ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБЩЕОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО
ОБРАЗОВАНИЯ

«Тульский государственный университет»

На правах рукописи

САВЕНКОВ ПАВЕЛ АНАТОЛЬЕВИЧ

ИНТЕЛЛЕКТУАЛЬНАЯ СИСТЕМА АНАЛИЗА ПОВЕДЕНЧЕСКОГО ПРОФИЛЯ ПОЛЬЗОВАТЕЛЯ С ИСПОЛЬЗОВАНИЕМ МАШИННОГО ОБУЧЕНИЯ

Специальность 2.3.5. Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей

Диссертация на соискание ученой степени
кандидата технических наук

Научный руководитель:

д.т.н., профессор

Ивутин А.Н.

Тула - 2022

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Актуальность темы исследования. Многократно возросшие вычислительные мощности позволили реализовать на практике интеллектуальные алгоритмы обработки данных, применение которых ранее было ограниченно ввиду невозможности оперативного получения результатов за приемлемое время. Особый интерес здесь представляют системы способные формировать нетривиальные выводы на основе накопленных массивов информации.

В настоящее время сформировался отдельный класс решений, осуществляющий мониторинг пользовательской активности, в основе которого лежат методы машинного обучения. На данный момент не существует устоявшегося термина для подобного класса решений, однако компания Gartner именует данные системы как UBA (англ. User Behavior Analytics — анализ поведения пользователей). На основе выполнения различных действий пользователей UBA системы формируется их поведенческий профиль, а отклонения от типовой модели поведения обнаруживаются на основе сравнения с историческими данными.

Однако, применение методов машинного обучения позволяет выявлять скрытые зависимости, ранее не обнаруживаемые существующими алгоритмами. Не является исключением и поиск изменений в поведении пользователей по большим наборам коротких текстовых данных в UBA системах. Смещение акцента анализа пользовательского профиля в сторону применения мобильных рабочих станций, делает актуальной задачу реализации подобных систем в мобильном исполнении. Однако пользовательские данные, обрабатываемые на мобильных рабочих станциях, существенно отличаются от аналогичных получаемых с персональных компьютеров сотрудников, что требует проведения дополнительных исследований связанных с выбором источников данных, их предобработки и очистки от информационного шума. При этом специфика мобильного поведения накладывает дополнительные условия на формирование данных и

не позволяет в полной мере использовать существующие наработки стационарных UBA систем. Диапазоны временного окна, объемы и тип накапливаемой информации так же являются параметрами, существенно влияющими на качество анализа.

Наряду с указанными выше особенностями мобильных UBA систем остается не до конца решенная задача устранения временного лага между совершенными нетиповыми действиями и реакцией на них при сохранении высокого качества получаемой результирующей информации, а также слабая подстройка современных UBA систем под изменения индивидуальных особенностей пользователей с течением времени. Кроме того, характерны высокие трудозатраты и высокая чувствительность получаемого результата ввиду ручной обработки больших массивов пользовательских данных экспертами при использовании существующих UBA систем.

Перспективным направлением развития мобильных UBA систем является формирование эталонного поведенческого профиля, представляющего динамически меняющуюся предобработанную ретроспективу пользовательской активности за определенный период времени, на основе которого производится поиск поведенческих отклонений, при помощи программных средств, использующих методы и алгоритмы машинного обучения.

Таким образом, существует противоречие, заключающееся в потребности оперативно идентифицировать нетиповые сценарии поведения пользователей в условиях доступности различных информационных каналов и отсутствии методов выделения полезной информации при постоянно возрастающем объеме информационного потока.

Предлагаемые в данной работе решения обеспечивают организацию сбора данных пользователей с мобильных устройств, их предварительную обработку, а применение методов машинного обучения в разработанной интеллектуальной системе, позволяет существенно снизить энтропию исходного объема информации за счет формирования краткого объема

высокоинформативных результирующих значений, при достаточно большом количестве анализируемых входных данных, благодаря чему появляется возможность оперативной идентификации нетиповых сценариев использования мобильного устройства пользователем.

Степень разработанности темы исследования. Актуальным вопросам поиска аномального поведения пользователей с применением программного обеспечения, методов машинного обучения и анализа естественного языка посвящены работы отечественных и зарубежных ученых И.В. Котенко, И.А. Ушакова, М.А. Поляничко, И.В. Машечкина, М.И. Петровского, В.О. Горохова, К.К. Отраднова, В.К. Раева, H Liu, B Lang., de Doncker E. и др. Развитием технологии «Большие данные» занимаются Николенко С.И., Цветков В.Я., Кузнецов С.Д., N.A Ghani, N. Usman, S. Usman, F. Khan, M.A. Jan и др.

В соответствии с вышеизложенным **научная задача** диссертации заключается в сокращении объема обрабатываемой экспертами текстовой информации о деятельности пользователей для повышения оперативности детектирования нетиповых сценариев использования мобильных устройств.

Объектом исследования являются интеллектуальные системы машинного обучения поведенческого анализа деятельности пользователей.

Предметом исследования выступают методы машинного обучения и анализа естественного языка, используемые в интеллектуальных системах машинного обучения поведенческого анализа деятельности пользователей.

Целью работы является сокращение объемов анализируемой экспертами вручную текстовой информации путем разработки интеллектуальной системы анализа поведенческого профиля пользователя с использованием машинного обучения.

Для достижения данной цели в работе поставлены и решены следующие **задачи**:

1. Обзор решений и методов анализа данных для идентификации изменений в поведении пользователей и их применение в современных интеллектуальных системах машинного обучения;

2. Разработка метода предварительной обработки накапливаемой текстовой информации из коротких выборок, обеспечивающего уменьшение информационного шума, отличающегося предварительным формированием оптимальной длинны коротких последовательностей пригодных для дальнейшего анализа длиной от 7 до 100 символов;

3. Разработка метода идентификации нетиповых сценариев использования мобильных устройств пользователями по наборам коротких текстовых данных, отличающегося применением методов машинного обучения, анализа естественного языка (bag-of-words, TF-IDF, Word2Vec, GloVe, BERT) и метрик сходства, обеспечивающего сокращение объемов анализируемой экспертами вручную текстовой информации, собираемой с мобильных устройств пользователей, а так же экономное потребление вычислительных ресурсов;

4. Разработка архитектуры программного комплекса идентификации нетиповых сценариев использования мобильных устройств пользователями, отличающейся расширяемой модульной структурой, обеспечивающей сбор биометрических данных, содержащих пользовательские поведенческие характеристики, и идентификацию нетиповых сценариев использования мобильного устройства на их основе;

5. Разработка программного комплекса, реализующего предложенные методы предварительной обработки накапливаемой текстовой информации из коротких выборок и идентификации нетиповых сценариев использования мобильных устройств, обеспечивающего сбор и анализ биометрических данных, уменьшение информационного шума и экономное потребление вычислительных ресурсов;

6. Экспериментальное исследование разработанных методов и программного комплекса идентификации нетиповых сценариев использования мобильных устройств и анализ полученных результатов.

Научная новизна включает новые научные результаты, полученные в работе, и заключается в следующем:

1. Разработан метод предварительной обработки накапливаемой текстовой информации из коротких выборок, обеспечивающий уменьшение информационного шума, отличающийся предварительным формированием оптимальной длинны коротких последовательностей пригодных для дальнейшего анализа длиной от 7 до 100 символов;

2. Разработан метод идентификации нетиповых сценариев использования мобильных устройств пользователями по наборам коротких текстовых данных, отличающийся применением методов машинного обучения, анализа естественного языка (bag-of-words, TF-IDF, Word2Vec, GloVe, BERT) и метрик сходства, обеспечивающий сокращение объемов анализируемой экспертами вручную текстовой информации, собираемой с мобильных устройств пользователей, а так же экономное потребление вычислительных ресурсов;

3. Предложена и реализована архитектура программного комплекса идентификации нетиповых сценариев использования мобильных устройств пользователями, отличающаяся расширяемой модульной структурой, обеспечивающая сбор биометрических данных, содержащих пользовательские поведенческие характеристики, и идентификацию нетиповых сценариев использования мобильного устройства на их основе.

Практическая значимость.

Разработан программный комплекс, реализующий предложенные методы предварительной обработки накапливаемой текстовой информации из коротких выборок и идентификации нетиповых сценариев использования мобильных устройств, обеспечивающий сбор и анализ биометрических

данных, уменьшение информационного шума и экономное потребление вычислительных ресурсов.

Предложенные метод обнаружения нетиповых сценариев использования мобильного устройства и метод предобработки могут использоваться как в новых системах поведенческого анализа, так и быть интегрированы в уже существующие решения, применяемые для контроля изменений в поведении пользователей и сбора их поведенческой информации.

Задачи, решаемые в данной диссертационной работе, охватывают ключевые сегменты рынка «Сейфнет» в области обработки естественного языка, использования методов машинного обучения и анализа больших данных. По стратегии научно-технологического развития Российской Федерации, задачи направлены на переход к передовым цифровым, интеллектуальным производственным технологиям, роботизированным системам, новым материалам и способам конструирования, создание систем обработки больших объемов данных, машинного обучения и искусственного интеллекта.

Проведенные исследования соответствуют перечню научных исследований и опытно-конструкторских разработок, расходы налогоплательщика на которые в соответствии с пунктом 7 статьи 262 части второй Налогового кодекса Российской Федерации включаются в состав прочих расходов в размере фактических затрат с коэффициентом 1.5, а именно, подпунктам 13, 14 пункта 2 раздела 2: «разработка интеллектуальных систем оперативного реагирования на чрезвычайные ситуации» и «разработка методов и программных средств интеллектуальных систем поддержки принятия решений, в том числе с элементами искусственного интеллекта».

Методы исследования. В работе использовались методы машинного обучения, анализа естественного языка, методология объектно-ориентированного проектирования и подходы, применяемые при разработке программного обеспечения.

Реализация и внедрение результатов работы. Работа по теме диссертации проводилась на кафедре вычислительной техники ТулГУ в рамках проекта ФСИ №13121ГУ/2018 «Разработка автоматизированной системы анализа и контроля деятельности сотрудников», гранта РФФИ №19-37-90111 «Использование методов и алгоритмов анализа данных и машинного обучения в информационных системах», гранта РФФИ №19-41-710003 «Методы и средства автоматической оптимальной адаптации последовательных алгоритмов для исполнения в гетерогенных вычислительных системах», проекта ФСИ «Акселерация ИИ» №4ГАИИС13-D7/72316 «Исследование применимости методов машинного обучения для поиска аномальной активности в поведении пользователей и разработка программного обеспечения для обнаружения поведенческих отклонений».

Полученные результаты исследований внедрены в деятельность ООО «Кадастр Экспресс», ООО «Гуд Фуд», а также используются в образовательном процессе кафедры «Вычислительная техника» ТулГУ, что подтверждается соответствующими актами внедрения.

Соответствие паспорту научной специальности 2.3.5 «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей». Диссертационная работа выполнена в соответствии с пунктом «4. Интеллектуальные системы машинного обучения, управления базами данных и знаний, инструментальные средства разработки цифровых продуктов».

Апробация работы. Основные положения и результаты работы докладывались и получили положительную оценку на следующих всероссийских и международных конференциях: Mediterranean Conference on Embedded Computing «MECO» (Черногория, 2018, 2019, 2020, 2021); Международная научно-практическая конференция молодых ученых «Прикладная математика и информатика: современные исследования в области естественных и технических наук» (Россия, Тольятти, 2019, 2020); International Conference on Swarm Intelligence «ICSI» (Сербия, 2020); 14th

International Symposium Intelligent Systems «Intels» (Россия, Москва, 2020); Оптико-электронные приборы и устройства в системах распознавания образов и обработки изображений «Распознавание» (Россия, Курск, 2019, 2021); Всероссийская научно-техническая конференция интеллектуальные и информационные системы «Интеллект» (Россия, Тула, 2016, 2019, 2021); «Всероссийская научно-практическая конференция им. Жореса Алфёрова» (Санкт-Петербург, 2021); Научно-техническая конференция «Инновационные наукоемкие информационные технологии» (Тула, 2017, 2019, 2020).

Личный вклад автора заключается в выполнении основного объема теоретических и экспериментальных исследований, а также в разработке архитектуры программного комплекса обнаружения нетиповых сценариев использования мобильных устройств пользователями, отличающегося расширяемой модульной структурой, обеспечивающего сбор биометрических данных, содержащих пользовательские поведенческие характеристики.

Автор лично выполнил оформление полученных результатов диссертационной работы в виде публикаций, научных докладов и свидетельств о государственной регистрации программ для ЭВМ.

Все выносимые на защиту научные результаты, в том числе постановка задач, разработка методов, организация экспериментов, анализ экспериментальных данных, основные научные результаты и выводы получены соискателем лично.

Положения, выносимые на защиту:

1. Обзор решений и методов анализа данных для идентификации изменений в поведении пользователей и их применение в современных интеллектуальных системах машинного обучения;

2. Разработка метода предварительной обработки накапливаемой текстовой информации из коротких выборок, обеспечивающего уменьшение информационного шума;

3. Разработка метода идентификации нетиповых сценариев использования мобильных устройств пользователями по наборам коротких

текстовых данных, с использованием методов машинного обучения и анализа естественного языка и метрик сходства, обеспечивающего сокращение объемов анализируемой экспертами вручную текстовой информации, а так же экономное потребление вычислительных ресурсов;

4. Разработка архитектуры программного комплекса идентификации нетиповых сценариев использования мобильных устройств пользователями, отличающейся расширяемой модульной структурой, обеспечивающей сбор биометрических данных, содержащих пользовательские поведенческие характеристики, и идентификацию нетиповых сценариев использования мобильного устройства на их основе;

5. Разработка программного комплекса, реализующего предложенные методы предварительной обработки накапливаемой текстовой информации из коротких выборок и идентификации нетиповых сценариев использования мобильных устройств, обеспечивающего сбор и анализ биометрических данных, уменьшение информационного шума и экономное потребление вычислительных ресурсов;

6. Экспериментальное исследование разработанных методов и программного комплекса идентификации нетиповых сценариев использования мобильных устройств и анализ полученных результатов.

