL: https://cyberleninka.ru/article/n/novyy-sotsialnyy-mir-opisanie-osnovnyh-sotsialnyh-setey/viewer (дата обращения: 23.10.2023).

24. Социальные сети в России: цифры и тренды, весна 2023. URL: https://vc.ru/social/727573-socialnye-seti-v-rossii-cifry-i-trendy-vesna-2023 (дата обращения: 15.11.2023).

25. Комаленков К. Топ-5 мошеннических схем с использованием искусственного интеллекта в 2023 году! URL: https://vc.ru/u/1902918-lirik-komalenkov-kirill/822889-top-5-moshennicheskih-shem-s-ispolzovaniem-iskusstvennogo-intellekta-v-2023-godu (дата обращения: 29.10.2023).

26. Оспанова М. Факты | В чём реальная опасность искусственного интеллекта. URL: https://factcheck.kz/texnologii/fakty-v-chyom-realnaya-opasnost-iskusstvennogo-intellektahttps://factcheck.kz/texnologii/fakty-v-chyom-realnaya-opasnost-iskusstvennogo-intellekta/ (дата обращения: 29.10.2023).

27. Вопилов Ю. Карты, деньги и дипфейки: как мошенники используют технологии для вымогательств. URL: https://incrussia.ru/understand/deepfake-brandmonitor/ (дата обращения: 29.10.2023).

28. Как распознать сгенерированную нейросетью. URL: https://ya.zerocoder.ru/kak-raspoznat-sgenerirovannuju-nejrosetju-kartinku/ (дата обращения: 13.11.2023).

29. Vanshika Jakhar. Top 12 AI Content Detector Tools in 2023. URL: https://www.safalta.com/online-digital-marketing/online-marketing-tools/top-ai-content-detector-tools-in-2023 (дата обращения: 13.11.2023).

Воронежский государственный технический университет
Voronezh State Technical University

Поступила в редакцию 07.11.2023

Сведения об авторах

Остапенко Александр Григорьевич – д-р техн. наук, профессор, заведующий кафедрой, Воронежский государственный технический университет, e-mail: alexanderostapenkoias@gmail.com

Боков Илья Александрович – студент, Воронежский государственный технический университет, e-mail: alexanderostapenkoias@gmail.com

Лихобабин Сергей Викторович – студент, Воронежский государственный технический университет, e-mail: alexanderostapenkoias@gmail.com

Ясенко Дмитрий Сергеевич – студент, Воронежский государственный технический университет, e-mail: alexanderostapenkoias@gmail.com

Чапурин Евгений Юрьевич – ассистент, Воронежский государственный технический университет, e-mail: alexanderostapenkoias@gmail.com

GOAL-SETTING OF PROJECT ACTIVITIES TO CREATE TOOLS FOR AUTOMATED IDENTIFICATION AND RISK ANALYSIS OF DESTRUCTIVE CONTENT AFFILIATED WITH CORPORATE PERSONNEL

A.G. Ostapenko, I.A. Bokov, S.V. Likhobabin, D.S. Yasenko, E.Y. Chapurin

Video, audio and graphic content of social networks are considered as a factor in ensuring information security of corporations. The goal-setting of project activities is carried out to create automated tools for identifying and risk analysis of the above-mentioned content, including parsing of social network resources, selection of collected content based on signs of destructiveness and their risk analysis to develop recommendations for delimiting access to corporate information. The relevance of the project activity is evaluated, analogues are investigated, architecture and algorithms of the created tools are proposed. The existing contradictions are formulated, the research tasks arising from them and the expected results with their corresponding novelty, practical value and theoretical significance. The prospects of organizing a risk analysis of the studied content and the use of its results to develop recommendations on the differentiation of corporate access to information are discussed.

Keywords: content, social network, risk, parsing, personnel, corporation, access.

Submitted 07.11.2023

Information about the authors

Alexander G. Ostapenko – Dr. Sc. (Technical), Professor, Head of Department, Voronezh State Technical University, e-mail: alexanderostapenkoias@gmail.com
Ilya A. Bokov – Student, Voronezh State Technical University, e-mail: alexanderostapenkoias@gmail.com
Sergey V. Likhobabin – Student, Voronezh State Technical University, e-mail: alexanderostapenkoias@gmail.com
Dmitry S. Yasenko – Student, Voronezh State Technical University, e-mail: alexanderostapenkoias@gmail.com
Evgeny Y. Chapurin – Assistant, Voronezh State Technical University, e-mail: alexanderostapenkoias@gmail.com