Публикации. По теме диссертации опубликовано 27 статей в различных научных изданиях, в том числе 6 в международных изданиях, входящих в системы цитирования Web of Science и Scopus, 9 статей в российских изданиях из перечня ВАК РФ для публикации основных научных результатов на соискание ученой степени кандидата наук.

На разработанные модули интеллектуальной системы машинного обучения получено 10 свидетельств о государственной регистрации программы ЭВМ.

Структура и объем диссертации. Диссертация состоит из введения, четырех глав, заключения, изложенных на 124 страницах текста, списка использованной научной литературы, включающего 117 наименований

научных трудов на русском и иностранных языках и 2 приложения, и содержит 36 рисунков и 9 таблиц.

Достоверность полученных результатов диссертационной работы. Выводы, полученные в диссертационной работе, являются достоверными и обоснованными, что подтверждается достаточным объемом проанализированных отечественных и иностранных источников по тематике научного исследования, актами о внедрении, выполненными НИР, обсуждением результатов работы на различных научных конференциях российского и международного уровня и согласованностью результатов теоретических и экспериментальных исследований.

1 АНАЛИЗ РАЗВИТИЯ ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМ МАШИННОГО ОБУЧЕНИЯ. ОБЗОР МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПОИСКА ОТКЛОНЕНИЙ В ПОВЕДЕНИИ ПОЛЬЗОВАТЕЛЕЙ

1.1 Существующие системы анализа данных пользователей

В связи с ростом объемов информации [25], обрабатываемой в различных информационных системах [26], появляется задача сокращении объема обрабатываемой экспертами вручную текстовой информации при анализе не типовых сценариев использования мобильных устройств пользователями [27].

Прогресс в сфере информационных технологий оказывает существенное влияние на манеры поведения пользователей и их взаимодействие друг с другом [28]. Наблюдающаяся тенденция автоматизации процессов взаимодействия пользователей, с использованием электронных средств и информационных технологий, задает темп дальнейшего развития автоматизированных систем и отказ от бумажных способов обмена информацией и источников [29].

Возможности существующих систем обработки и анализа текстовых данных достаточно широки, однако большинство решений сосредоточены на обработке информации, собираемой с персональных компьютеров сотрудников, что не позволяет идентифицировать нетиповые сценарии использования мобильных устройств пользователями ввиду наличия определенной специфики [30-33].

Выявление факта обработки сотрудником определенных текстовых данных, целей взаимодействия с мобильным устройством и своевременная идентификация нетиповых сценариев использования мобильного устройства, остается затрудненным в виду недостаточной целевой направленности решений, методов и алгоритмов.

Существуют следующие классы программных решений, осуществляющие анализ данных пользователей:

- DLP (англ. Data Loss Prevention) — специализированное программное обеспечение, предназначенное для защиты компании от утечек информации [34];
- SIEM (англ. Security information management) — программное обеспечение, обеспечивающее анализ событий безопасности в реальном времени, исходящих от сетевых устройств и приложений, и позволяющее осуществлять реагирование на них до наступления существенного ущерба [35];
- ECM (англ. Enterprise content management) — программное обеспечение предназначенное для управления корпоративным контентом. По определению Gartner — стратегическая инфраструктура и техническая архитектура для поддержки единого жизненного цикла неструктурированной информации (контента) различных типов и форматов [36];
- SOAR (англ. security operations, analytics and reporting) — специализированный инструментарий, позволяющий сводить данные об угрозах безопасности из разных источников для последующего анализа [37];
- UBA (англ. User behavior analytics) — система осуществляющая реагирование на различные внутренние изменения в организации, отслеживающая поведение пользователей. Данные системы, на основе поступающих разнородных данных, формируют и анализируют модели поведения человека для дальнейшего обнаружения аномалий, указывающих на потенциальные отклонения от эталонного поведенческого профиля пользователя [38].

Несмотря на то, что перечисленные системы имеют ряд сходств и могут включать в себя некоторые возможности поиска аномального поведения

пользователей, основным типом систем поведенческого анализа являются UBA решения.

UBA-системы, в отличие от DLP, SIEM, ECM, SOAR осуществляют мониторинг широкого спектра действий пользователя и формируют результат, на основе исторических данных о правомерной работе пользователя, а не на основе сформированных экспертами политик безопасности. Цель UBA-систем состоит не в блокировке действий пользователей, а в предоставлении данных аналитической службе с описанием того, почему выявленные действия являются аномальными для конкретного пользователя [39]. Системы DLP, SIEM, ECM, SOAR не предназначены для решения данных задач, так как реагируют на статические угрозы, заданные экспертами, например выгрузку пользователем данных на сторонний ресурс или их копирование с рабочего устройства на внешний съемный носитель.

SIEM системы имеют возможность сбора и агрегации данных, получаемых из различных программных инструментов и ИТ-систем, анализируют их и отправляют оповещения в режиме реального времени для групп безопасности. Традиционно SIEM решения не имеют в своем составе технологии поведенческой аналитики. Их идентификация осуществляется при помощи определяемых администратором системы правил корреляции.

Решения UBA разрабатываются для устранения данного недостатка и уже доказали свою эффективность в идентификации неизвестных ранее отклонений [40]. В последние годы аналитики определили, что расширение базовых возможностей SIEM систем, при помощи решений другого класса, поможет более эффективно осуществлять идентификацию отклонений различного уровня. На сегодняшний день многие SIEM системы интегрируют в свои продукты встраиваемые UBA решения.

Для выявления признаков нетиповой активности UBA системы на основе методов машинного обучения формируют модели поведения пользователей и выявляют отклонения различного рода. Решения класса UBA могут осуществлять анализ обрабатываемых на устройстве данных, контроль

используемых ими устройств, мониторинг работающих приложений и т.д. UBA системы формируют модель, в которой доступ к данным, активность рабочих станций и сетевая активность привязывается к конкретным пользователям.

Отличительной особенностью UBA-решений от других систем является наличие возможности оценки уровня отклонений в поведении пользователя, по которому администратор системы может получить краткую обобщенную информацию о деятельности конкретного пользователя для своевременного реагирования [41].

Системы класса UBA решают следующие задачи:

- Аналитика данных, полученных из различных источников, с использованием методов и алгоритмов машинного обучения;
- Своевременное обнаружение отклонений от сформированного ранее эталонного поведенческого профиля пользователя;
- Консолидация данных, полученных из разных источников;
- Установка приоритетов и маркеров для различных событий;
- Уменьшение объемов, анализируемой вручную экспертами, информации за счет предоставления краткой, агрегированной, структурированной информации по найденным отклонениям.

Любое ядро современной UBA системы включает в себя методы и алгоритмы для работы с большими данными [42]. UBA решения формируют модель эталонного поведения пользователя при его взаимодействии с устройством. Кроме того, данные системы имеют возможность формирования поведенческого профиля не одного пользователя, а целой группы лиц. UBA решения могут быть реализованы в виде полноценной самостоятельной системы или же могут дополнять функционал существующих систем, которые не имеют в своем составе данных возможностей.

В качестве информационных ресурсов могут выступать не только структурированные данные журналов системы, но и такие источники как:

- Системы DLP, SIEM и др.;
- Переписка пользователей;
- Входящие исходящие сообщения;
- Набираемый текст;
- Перемещения (GPS);
- Время использования приложений;
- Социальные сети.

Общий принцип работы систем поведенческого анализа может быть представлен в виде следующей схемы (рис. 1.1).

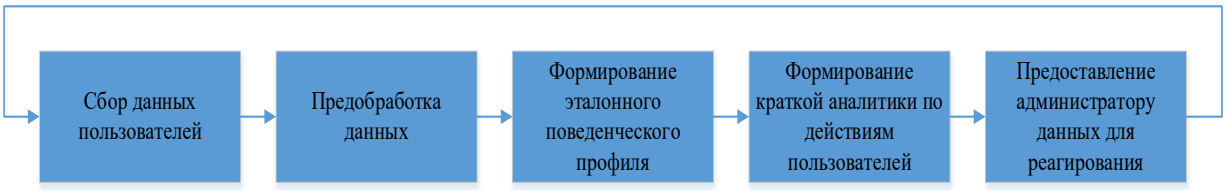


Рис. 1.1 – Схема общего принципа работы систем поведенческого анализа

Определение UBA систем компанией Gartner включает в себя три основных атрибута (рис 1.2):

1. Варианты использования - решения UBA сообщают о поведении объектов и пользователей системы. Они обнаруживают, отслеживают и предупреждают администратора системы об отклонениях. Решения UBA предоставляют множество вариантов использования, в отличие от систем, которые выполняют специализированный анализ, такой как мониторинг доверенных хостов, обнаружение мошенничества и т. д.
2. Источники данных — решения UBA могут получать данные из различных источников. Такими источниками могут так же выступать и системы SIEM, DLP, SOAR, ECM и др.
3. Аналитика — решения UBA идентифицируют отклонения при помощи аналитических методов, включая машинное обучение, статистические модели, правила и сигнатуры угроз.

Рис. 1.2 – Диаграмма основных атрибутов современной UBA системы

Аналитика больших данных позволяет компаниям извлекать выгоду из неструктурированных и слабоструктурированных массивов данных [43].

Следующие возможности отличают системы анализа поведения пользователей от инструментов информационной безопасности:

- Отчетность и визуализация — визуализация больших данных, для извлечения требуемой информации и ее предоставление в наглядном виде;
- Хранилище больших данных — эффективное хранение данных, их обработка и анализ с использованием масштабируемых систем;
- Контекст — возможность анализа данных с учетом контекста и привязки к пользователю или целой группе лиц;
- Широкие функциональные возможности — применение исторических данных для идентификации нетиповых сценариев использования устройства.

В настоящее время существует большой объем зарубежных и отечественных решений, предназначенных для анализа данных пользователей:

- Определение скомпрометированных пользовательских аккаунтов:
  - Secure Portal (Group IB);
  - Kasperskiy Fraud Prevention (Лаборатория Касперского);
  - Exabeam Advanced Analytics (Exabeam);
  - Splunk UBA (Splunk).
- Определение и предотвращение инсайдерских угроз:
  - Solar Dozor (Solar Security);
  - Гарда БД (МФИ Софт);
  - Контур информационной безопасности КИБ (SearchInform);
  - Microsoft Advanced Threat Analytics (Microsoft);
  - ObserveIT Insider Threat Intelligence (ObserveIT).
- Мониторинг сотрудников и их прав доступа:
  - StaffCop Enterprise (Атом Безопасность);
  - HPE AcrSight UBA (HPE/MicroFocus);
  - IBM QRadar UBA (IBM).

Основными недостатками существующих решений, осуществляющих сбор данных пользователей, является отсутствие возможности получения администратором системы краткой агрегированной информации по действиям выбранного человека или группы лиц за определенный временной период и автоматического оповещения при идентификации системой нетипового сценария использования устройства, а также:

- высокая стоимость;
- высокие трудозатраты на настройку и обслуживание системы;
- высокая чувствительность результатов, связанная с формированием экспертами жестких статических правил и политик;

- высокий уровень требований к квалификации обслуживающего персонала.

Достоинствами программных комплексов анализа данных пользователей являются [44]:

- гибкость в настройке и интеграции с имеющимися UBA решениями;
- широкий функционал, предоставляющий администратору системы высокоинформативные результаты.

Таким образом, существующие системы в области анализа данных пользователей, не обеспечивают в полном объеме решение ключевых задач (не идентифицируют нетиповые сценарии использования мобильного устройства сотрудником), а также используют для решения данных вопросов возможности интеграции с имеющимися UBA системами для расширения собственного функционала.

1.2 Анализ методов поиска отклонений в поведении пользователей по наборам текстовых данных

Обработка больших объемов текстовых данных пользователей с целью извлечения из них знаний, в общем случае подразделяется на следующие классы [45]:

- Классификация (обучение с учителем);
- Определение тональности, определение эмоциональной окраски;
- Кластеризация (обучение без учителя);
- Реферирование/аннотирование;
- Генерация текста;
- Сравнение, идентификация изменений.

Решение задачи классификации подразумевает наличие обучающей выборки, однако в связи с тем, что поведение пользователя не имеет четких факторов, идентифицирующих сценарий использования мобильного

устройства как типовой или нетиповой, формирование такого набора данных является невозможным.

Определение тональности, эмоциональной окраски является частным случаем задачи классификации и имеет те же свойственные методу недостатки.

При кластеризации текстов появляется возможность деления пользователей без обучающей выборки на разные неименованные группы, что в свою очередь не формирует полного представления о поведении конкретного пользователя.

Реферирование и аннотирование текстов позволяет извлекать основную информацию из больших объемов данных, что в свою очередь предоставляет администратору системы возможность быстрого своевременного изучения ключевых фраз, используемых пользователем. Однако по набору ключевых фраз, без определения администратором системы правил корреляции, невозможно автоматизировано определить правомерность активности деятельности пользователя, а определение таких правил делает невозможным выявление нетиповых сценариев использования устройства и требует их постоянного обновления.

Генерация текста не используется в выявлении отклонений в поведении и применяется в основном при построении чат ботов или иных консультационных автоматизированных систем [46].

Для решения задачи анализа пользовательских текстов на предмет сходства и, следовательно, идентификации нетиповых сценариев использования мобильного устройства пользователем по текстовым выборкам целесообразно использовать векторизацию документов и их дальнейшее сравнение при помощи метрик сходства [47].

Сравнение пользовательских текстов и выявление нетиповых сценариев использования мобильного устройства пользователем по кратким наборам текстовых данных является актуальным. Применение данного подхода позволяет сократить объем обрабатываемой экспертами вручную текстовой

информации при анализе не типовых сценариев использования мобильного устройства пользователем, акцентировав внимание только на данных деятельности пользователей, осуществляемой в определенные временные периоды и имеющий явные отклонения от эталонного поведения.

Реферирование и аннотирование пользовательских текстов может быть дополнительно применено для получения более подробной картины об изменениях в поведении пользователей и дальнейшего принятия управленческих решений экспертом [48].

Комбинированное использование методов анализа естественного языка для формирования векторных представлений и метрик сходства позволяет существенно сократить объем обрабатываемой экспертами вручную текстовой информации, анализируемой для выявления отклонений в поведении пользователей и идентификации нетиповых сценариев.

Для выявления уровня различий между двумя пользовательскими наборами текстовых данных, при помощи методов анализа естественного языка, требуется их преобразование к числовому векторному виду и построение словаря для дальнейшего сравнения при помощи метрик сходства. После разбиение исходных пользовательских текстов на токены, требуется представить их в числовом виде тем самым осуществив векторизацию [49].

Одним из базовых методов векторизации считается числовое кодирование. При данном типе векторизации каждому токену назначается свой код, отображающий частоту использования слова в тексте. Другим подходом является сопоставление токена набору чисел (вектора). Существует несколько вариантов отображения токена в векторном виде, ими являются разреженные вектора типа «One Hot Encoding» и плотные векторные представления Embedding [50]. Ограничением представления One Hot Encoding является потеря информации о позиции слова в тексте. Размерность такого вектора ограничивается объемом словаря, а все значения равны нулю, кроме того, который соответствует определенному токену. В случае представления токенов в виде Embedding векторов, их размерность гораздо

ниже, чем у «One Hot Encoding», а в качестве числовых значений могут быть использованы не только нули и единицы, но и любые целые и дробные числа.

При построении векторов формата «One Hot Encoding», для повышения быстродействия, его длина ограничивается, путем использования словарей, содержащих наиболее часто встречаемые слова, так как с увеличением длины вектора уменьшается и быстродействие его дальнейшей обработки. Примерами таких словарей могут служить «Oxford 3000» и «Merriam Webster 3000 Core Vocabulary Words» [51]. Недостаток использования разреженных векторов «One Hot Encoding» заключается не только в его высокой размерности, но и затрудненным хранения в памяти и недостаточно эффективной обработке современными процессорами и графическими ускорителями. При использовании «Embedding» векторов количество обрабатываемых данных уменьшается, что положительно сказывается на производительности рабочей станции, обрабатывающей массивы пользовательских данных [52]. Однако при использовании таких представлений не ясно какие числовые значения должны формировать данный вектор, так как значения плотного векторного представления определяется в процессе обучения нейронной сети, что дополнительно требует наличия размеченного набора данных для обучения. На первом этапе элементы вектора инициализируются случайными числами, на следующем осуществляется обучения с учителем при помощи метода обратного распространения ошибки, в процессе которого значения плотного векторного представления итерационно изменяются пока в них не сформируются значения, требуемые для решения определенной задачи.

Для формирования «Embedding» векторов требуются огромные массивы размеченных данных и как правило большие вычислительные и значительные временные ресурсы. По данной причине одним из подходов в решении данной проблемы является использование предварительно обученных векторных представлений слов таких как: GloVe (Global Vectors) [53] – сформирован

Стэндфортским университетом, Word2Vec [54] – разработана компанией Google, RusVectores и др.

Для формирования векторных представлений пользовательских текстов требуется наличие словаря, содержащего токен и его числовое значение. Формируемый по пользовательскому тексту вектор может быть преобразован из текстовой формы при помощи таких методов как:

- Мешок слов (англ. bag-of-words, BOW) [55];
- TF-IDF (англ. TF — term frequency, IDF — inverse document frequency)[56];
- Word2Vec;
- GloVe;
- BERT.

Вопросы использования представленных методов и формирование модели поведения будут рассмотрены в следующих главах.

1.3 Нормализация данных в задачах поиска аномального поведения пользователей по их текстовым наборам

Процесс формирования результатов для предоставления администратору системы (эксперту) информации о нетиповых сценариях использования мобильных устройств в общем случае состоит из нескольких стадий. Исходными данными могут являться сообщения в мессенджерах, электронная почта или другой вводимый и получаемый набор текстов. Данные наборы требуют предварительной обработки для получения корректных результатов анализа.

В зависимости от типа входных данных, используются различные принципы предварительной обработки текста. Для уменьшения размерности словаря и повышения корректности результатов анализа применяется приведение токенов к единому регистру, что позволяет записывать в словарь одинаковые по написанию, но различные по регистру токены как один. Для

разбиения исходных текстов на слова или буквы (токены), для дальнейшего построения словаря применяется токенизация. Без построения словаря невозможно составить числовые векторы обрабатываемых текстов, в связи с отсутствием числового представления каждого токена. Для восстановления исходной формы слова, применяется стемминг и лемматизация [57]. При удалении стоп слов удаляются часто используемые слова в языке, это могут быть как предлоги и другие часто используемые токены. Идея, лежащая в основе удаления стоп слов, заключается в том, что, в исходном тексте удаляются слова с низкой информативностью, что позволяет составить словарь для дальнейшей векторизации с наиболее высокоинформативными словами. Для удаления шума осуществляется удаление спец символов, цифр, знаков, которые могут мешать дальнейшему анализу текстов.

Нормализация текста является важным этапом предобработки зашумленных данных, таких как комментарии в социальных сетях, комментарии в блогах, где преобладают сокращения, орфографические ошибки и смайлы. Так же нормализация повышает точность классификации в крайне неструктурированных текстах [58]. Данные подходы помогают формировать векторы меньшей размерности, тем самым создавать менее нагруженную и более быстродействующую модель.

В связи с тем, что методы анализа естественного языка используют словарь для построения числовых векторов текстовых представлений, требуется его формирование из полученной текстовой выборки пользовательских текстов. Без создания словаря токенов невозможно осуществить анализ содержимого на предмет наличия отклонений в поведении пользователя.

Создание словаря без предварительной обработки текстов возможно, однако качество результатов дальнейшего анализа будет значительно снижено, в связи с присутствием в словаре специальных системных символов, знаков пунктуации. Так же снижение качества дальнейшей обработки происходит из-за присутствия в словаре слов, состоящих из букв в различных

регистрах, так как два одинаковых слова в различных регистрах имеют два разных идентификатора в словаре и являются для методов анализа естественного языка различными токенами.

Порядок выполнения операций предварительной обработки не имеет значения. Создание словаря проводится только после проведения предварительной обработки исходного текста.

1.4 Постановка задачи исследований диссертационной работы

Задача сокращения объема обрабатываемой экспертами вручную текстовой информации, при анализе нетиповых сценариев использования мобильного устройства пользователем, может быть сформулирована следующим образом: даны наборы текстов пользователя, на естественном языке, последовательно выбранные за одинаковые по длительности временные интервалы. Требуется определить сходство между текстами и установить нормальный порог отличий между двумя сформированными результирующими векторами. Если разница между сходством выборок превышает установленный порог, то считается, что в поведенческих характеристиках пользователя, которому принадлежат данные текстовые наборы, содержатся отклонения, а сценарий использования мобильного устройства является нетиповым.

Анализ пользовательских текстов требуется для идентификации нетиповых сценариев использования мобильных устройств пользователями. В связи с этим требуется создание метода, позволяющего идентифицировать нетиповые сценарии использования мобильного устройства пользователем по наборам коротких текстов, реализующего предобработку данных и дальнейшее сравнение векторных представлений текстовых наборов при помощи метрик сходства и предоставляющего эксперту краткую агрегированную результирующую информацию о деятельности в выбранном временном периоде.

Практическое решение вопросов повышения быстродействия систем поиска отклонений в поведении пользователей связано с большой размерностью формируемых словарей, присутствием зашумленных данных и анализа текстовых наборов в краткие временные промежутки. Это порождает задачу поиска путей снижения размерности формируемых словарей для дальнейшего снижении времени анализа и разработки эффективных методов для своевременной идентификации отклонений в поведении пользователей [59].

Задача идентификации нетиповых сценариев использования мобильных устройств пользователями по наборам коротких текстов является затруднительной ввиду присутствия в исходных текстовых массивах большого количества зашумленных данных.

В случае использования мобильных устройств как источника получения текстовой информации, для обеспечения высокого уровня быстродействия требуется определить длины выбираемых и в дальнейшем анализируемых текстовых сообщений пользователей, так как от их длины напрямую зависит обрабатываемый объем информации и количество зашумленных данных требующих дополнительную предобработку.

Таким образом, в ходе проводимых исследований с целью сокращения объемов, анализируемых вручную экспертами текстовых данных пользователей и идентификации нетиповых сценариев использования мобильных устройств, необходимо решить следующие частные задачи (рис 1.3):

1. Осуществить выбор методов машинного обучения, анализа естественного языка, метрик сходства и расстояния, применительно к задаче поиска отклонений в поведении пользователей по коротким наборам зашумленных текстовых данных;
2. Определить оптимальную длину строк единичных пользовательских текстов и размеры временных диапазонов для формирования

выборки, на основе собранного впервые набора текстовых данных пользователей;

3. Реализовать экспериментальный образец программного комплекса, осуществляющий сбор текстовых данных пользователей, их хранение, дальнейшую обработку, и анализ с целью обнаружения нетиповых сценариев использования мобильного устройства пользователем;
4. Выполнить экспериментальную проверку разработанного программного комплекса идентификации нетиповых сценариев использования мобильных устройств по наборам коротких текстов, на основе которой разработать методику применения реализованных методов.

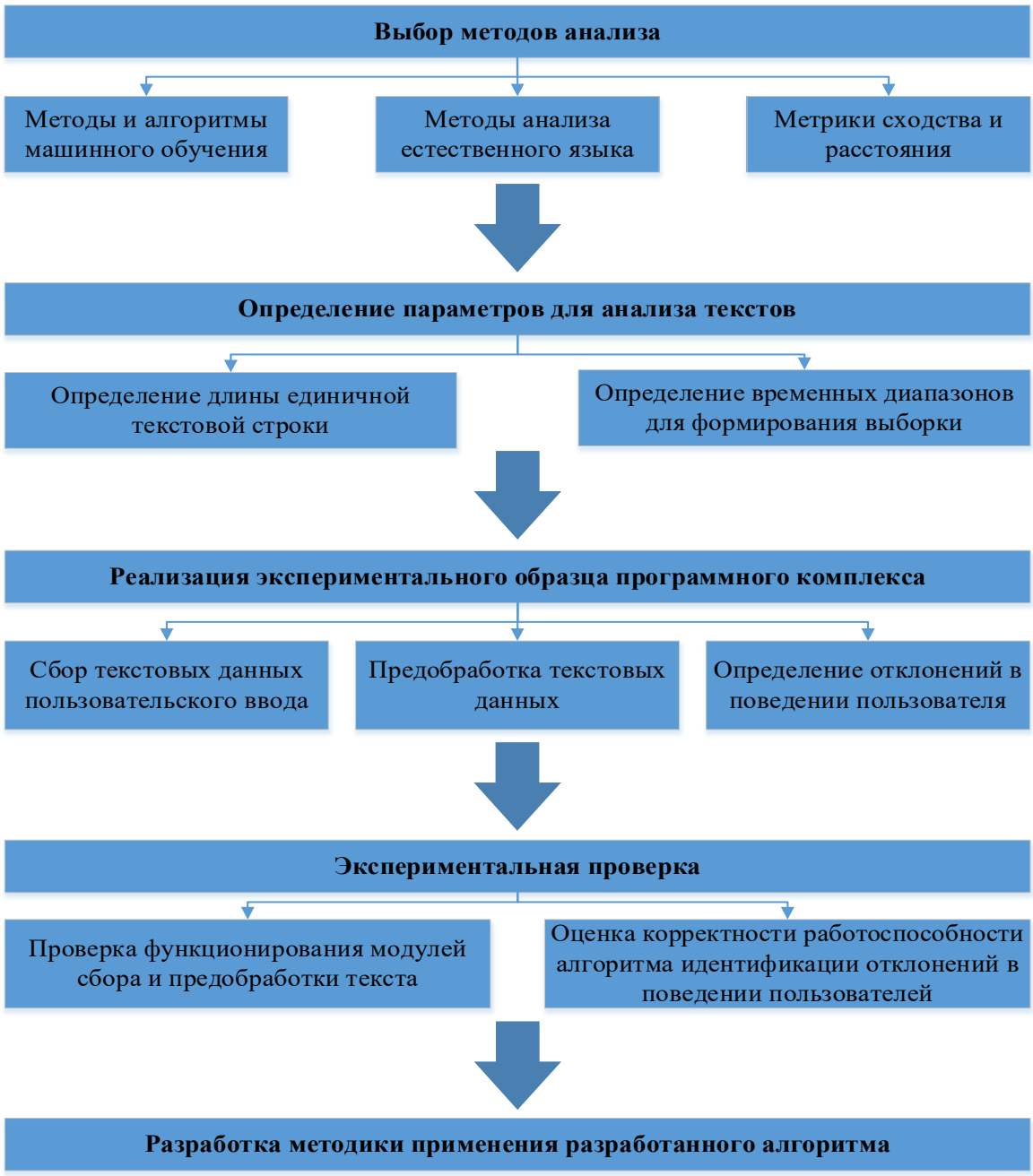


Рис. 1.3 – Структура задач исследования

В данной работе был использован впервые собранный набор текстовых данных, состоящий из 4 953 300 сообщений, набираемых пользователями на мобильных устройствах.

Для получения результирующих значений анализа и агрегации больших массивов текстовых данных пользователей будут использованы методы машинного обучения, анализа естественного языка, метрики сходства и расстояния. Применительно к задаче идентификации нетиповых сценариев использования мобильных устройств пользователями, перспективными

являются подходы, обеспечивающие высокий уровень быстродействия при достаточной возможности идентификации таких сценариев.

По данной причине в следующей главе рассматриваются особенности применения методов машинного обучения, анализа естественного языка, метрик сходства и расстояния применительно к задаче сравнения наборов коротких текстов.

1.5 Выводы

В результате проведенного обзора современных систем анализа данных и поведения пользователей, были определены ключевые решения и подходы, используемые в обработке, анализе и сравнении текстовых данных, для идентификации нетиповых сценариев использования мобильных устройств пользователями и сформированы следующие выводы:

1. При сравнении возможностей и функционала систем анализа данных пользователей показана важность и актуальность развития и разработки как самостоятельных UBA систем, так и решений для интеграции с другими программными продуктами;
2. Опираясь на результаты исследований выделены существующие проблемы в области систем анализа поведения пользователей, основанные на специфике обрабатываемых наборов коротких текстовых данных;
3. Была выявлена целесообразность разработки программного комплекса сбора данных и анализа поведенческих характеристик пользователей для оценки нетиповых сценариев использования мобильного устройства, по кратким наборам пользовательских текстов;
4. На основе обзора существующих систем выявлена потребность определения этапов и параметров предобработки данных;
5. В результате проведенного анализа систем анализа поведения выявлено, что в качестве исходных данных при анализе текстов может быть использована переписка в мессенджерах, электронная почта, социальные сети и другие текстовые наборы данных;

6. В качестве технологий поиска аномалий в поведении пользователей целесообразно применять методы машинного обучения, методы анализа естественного языка, метрики сходства и расстояния.

Полученные выводы являются основанием проведения дальнейших исследований.

2 ФОРМИРОВАНИЕ МОДЕЛИ ПРЕДСТАВЛЕНИЯ ПОВЕДЕНИЯ ПОЛЬЗОВАТЕЛЯ

В диссертационной работе осуществляется исследование и разработка методов идентификации нетиповых сценариев использования мобильных устройств пользователей при помощи методов машинного обучения и анализа естественного языка по их наборам коротких текстовых данных. Наборы текстовых данных пользователей являются неструктурированными ввиду специфики ввода текстов на мобильных устройствах. Для обработки таких данных при помощи методов машинного обучения требуется их предварительная обработка и приведение к числовому виду. Формирование набора признаков для дальнейшего анализа заключается в извлечении знаний из наборов пользовательских текстов и их представление в векторном виде, содержащим признаки и их весовые значения. Набор пользовательских сообщений $N = (t_1, \ldots, t_i)$ может быть представлен в виде числовой матрицы. Каждый пользовательский текст t_i в свою очередь может быть представлен в виде числового вектора фиксированной длины $t_i = [a_{1i}, a_{2i, \ldots} a_{ji}]$ содержащего соответствующие весовые значения.

Формирование результирующих векторных представлений для дальнейшего извлечения признаков осуществляется при помощи рассмотренных ранее методов bag-of-words, TF-IDF, Word2Vec, GloVe, BERT.

2.1 Предварительная обработка текстовых данных и их очистка от информационного шума

Целью предварительной обработки текста, является его очистка от информационного шума с целью повышения качества идентификации нетиповых сценариев использования мобильных устройств пользователями. Помимо повышения качества идентификации, так же повышается и быстродействие обрабатывающих методов и алгоритмов, в связи с уменьшением количества обрабатываемых данных, векторов, и уменьшения

их размерности. Для очистки пользовательских данных применяются следующие принципы предобработки текстов [60-64]:

1. *Удаление стоп-слов.* При удалении стоп слов удаляются часто используемые слова в языке. Это могут быть как предлоги и другие часто используемые токены. Идея, лежащая в основе удаления стоп слов, заключается в том, что, в исходном тексте удаляются слова с низкой информативностью, что позволяет составить словарь для дальнейшей векторизации с наиболее высокоинформативными словами. Удаление стоп слов помогает так же формировать векторы меньшей размерности, тем самым формировать менее нагруженную модель;

2. *Лемматизация* схожа со стеммингом, однако при лемматизации восстанавливается исходная форма слова, а при стэмминге слово обрезается и корень может быть найден некорректно;

3. *Приведение токенов к единому регистру* применяется для уменьшения размерности словаря и повышения корректности результатов анализа. Если не использовать приведение токенов к нижнему регистру, одинаковые по написанию слова, но различные по регистру будут записаны в словарь как отдельные токены;

4. *Удаление шума. У*даление спец символов, цифр, знаков, которые могут мешать дальнейшему анализу текстов.

Нормализация текста является важным этапом для зашумленных текстов, таких как комментарии в социальных сетях, комментарии в блогах, где преобладают сокращения, орфографические ошибки и смайлы. Так же нормализация повышает точность классификации в неструктурированных текстах [65]. Создание словаря без предварительной обработки текстов возможно, однако качество результатов дальнейшего анализа будет значительно снижено [66], в связи с присутствием в словаре специальных системных символов, знаков пунктуации. Так же снижение качества дальнейшей обработки происходит из-за присутствия в словаре слов,

состоящих из букв в различных регистрах, так как два одинаковых слова в различных регистрах имеют два разных идентификатора в словаре и являются для методов анализа естественного языка различными токенами. Порядок выполнения операций предварительной обработки не имеет значения. Создание словаря проводится только после проведения предварительной обработки исходного текста.

2.2 Определение длины анализируемой строки пользовательских текстов

В ходе экспериментального исследования было установлено, что наиболее подходящей длиной пользовательских текстов, для извлечения из них высокоинформативных результатов и идентификации отклонений в поведении является диапазон от 7 до 100 символов [67].

Установлено, что пользовательские тексты длиной менее 7 символов, чаще всего, состоят из стоп-слов и различного информационного шума в виде ошибочно набранных фраз и распространённых бесконтекстных словосочетаний (рис 2.1), имеющихся в выборках 98% пользователей, что не позволяет идентифицировать нетиповые сценарии использования мобильного устройства.

	Text	ApplicationName		Text	ApplicationName		Text	ApplicationName		Text	ApplicationName
49	86	Zenly	181	а	TikTok	556	wink	Samsung Internet	979	Да	WhatsApp
50	8684	Zenly	182	Сам	VK	557	?goo	Почта Mail.ru	980	днс	Samsung Internet
51	eno	Яндекс.Музыка	183	ору	TikTok	558	?g	Почта Mail.ru	981	Ок	WhatsApp
52	•4•	Яндекс.Музыка	184	ужс	TikTok	559	ды	VK	982	Не	VK
53	ощон	Google Play Маркет	185	ужс	TikTok	560	Ужас	VK	983	Ага	VK
54	зани	VK	186	но	TikTok	561	Дома	WhatsApp	984	•8	Homsbox
55	длд	Samsung Internet	187	но	TikTok	562	Дома	VK	985	днс	Samsung Internet
56	sma	Google Play Маркет	188	ахах	TikTok	563	?	VK	986	Да	WhatsApp
57	е	Smart Life	189	)	TikTok	564	Да	VK	987	Ну	VK
58	ук	Google Play Маркет	190	????	TikTok	565	Да	VK	988	Дааа	VK
59	ук	Google Play Маркет	191	да?	TikTok	566	be	VK	989	Opel	Infocar
60	Ужс	VK	192	ЧТ...	TikTok	567	Ты	VK	990	47	Infocar
61	84	СберБанк	193	(	TikTok	568	Крч	VK	991	Нет.	Viber
62	700	СберБанк	194	ну	TikTok	569	??	Telegram	992	Чка	VK
63	и ти	СберБанк	195	ну,	TikTok	570	Play	Загрузчик OnePlus	993	И ч/	VK
64	800	СберБанк	196	х	TikTok	571	in	Google Play Маркет	994	kiss	TikTok
65	gear	Google Play Маркет	197	хотя	TikTok	572	??	Telegram	995	kiss	TikTok

Рис. 2.1 – Выборка вводимых текстов длиной менее 7 символов

Тексты, длина которых более 100 символов (рис. 2.2), в основном, содержат пользовательские заметки, многократно скопированный текст, набираемые и отправляемые в деловом стиле сообщения, информационную

рассылку, скопированные Web ссылки и т.д. Данные сообщения так же негативно влияют на качество дальнейшего анализа. По данным сообщениям из наборов пользовательских текстов невозможно извлечь уникальную поведенческую информацию ввиду ее отсутствия [68].

286	первые два слайда - суть проекта смета - 3 слайд (эксперты 10 человек по 5000 рублей, денежные призы победителям, брендированные сувениры лучшим участникам, съемка фильма) как будем искать участ
331	Освещение студенческих событий: Специфика работы в соц сетях. Освещение мероприятий. Актуальность (подача информации, оформление) Если прошло мероприятие, новость либо день в день, либо утр
543	@all ???? Еще раз приветствую Кратко и по делу: кидаю вам презентацию, как и обещала, там вы найдете все то, о чем я сегодня рассказывала. Если кого-то не было - советую ознакомиться (особенно, если
544	Подскажите, а Вы работаете с удалением негативных отзывов? просто есть конкретные примеры, где действительно ситуация была плохая, а есть отзывы, в которых говорится о том, чего в действительно
545	здравствуйте, я получил Ваш заказ, но был очень расстроен. потому что он не соотвествует заявленным характеристикам. В этих лампах нет CanBus. Они очень часто моргают
663	Друг, добрый день, у меня к тебе просьба. У меня нет знакомых людей, владеющих китайским языком. Мог ли бы ты перевести мне фразу на китайский язык? " Добро пожаловать в музей" Спасибо
664	Good afternoon, friend, I have a request for you. I don't know people who speak Chinese. Could you translate the phrase for me into Chinese? "Welcome to the museum" Thank you
672	Поздравляю тебя с днём рождения!!! Оставайся такой же прекрасной мамой и лучшей женой! Пусть по жизни с тобой в ногу идёт удача, а белая полоса никогда не заканчивалась! Поздравляю!??????????
673	Привет, друг. снимаю тебе видео с проблемой твоей новым датчиком кислорода. Ошибка светится красным. Ошибка P0136-01 - O2 Sensor Circuit Open (Bank 1 Sensor 2)
1116	Здравствуйте! Прошу вернуть мне деньги за заказ, мне пришло совершенно не то, что я заказал. Размер не тот, который я заказал, рисунок не совпадает с тем, что я заказывал. Рисунок
1211	Работа таргетолога в ВК. Таргетированная реклама Вконтакте, работа таргетолога, включая рекламный бюджет на продвижение. По договору №3/ .-76 от 10 ноября 2021 года. ИГК 0000000009 \TC
1212	Работа таргетолога по контекстной рекламе Яндекс. Директи и таргетированной рекламе Youtube. Работа таргетолога, директолога по продвижению контента, включая рекламный бюджет на продвижение
1213	Оператор для съемок анимации. Оператор для съемок анимации в технике перекладка на станке.№ 1-2021-ИРИ от 16 июня 2021 года ИГК № 000000A309121P090002
1214	https://login.aliexpress.com/?flag=1&return_ ao-global.aliexpress.com/export/ae/EditRecipientWireless?_' ~3&refreshConfirmOrder=true&spm=a1z65.placeorder.0.0&_currency=RUB&_lang=
1215	Бритва Зубная щётка Зарядка от ноута Зарядка от тедефона Наушники Павербанк Часы Ноут Подставка Мышка Духи Банки Студак Ключи Тапки Документы (бумаги) Футболка

Рис. 2.2 – Выборка вводимых текстов длиной более 100 символов

В процентном соотношении количество предложений, имеющих длину в диапазоне от 1 до 7 символов составляет 25,55% (1 256 701 записей), от 7 до 100 символов ~ 69% (3 294 412 записей), от 100 до 200 символов 1,85% (87 331 записей), более 200 символов ~1,98% (93 511 записей) от общего объема набора пользовательских текстовых данных ввода (4 731 955 записей). Диаграмма распределения длин пользовательских текстов в анализируемом наборе данных представлена на рис. 2.3.

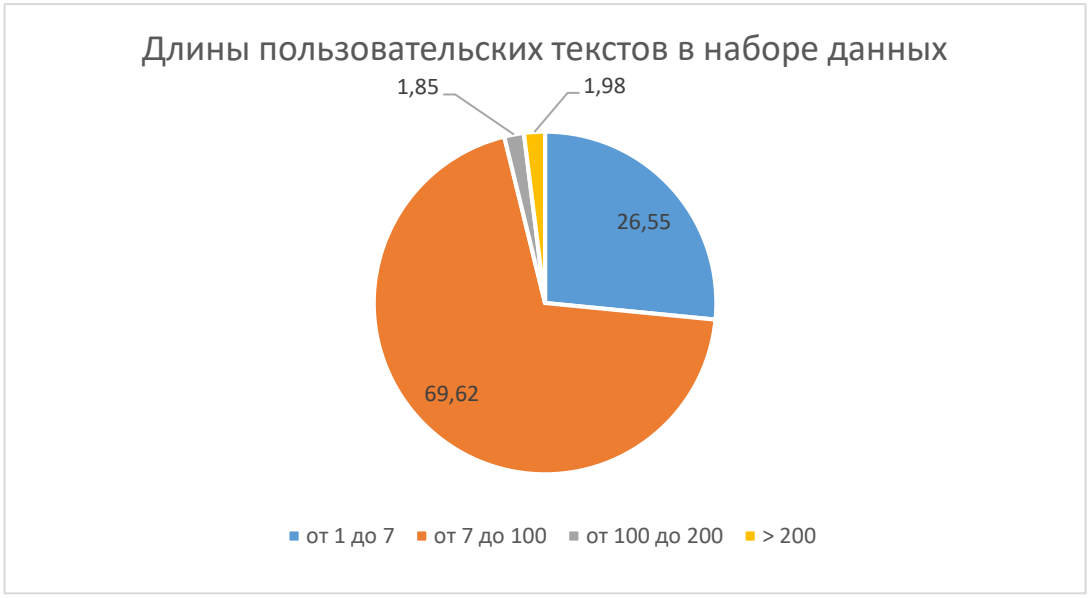


Рис. 2.3 –Диаграмма соотношения длин пользовательских текстов

По количеству символов предложения длиной от 7 до 100 символов занимают основную часть от всей выборки [69], что позволяет использовать их для дальнейшего анализа без снижения качества результатов (рис.2.4)

Рис 2.4 –Диаграмма концентрации количества пользовательских текстов и их длин

Таким образом текстовые сообщения длиной более 100 символов встречаются в выборке крайне редко ввиду специфики пользовательского ввода на мобильных устройствах и не учитываются при анализе нетиповых сценариев использования мобильного устройства пользователем.

2.3 Частотные модели векторного представления

2.3.1 Модель представления «мешок слов»

Одним из наиболее простых методов отображения текстовых данных в виде векторного представления является «Мешок слов» (bag-of-words). Как было отмечено в первой главе, данный метод формирует вектор без дополнительной нормализации повторяющихся слов [70]. Метод «мешок слов» является упрощенным способом представления текста в числовой вид, использующийся в обработке естественного языка и информационном поиске.

В данной модели текст представляется в виде мультимножества находящихся в нем слов (Таблица 2.1).

Таблица 2.1. Формирование векторов по пользовательским текстам методом «Мешок слов»

Индекс вектора	Слово/термин	Значение вектора по документу (кол-во упоминаний слова в тексте)	
		Документ 1 «системы поведенческого анализа пользователей»	Документ 2 «решения для поведенческого анализа данных пользователей»
1	системы	1	0
2	поведенческого	1	1
3	анализа	1	1
4	пользователей	1	1
5	решения	0	1
6	для	0	1
7	данных	0	1

При такой языковой обработке формируется векторное представление из текстовых данных для отображения лингвистических свойств текстов. Недостатком модели «Мешок слов» является отсутствие возможности сохранения семантики слов и их последовательности. Так же для дальнейшего сравнения документов требуется словарь, и размерность таких векторов может быть крайне высокой, так как в случае формирования словаря по имеющимся текстам, его объем будет увеличиваться относительно количеству новых слов в векторизуемых данных. Для уменьшения размерности анализируемых векторов можно использовать словари наиболее часто употребляемых слов или принудительно сокращать длину векторов перед анализом.

2.3.2 Модель представления «TF-IDF»

При использовании статистической меры TF-IDF появляется возможность автоматической оценки важности слова в документе, являющегося частью коллекции набора текстовых данных.

Метод bag of words формирует векторы частоты используемых слов по документам без корректировки веса термина в случае его повторяемости в различных документах. В отличии от метода bag of words, TF-IDF изменяет вес слов, наиболее часто встречающихся в текстах. Общая формула метода извлечения признаков TF-IDF представлена далее [71]:

$$TF - IDF = TF \cdot IDF \tag{2.1}$$

где:

TF (Term Frequency) – отношение числа вхождений некоторого слова к общему числу слов в документе;

IDF (Inverse Document Frequency - Обратная частота документа) — это мера того, сколько информации предоставляет слово, является ли оно общим или редким во всех документах.

TF рассчитывается по следующей формуле:

$$tf(t,d) = \frac{n_t}{\sum_k n_k} \tag{2.2}$$

где n_t – число вхождений слова t в документ,

Знаменатель – общее число слов в данном документе.

IDF используется для расчета веса редких слов во всех документах в корпусе. Слова, которые редко встречаются в корпусе, имеют высокий балл IDF. Формула нахождения обратной частоты документа представлена далее:

$$IDF(t,D) = \log \frac{|D|}{|\{d \in D : t \in d\}|} \tag{2.3}$$

где:

|D| – общее число документов в корпусе.

$|\{d \in D: t \in d\}|$ - количество документов где слово t используется ($tf(t,d) \neq 0$).

Если слово не находится в корпусе, то это приведет к делению на ноль. По данной причине принято преобразовывать знаменатель к виду

$1 + |\{d \in D: t \in d\}|$.

В общем виде формула расчета TF-IDF выглядит следующим образом:

$$tfidf(t,d,D) = tf(t,d) \cdot idf(t,D) \quad (2.4)$$

Высокий вес в tf – idf формируется при высокой частоте термина в данном документе и низкой частоты использования термина в совокупности всех документов. Таким образом фильтруются общие термины [72]. TF-IDF дает большие значения для менее частых слов в корпусе документа. Значение TF-IDF высокое, когда оба значения IDF и TF высокие, т.е. слово встречается редко во всем документе, но часто встречается в текущем документе. TF-IDF, как и bag of words не учитывает семантическое значение слов.

2.4 Нейросетевые модели векторного представления

2.4.1 Модель представления Word2Vec

Для повышения качества сравнения векторов и дальнейшего поиска аномального поведения требуется учитывать семантику слов в анализируемых предложениях. Для сохранения семантических связей применяются модели Word2Vec [73]. Каждое слово в словаре кодируется не частотным признаком, а вектором Embedding, с сохраненной семантической связью.

Архитектура Word2Vec состоит из трех слоев. Входной слой принимает одно слово в формате one hot encoding (каждое слово кодируется бинарным вектором, содержащим одну единицу, которая представляет позицию слова в словаре). Длина вектора one hot encoding равна длине словаря. Второй слой – слой Embedding, представляет собой матрицу размерностью NxP, где N размер словаря, P-гиперпараметр подбираемый эмпирически. Выходной слой размером Nx1, где N размер словаря. Каждый из нейронов данного слоя выдает вероятность принадлежности входящего слова к другим словам. На рис. 2.5 представлена визуализация модели skip-gram Word2Vec [74].

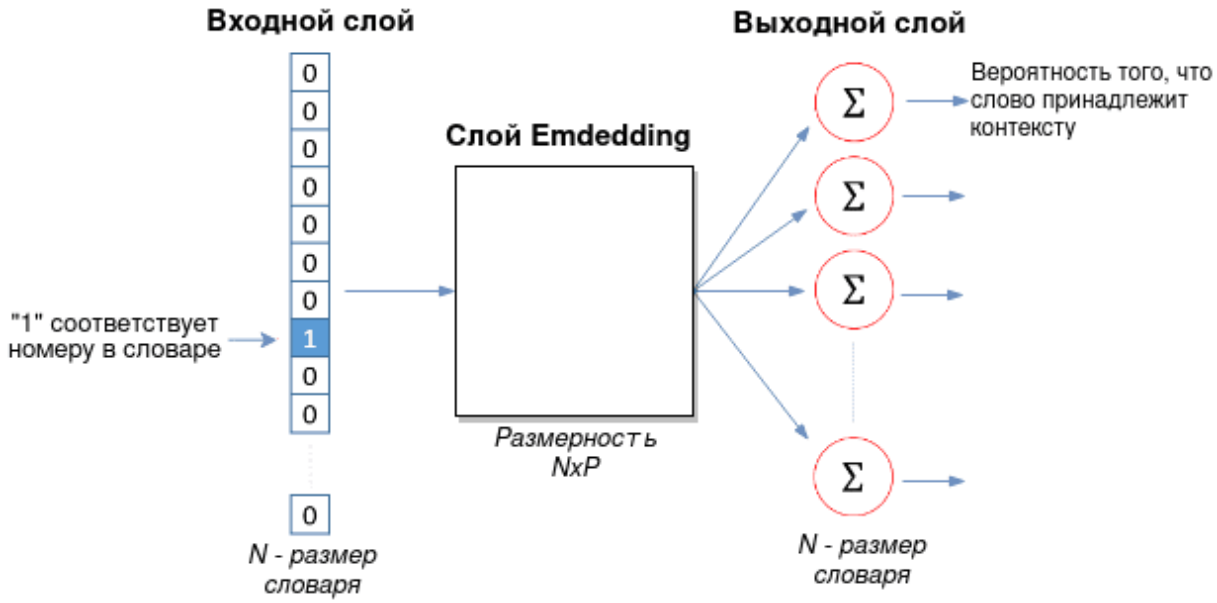


Рис 2.5 – Визуализация применяемой модели skip-gram Word2Vec

2.4.2 Модель распределенного представления слов GloVe

Помимо представленной ранее модели word2vec имеются так же и другие embedding модели. Самой популярной альтернативой Word2Vec является GloVe (Global Vectors) модель, предложенная лабораторией Стендфордского университета, сочетающая в себе черты SVD разложения и Word2Vec. Метод GloVe предоставляет возможность получения семантических связей между словами из матрицы совместной встречаемости.

Имея корпус, содержащий V слов, матрица совместного использования X будет иметь вид VxX, где i-я строка и j- й столбец обозначает, сколько раз слово i встречалось вместе со словом j. Пример матрицы совместной встречаемости может выглядеть следующим образом. На рис. 2.6 представлен пример матрицы совместного использования.

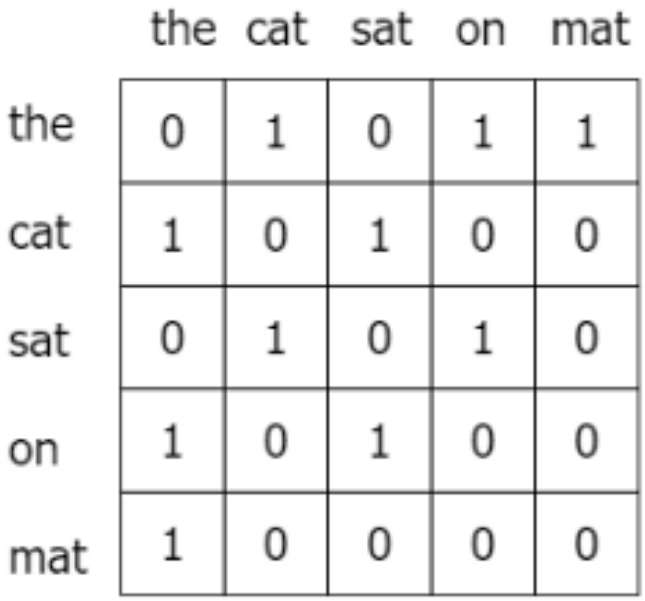


Рис. 2.6 – Матрица совместного использования

Применение данной модели так же возможно для формирования эталонного поведенческого профиля пользователя и идентификации нетиповых сценариев использования им мобильного устройства.

2.4.3 Модель представления BERT

BERT технология, основанная на нейронных сетях, используется для поиска различий слов в контексте, при обработке естественного языка [75]. Технология BERT основана на архитектуре механизма внимания. Благодаря данному механизму модели, построенные на данной архитектуре, лучше находят закономерности необходимые для решения задач. Энкодер получает на вход и обрабатывает набор векторов, проводя их через слой внутреннего внимания и далее – через нейронную сеть прямого распространения, пока не передает свой выход следующему энкодеру. Механизм внимания значительно улучшает качество работы метода, позволяя концентрироваться на релевантных частях входных последовательностей.

2.5 Выводы

В результате анализа моделей представления поведения пользователей, для формирования поведенческого профиля и идентификации нетиповых сценариев использования устройств, были получены следующие выводы:

1. Для повышения качества результатов идентификации нетиповых сценариев использования мобильного устройства пользователем требуется предварительная обработка пользовательских текстов и их очистка от информационного шума, включающая в себя удаление стоп – слов, стемминг, лемматизацию, приведение токенов к единому регистру и удаление шума. Получено, что порядок выполнения операций предварительной обработки не влияет на результаты анализа;

2. Было установлено, что наиболее подходящей длиной пользовательских текстов для идентификации нетиповых сценариев использования мобильного устройства является диапазон от 7 до 100 символов, составляющий ~ 69% общего объема выборки и включающий в себя только релевантные токены, содержащие поведенческую информацию, а не ошибочно набранный или скопированный текст;

3. В результате анализа частотных и нейросетевых моделей представления получено, что частотные BOW и TF-IDF не учитывают семантику слов пользовательских текстов, что негативно сказывается на результатах идентификации нетиповых сценариев использования мобильных устройств пользователями. Установлено, что для получения высокоинформативного результата идентификации нетиповых сценариев использования мобильных устройств требуется использовать нейросетевые модели (Embedding);

4. Установлено, что в интеллектуальных системах применяются модели Embedding такие как Word2Vec, Glove и BERT, учитывающие семантику слов. Однако технология BERT лучше находит закономерности необходимые для решения задач, ввиду наличия «механизма внимания».

3 МЕТОД ИДЕНТИФИКАЦИИ НЕТИПОВЫХ СЦЕНАРИЕВ ИСПОЛЬЗОВАНИЯ МОБИЛЬНЫХ УСТРОЙСТВ ПОЛЬЗОВАТЕЛЯМИ

Идентификация нетиповых сценариев использования мобильных устройств по заранее сформированным константным шаблонам поведения невозможна ввиду того, что выполняемые пользователями задачи динамически изменяются на протяжении всего времени взаимодействия с мобильным устройством, что требует постоянной актуализации поведенческих данных [76]. Даже незначительная смена рода деятельности отражается на поведенческом профиле пользователя. Идентификация нетиповых сценариев позволяет сократить объем обрабатываемой экспертами информации за счет акцентирования их внимания на временных промежутках, в которых была обнаружена нетиповая активность, сформировав при этом облако ключевых тематик общения в пределах данных интервалов.

В связи со спецификой использования мобильных устройств, временной интервал формирования текстовых наборов пользовательских данных, для последующей идентификации типа сценария использования, может быть вариативен. Для получения корректных результатов анализа следует формировать векторные представления по двум последовательно выбранным текстовым наборам данных в одинаковом временном интервале.

Формирование векторных представлений осуществляется при помощи методов анализа естественного языка с учетом присутствия в данных информационного шума и специфики наборов коротких пользовательских текстов, а их сравнение при помощи выбранных метрик сходства. После сравнения и получения значения сходства осуществляется проверка вхождения полученного значения в доверительный диапазон. В результате работы метода идентификации нетиповых сценариев использования мобильного устройства эксперт получает перечень временных интервалов, а также массив облаков тегов, с нетиповой активностью.

3.1 Формирование наборов пользовательских текстов

В рамках диссертационной работы, был впервые собран набор текстовых данных, состоящий из 4 953 300 реальных сообщений, набираемых пользователями на мобильных устройствах. Сбор текстовых данных осуществлялся с согласия пользователей в фоновом режиме при помощи предустановленного на их мобильное устройство впервые разработанного приложения агента [77]. В сборе набора данных приняло участие около 200 пользователей. Экспериментальные исследования идентификации нетиповых сценариев использования мобильного устройства пользователем проводились на впервые собранном наборе реальных данных.

3.2 Определение временных диапазонов выборки

Для получения своевременной и корректной идентификации нетиповых сценариев использования мобильных устройств пользователей, требуется определить временные диапазоны выборки для анализируемых текстовых наборов. Для идентификации нетиповых сценариев, сравниваемые выборки, полученные по двум последовательно выбранным временным интервалам, должны быть достаточными и одновременно не избыточными по объему.

При избыточном объеме анализируемых текстовых данных наблюдается снижение качества идентификации нетиповых сценариев использования мобильного устройства в связи формированием более нормализованного во времени поведенческого вектора. Также увеличение объема текстовых данных зависит от длительности их сбора с клиентских устройств, что прямо сказывается на актуальности получаемых экспертом результатов анализа нетиповых сценариев использования мобильного устройства. Актуальность получаемых данных позволяет эксперту осуществлять своевременное реагирование и оперативно формировать управленческие решения.

При недостаточном объеме формируемой текстовой выборки, частота появления специфических определенному человеку слов снижена, и

определенные токены, формирующие эталонный поведенческий профиль пользователя, могут присутствовать лишь в малом количестве или отсутствовать вовсе, что в свою очередь негативно сказывается на качестве идентификации нетиповых сценариев использования устройств и возможности формирования и актуализации корректного и полного эталонного поведенческого профиля.

Для получения временных границ диапазона выборки требуется определить максимальный временной интервал, при котором получаемая администратором системы информация будет являться актуальной и позволять своевременно формировать управленческие решения в отношении целевых пользователей системы. Анализ текстовых данных в диапазоне более двух месяцев сводит актуальность получаемых результатов к минимуму и не позволяет эксперту оперативно принимать взвешенные управленческие решения [78] по отношению целевого пользователя системы, в связи с несвоевременным или сильно отсроченным получением результирующих данных [79]. Поведенческие характеристики пользователя, использующего мобильное устройство, меняются в зависимости от решаемой задачи. В случае невыполнения пользователем определенных целевых рабочих задач, администратор системы получит результат только по истечении выбранного временного периода анализа, что в свою очередь при выборе интервала более двух месяцев не позволит оперативно принимать управленческие решения и сводит ценность данного анализа к нулю.

Исходя из данных положений к формированию диапазона анализируемой выборки предлагается использовать два одинаковых временных интервала для последовательного извлечения двух наборов пользовательских текстов.

Количество данных, ежедневно собираемых с мобильных устройств, варьируется в зависимости от частоты использования мобильного устройства пользователем. В связи с этим невозможно получить две достаточные по объему выборки в случае, если интервал является коротким, так как одна

выборка может значительно отличаться по длине от последующей или вовсе быть пустой, что сделает дальнейший анализ невозможным, а получаемые результаты некорректными. Данная проблема решается путем увеличения интервала выборки. Результаты эксперимента представлены в таблице 3.1.

Таблица 3.1. Длины последовательно выбранных текстовых наборов

№	Диапазон одной выборки (дней)	ID пользователя	Объемы выборок (символов)	Сходство в длинах выборок в процентах
1	1	144	1550/3192	~ 48,5%
2			319/2435	~ 13,1%
3		155	2134/2212	~ 96,4%
4			1854/2631	~ 70,4%
5		186	154/2872	~ 5,3%
6			372/3765	~ 9,8%
7		175	76/3187	~ 2,3%
8			2467/3326	~ 74%
9	7	144	11003/13941	~ 78,92%
10			10236/12721	~ 80,4%
11		155	2969/4731	~ 62,75%
12			4371/5125	~85,28%
13		186	6786/6837	~99,2%
14			6434/6692	~96,2%
15		175	1192/3570	~33,38%
16			1832/2616	~ 70%
17	14	144	29865/49015	~ 60,9%
18			11023/24892	~ 44,2%
19		155	17528/21635	~ 81%
20			23562/24177	97,4%
21		186	14787/16043	~92,17%
22			12870/14719	~87,43%
23		175	12732/13786	~92,35%
24			12094/13874	~87,17%
25	28	144	37088/42144	~88%
26			34712/36072	~96,22%
27		155	35916/36981	~97,1%
28			33272/34036	~97,75%
29		186	44926/47031	~95,52%
30			41765/43912	95,11%
31		175	37815/39771	~95,08%
32			33036/35916	~91,98%

Исходя из полученных значений, можно сделать выводы о том, что анализ двух последовательно выбранных текстовых наборов с интервалом один день невозможен, так как наблюдается недостаточность данных

(выборки № 2, 5, 6, 7) в определенные выбираемые дни ввиду специфичности использования мобильных устройств пользователями (отсутствие взаимодействия или частичное взаимодействие пользователя с мобильным устройством), что не позволяет использовать анализ раз в два дня. При получении двух последовательно выбранных наборов текстовых данных пользователей за семь дней так же присутствуют недостаточно полные наборы (выборка № 15), однако частота их появления ниже, чем при формировании выборок за один день. Формирование двух последовательных выборок с интервалами в 14 или 28 дней показывают наиболее высокое сходство в длинах пользовательских текстов. Дальнейшие эксперименты, связанные с увеличением размера формируемых текстовых наборов для поиска отклонений в поведении пользователей, не проводились ввиду отсутствия актуальности у формируемых результатов анализа с большими временными интервалами.

3.3 Сравнение векторных представлений с использованием косинусного сходства

Для осуществления анализа пользовательских текстовых наборов, преобразованных в векторное представление, при помощи методов анализа естественного языка и дальнейшей идентификации нетиповых сценариев использования мобильных устройств, предлагается использовать косинусную меру как основной фактор идентификации. Применение косинусной меры сходства обусловлено тем, что данная метрика используется для анализа текстов, что подтверждается различными российскими и зарубежными научными исследованиями [80]. Косинусное сходство отображает косинус угла между двумя векторами, спроецированными в многомерном пространстве. Чем меньше величина угла, тем выше оценивается сходство. Так как даны два вектора признаков, A и B, косинусное сходство можно найти используя скалярное произведение и норму:

$$similarity(A,B) = \frac{A \cdot B}{\| A \| \times \| B \|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}} \quad (3.1)$$

В результате применения косинусной меры сходства, для сравнения пользовательских наборов текстовых данных, формируется числовой результат, определяющий значение сходства между двумя сравниваемыми векторами. Получим значения результатов сравнения векторных представлений пользовательских наборов текстовых данных, за различные выбранные временные интервалы (таблица 3.2).

Таблица 3.2. Результирующие значения косинусного сходства пользовательских текстов по методам анализа естественного языка

№	ID	Длина текста в выборке (символов)	Результирующие значения косинусного сходства по методам анализа естественного языка				
			BOW	TF-IDF	Word2Vec	Glove	BERT
1	144	30618; 22866	0,7632	0,8911	0,9941	0,9917	0,9491
2		8105; 1418	0,0533	0,0609	0,9262	0,8693	0,9471
3		20858; 20703	0,5279	0,6972	0,9692	0,9669	0,9798
4		63591; 2700	0,0290	0,0590	0,6937	0,6292	0,8952
Выборка двух интервалов по 14 дней							
5	155	6459; 4287	0,2921	0,3730	0,9548	0,9415	0,9687
6		19646; 53203	0,1291	0,1462	0,9317	0,9102	0,9571
7		8552; 18639	0,1329	0,1561	0,9220	0,9005	0,9354
8		21843; 15485	0,1182	0,1445	0,9025	0,8639	0,9277
Выборка двух интервалов по 28 дней							
9	175	13761; 28803;	0,5732	0,6066	0,9861	0,9774	0,9606
10		13336; 18304	0,5986	0,6622	0,9939	0,9884	0,9490
11	186	34891; 63422;	0,5334	0,6044	0,9479	0,9490	0,9601
12		57223; 38090	0,7901	0,8740	0,9971	0,9959	0,9422
Выборки за 7, 14, 28 дней с использованием мобильного устройства другим пользователем и заведомо известным присутствием аномалий в поведении							
13	145	7835; 9601	0,0328	0,0921	0,0976	0,1088	0,1124
14	169	22765; 18939	0,3672	0,3941	0,4102	0,4298	0,4002
15	198	53271; 52011	0,4112	0,4682	0,4778	0,4456	0,4573

Визуализация изменений в поведении пользователей представлена в виде графиков на рис. 3.1-3.4.

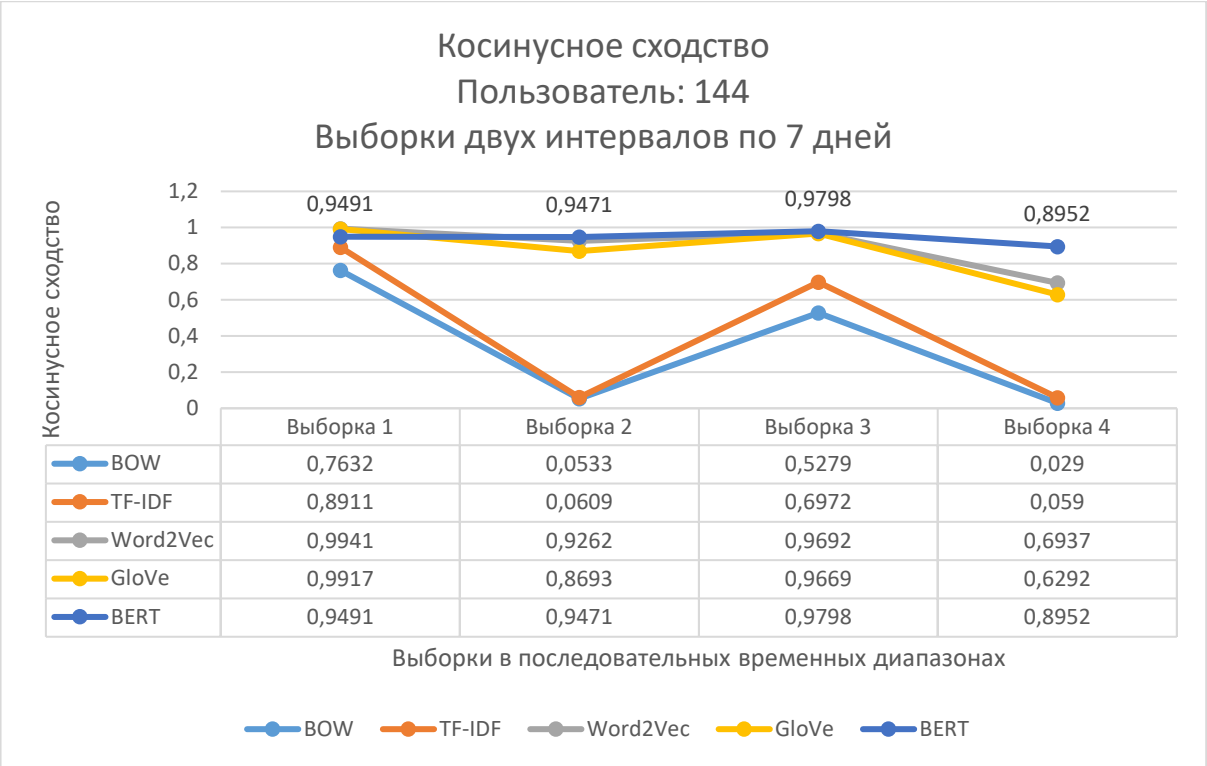


Рис. 3.1 – График изменения значений косинусного сходства пользователя №144

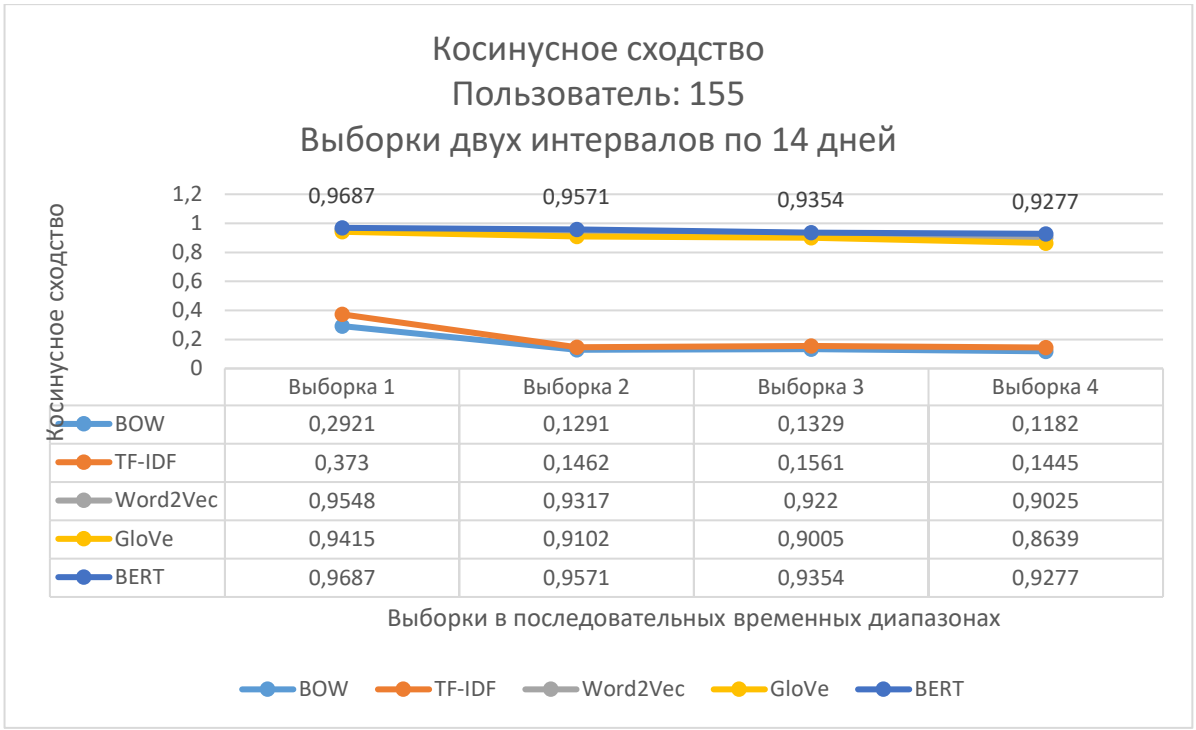


Рис. 3.2 – График изменения значений Косинусного сходства пользователя №155

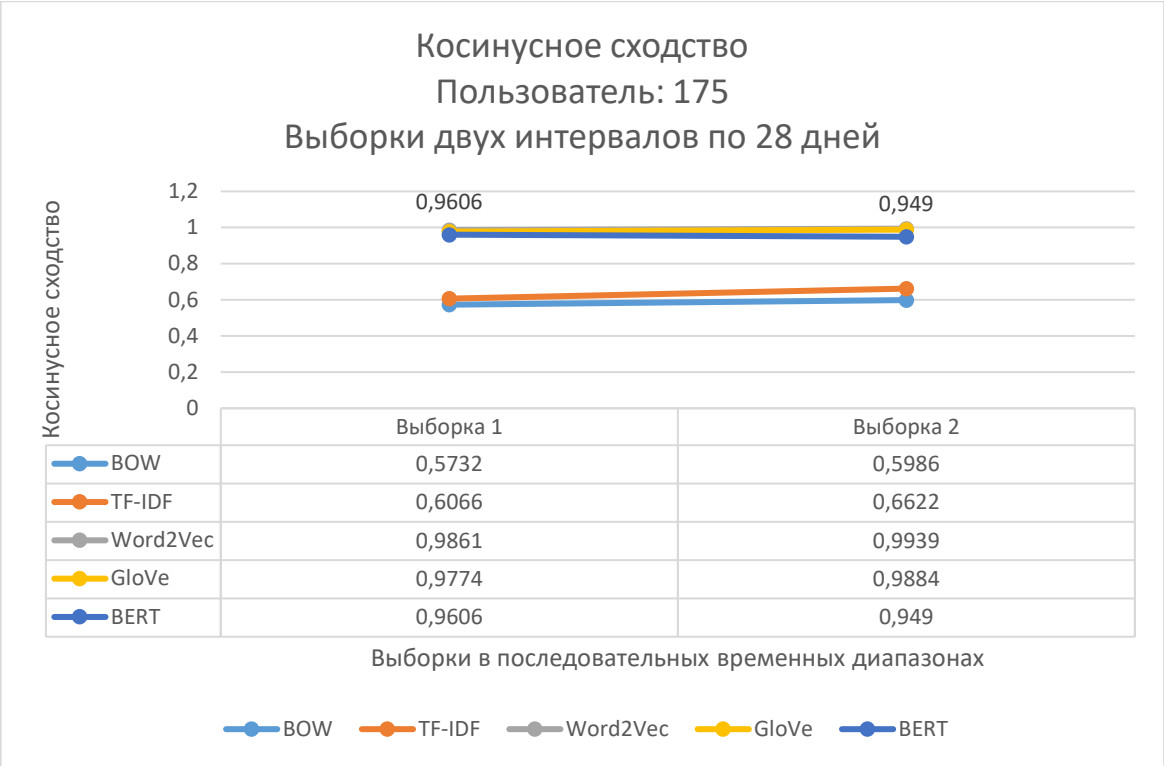


Рис 3.3 – График изменения значений Косинусного сходства пользователя №175

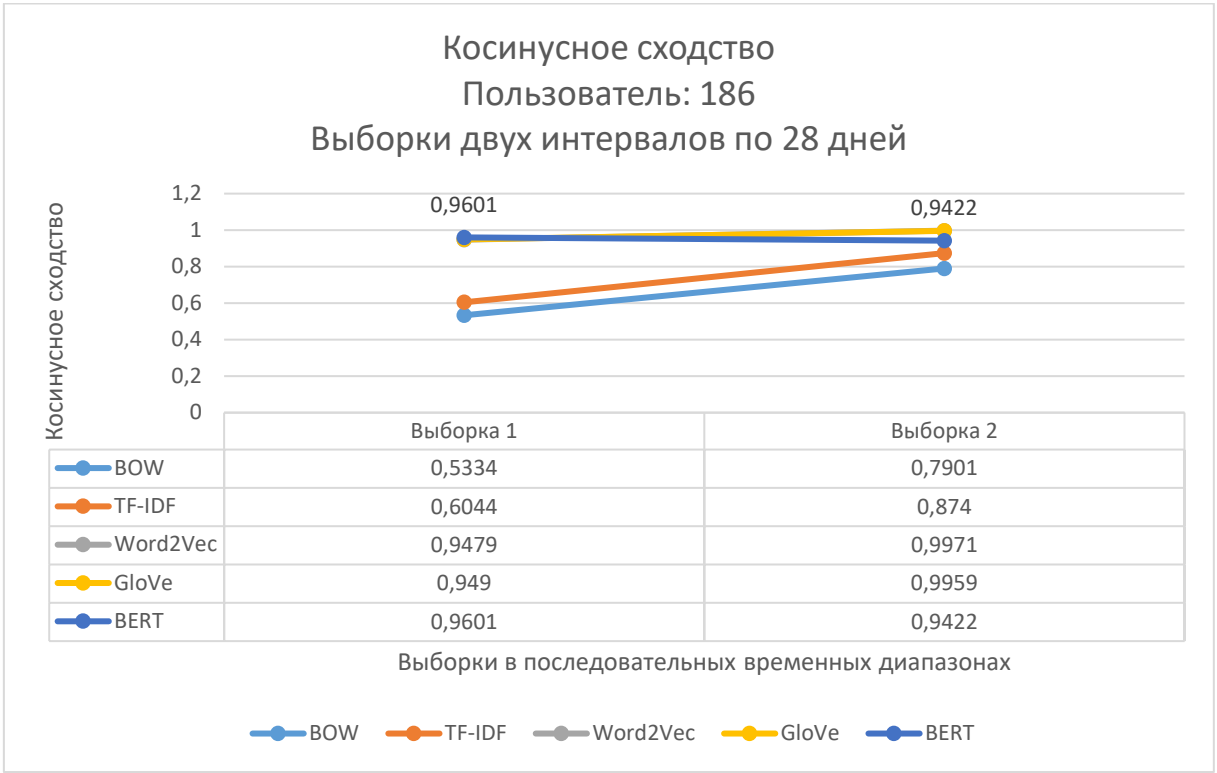


Рис 3.4 – График изменения значений Косинусного сходства пользователя №186

В результате проведенного эксперимента установлено, что при выборке двух интервалов по 7 дней, недостаточно данных в выборках №2 и №4, что негативно сказывается на результирующих значениях сравнения векторных представлений. В экспериментах №5, 8 при выборке двух интервалов по 14 дней данная проблема не зафиксирована, как и при интервалах по 28 дней (эксперименты №9, 10, 11, 12). В экспериментах на данных, в которых заранее был смоделирован нетиповой сценарий использования устройства (№13, №14, №15) на двух выборках по 7, 14 и 28 дней видны низкие результаты сравнения независимо от способа векторизации исходных текстовых наборов. Таким образом благодаря формированию векторных представлений и их анализа при помощи косинусного сходства, возможна идентификация нетиповых сценариев использования мобильного устройства.

3.4 Анализ векторных представлений при помощи Евклидова расстояния

Идентификация нетиповых сценариев использования мобильных устройств пользователями является многофакторной и требует использования различных подходов к анализу для получения высоко результативных значений. Наборы пользовательских текстов, выбранные за два последовательных временных интервала одной длины, далеко не всегда имеют одинаковую или схожую размерность, ввиду специфики использования мобильных устройств пользователями, наблюдаемой при сборе данных. Оценка частоты использования мобильного устройства пользователем является одним из факторов определения нетипового сценария. В случае оценки разности длин выборок пользовательских текстов возможно лишь базовое определение возможности последующего анализа в случае отсутствия сильного расхождения в длинах полученных текстовых наборов пользовательских данных. Однако анализ частоты использования устройства лишь по исходным длинам полученных выборок не является корректным, ввиду невозможности оценки поведенческих характеристик пользователя. Для

определения изменений частоты использования устройства пользователем, с учетом его поведенческих характеристик, в данной работе применяется Евклидова метрика [81].

Евклидово расстояние между двумя точками x, y в n-мерном пространстве определяется как:

$$r(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (3.2)$$

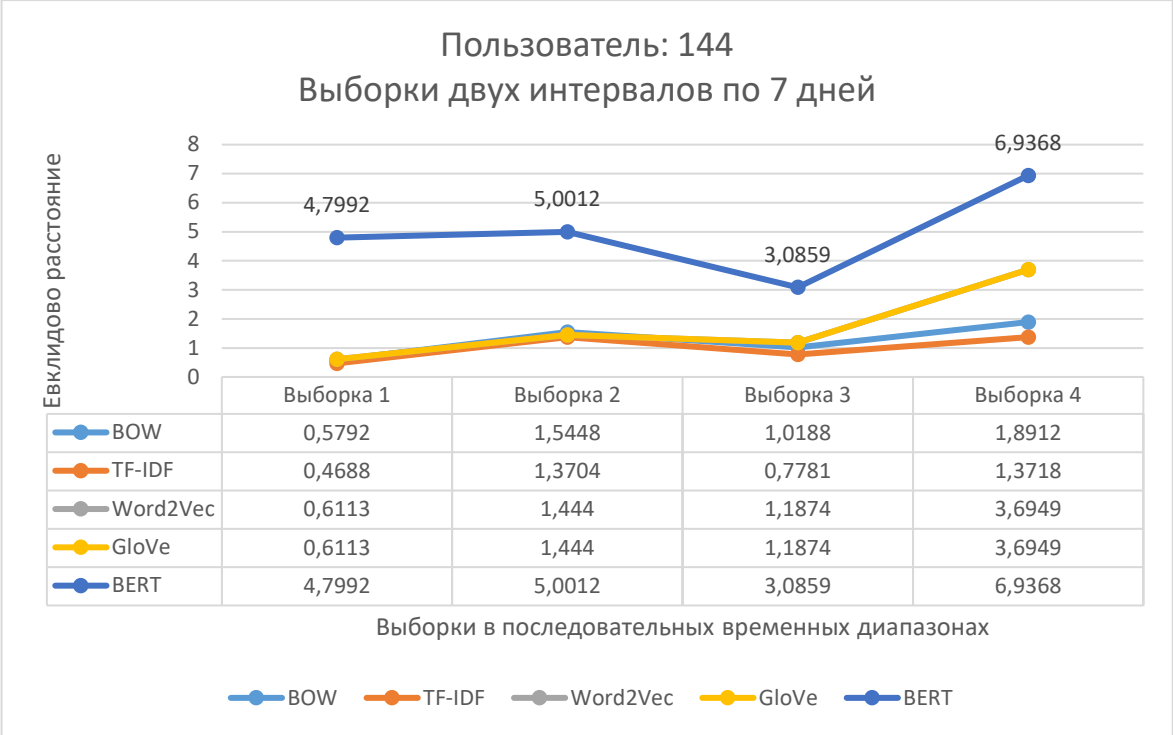
Сравнение значений Евклидова расстояния возможно только при использовании схожих длин диапазонов выборки пользовательских текстов, в ином случае полученные результирующие значения могут быть некорректными.

Для оценки изменений частоты использования устройства пользователями было проведено экспериментальное исследование, результаты которого представлены в таблице 3.3.

Таблица 3.3. Экспериментальные результирующие значения Евклидова расстояния для оценки частоты использования мобильных устройств

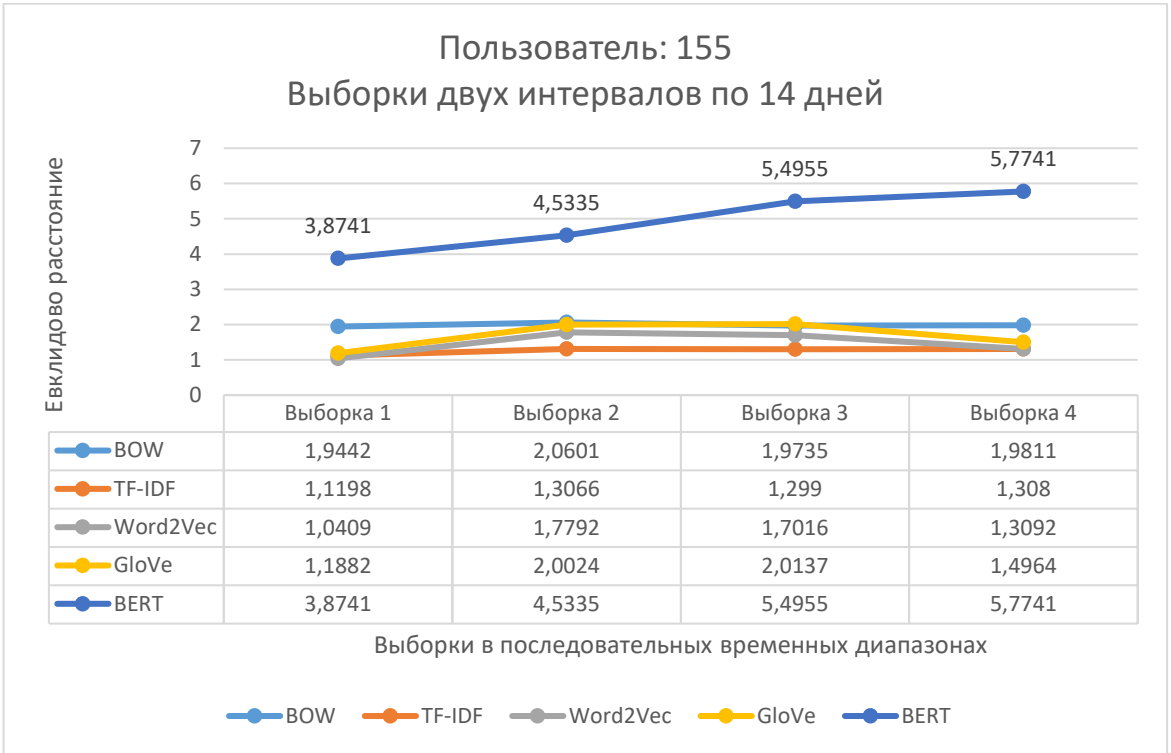
№	ID	Длина текста в выборке (символов)	Результирующие значения Евклидова расстояния по методам анализа естественного языка				
			BOW	TF-IDF	Word2Vec	Glove	BERT
Выборка двух интервалов по 7 дней							
1	144	30618; 22866	0,5792	0,4688	0,6113	0,7323	4,7992
2		8105; 1418	1,5448	1,3704	1,4440	1,8162	5,0012
3		20858; 20703	1,0188	0,7781	1.1874	1,3489	3,0859
4		63591; 2700	1,8912	1,3718	3,6949	5,2070	6,9368
Выборка двух интервалов по 14 дней							
5	155	6459; 4287	1,9442	1,1198	1,0409	1,1882	3,8741
6		19646; 53203	2,0601	1,3066	1,7792	2,0024	4,5335
7		8552; 18639	1,9735	1,2990	1,7016	2,0137	5,4955
8		21843; 15485	1,9811	1,3080	1,3092	1,4964	5,7741
Выборка двух интервалов по 28 дней							
9	175	13761; 28803;	1,4122	0,8869	0,6826	0,8974	4,2848
10		13336; 18304	1,3011	0,8218	0,5201	0,6694	4,8907
11	186	34891; 63422	1,0122	0,8894	1,5340	1,6598	4,3463
12		57223; 38090	0,8765	0,5019	0,4025	0,4969	5,1485
Выборки за 7, 14, 28 дней с использованием мобильного устройства другим пользователем и заведомо известным присутствием аномалий в поведении							
13	145	7835; 9601	1,7998	1,5442	1,6219	1,7182	7,9261
14	169	22765; 18939	3,9264	3,7478	3,6981	3,6502	9,9192
15	198	53271; 52011	5,1252	4,7688	4,7302	4,5501	8,9673

Графическое отображение изменений Евклидова расстояния так же представлено на рис. 3.5-3.8.



	Выборка 1	Выборка 2	Выборка 3	Выборка 4
BOW	0,5792	1,5448	1,0188	1,8912
TF-IDF	0,4688	1,3704	0,7781	1,3718
Word2Vec	0,6113	1,444	1,1874	3,6949
GloVe	0,6113	1,444	1,1874	3,6949
BERT	4,7992	5,0012	3,0859	6,9368

Рис. 3.5 – График изменения значений Евклидова расстояния пользователя №144



	Выборка 1	Выборка 2	Выборка 3	Выборка 4
BOW	1,9442	2,0601	1,9735	1,9811
TF-IDF	1,1198	1,3066	1,299	1,308
Word2Vec	1,0409	1,7792	1,7016	1,3092
GloVe	1,1882	2,0024	2,0137	1,4964
BERT	3,8741	4,5335	5,4955	5,7741

Рис. 3.6 – График изменения значений Евклидова расстояния пользователя №155

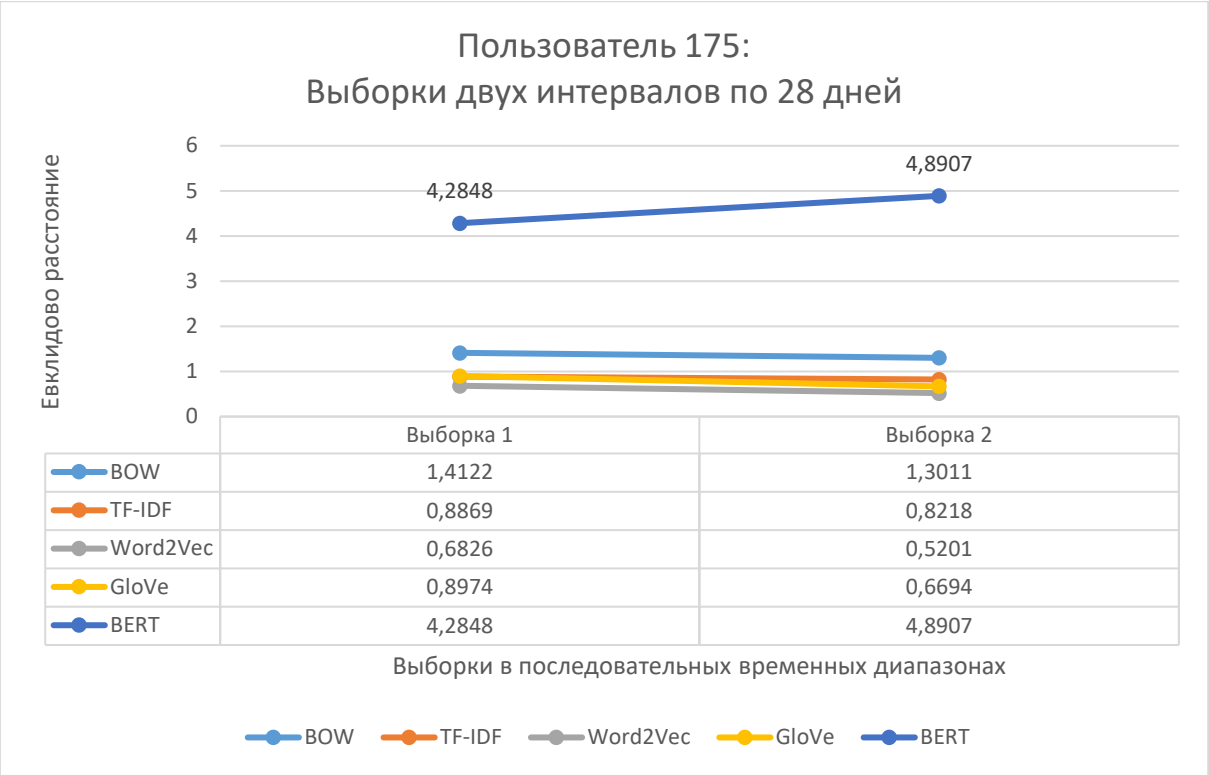


Рис. 3.7 – График изменения значений Евклидова расстояния пользователя №175

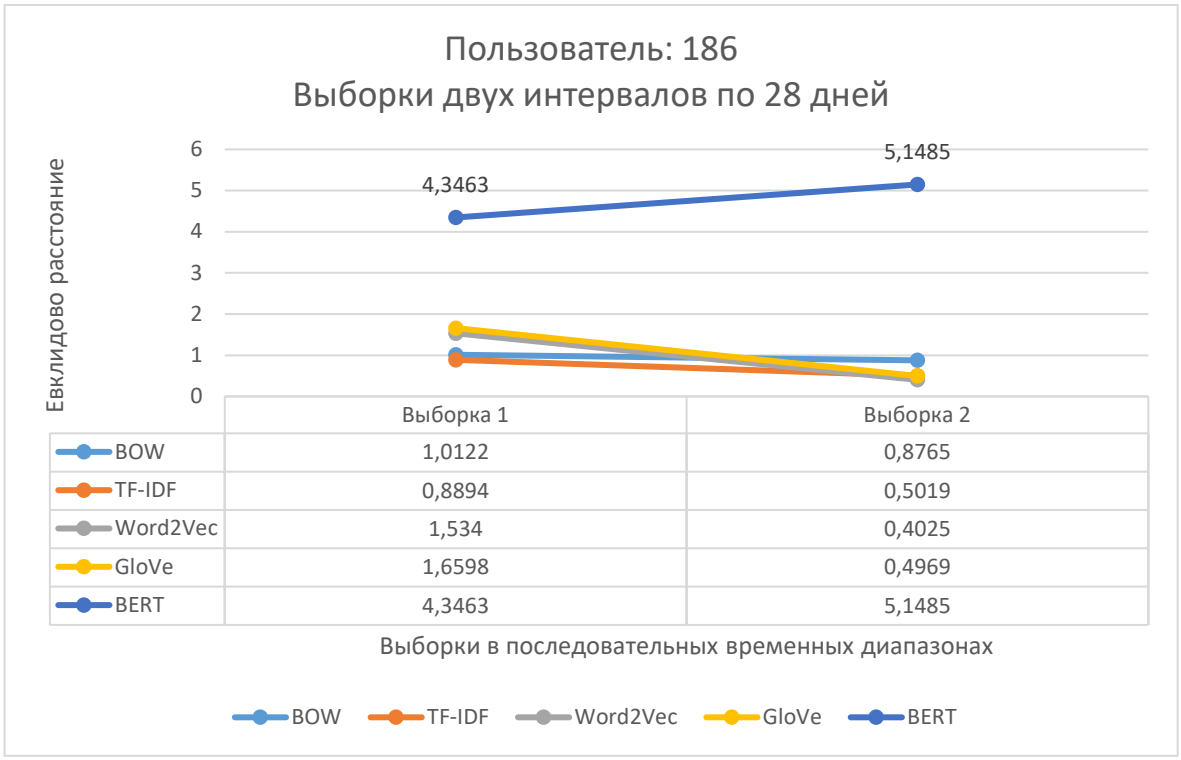


Рис. 3.8 – График изменения значений Евклидова расстояния пользователя №186

Евклидово расстояние становится не нулевым в случае, если абсолютное количество вхождений слов в анализируемых текстах отличается. Это может указывать на изменения в частоте использования мобильного устройства с целью переписки, а также на возможность получения ложного результата анализа в связи с недостаточным объемом текстовых данных, для формирования одного из векторов.

Значительное отличие результирующих значений Евклидова расстояния, наблюдаемое в экспериментах №2, 4, 6, 7, 11 возникает из-за высокой разницы в длинах полученных исходных текстовых наборов за определенный период и изменения частоты использования мобильного устройства пользователем. Эксперименты № 13, 14, 15 показывают постоянно высокие значения Евклидова расстояния при отсутствии значительных отличий в разнице длин формируемых текстовых выборок пользовательских текстов, что говорит о возможном наличии аномалий в поведении данных пользователей за выбранный временной период.

3.5 Метод идентификации нетиповых сценариев использования мобильного устройства

Для идентификации нетипового сценария использования мобильного устройства требуется установить порог отличий значений косинусной меры сходства и евклидова расстояния, примененных к двум векторам, полученным при последовательной выборке пользовательских текстов за определенный временной интервал.

Формулы определения диапазона верхней D_h и нижней D_l границ нормального поведения можно представить в следующем виде:

$$D_h = \frac{1}{n}\sum_{i=1}^{n} x_i + c \tag{3.3}$$

$$D_l = \frac{1}{n}\sum_{i=1}^{n} x_i - c \tag{3.4}$$

где:

n – размерность серии значений косинусной меры сходства полученная при анализе выборок пользовательских текстовых наборов;

x_i – значение косинусной меры сходства;

c – порог чувствительности, задаваемый администратором системы вручную.

Для идентификации нетиповых сценариев использования мобильного устройства пользователем, на анализируемом временном промежутке t, требуется определить принадлежит ли полученное значение сходства x, множеству значений нормального диапазона $D=\{x \in \mathbb{R}: D_l \leq x \leq D_h\}$. Сценарий считается нетиповым в случае, если полученное значение сходства выходит за пределы нормального диапазона отклонений $(x \notin D)$. Метод идентификации нетиповых сценариев использования мобильных устройств можно представить в виде схемы, изображенной на рис. 3.9.

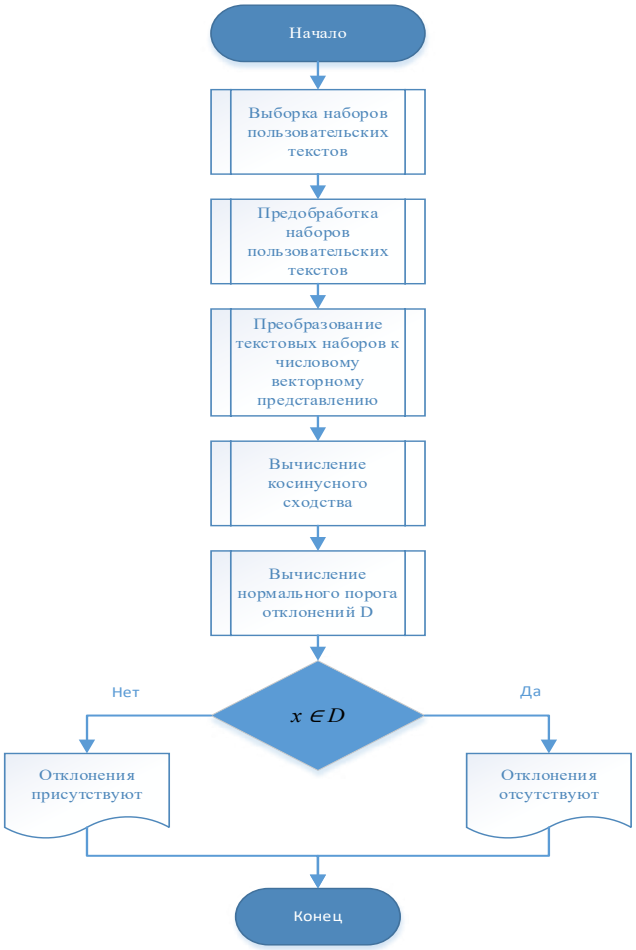


Рис. 3.9 – Метод идентификации нетиповых сценариев

Таким образом идентификация нетиповых сценариев использования мобильного устройства возможна при вычислении сходства между двумя последовательно выбранными текстовыми наборами в одинаковых временных интервалах, представленных в виде векторного представления при помощи методов анализа естественного языка, и ее сравнения с вычисляемым диапазоном значений нормального поведения.

Для идентификации нетиповых сценариев использования мобильных устройств пользователями, требуется сформировать текстовые наборы с учетом выбранных временных интервалов и осуществить предобработку пользовательских текстов и их анализ. Формирование анализируемых текстовых наборов осуществляется на основе выборки заранее полученной информации по целевому пользователю после чего производится их предобработка с выделением записей в интервале от 7 до 100 символов и их дальнейшая очистка от информационного шума. При помощи методов анализа естественного языка осуществляется формирование векторных представлений пользовательских текстов и их анализ при помощи косинусного сходства и евклидова расстояния. Для идентификации сценария использования мобильного устройства как типового или нетипового определяется значение величины уровня отклонений и проверка вхождения полученного значения сходства в нормальный диапазон [82]. В результате проведения данных операций определяется является ли сценарий использования мобильного устройства пользователем нетиповым [83]. В случае если сценарий является нетиповым администратор получает краткую агрегированную информацию о деятельности пользователя (аннотирование), полученную в анализируемом временном диапазоне.

Таким образом, сравнение пользовательских текстовых наборов на предмет сходства и последующая идентификация нетиповых сценариев использования мобильных устройств пользователями возможны при помощи применения косинусной меры сходства и евклидова расстояния и определения величины уровня отклонений с последующим аннотированием и

предоставлением эксперту краткой агрегированной результирующей информации [84].

3.6 Экспериментальное исследования метода идентификации нетиповых сценариев использования устройства

В результате проведенного экспериментального исследования разработанного метода идентификации нетиповых сценариев использования мобильного устройства впервые были получены значения сокращения объемов анализируемой вручную информации экспертами. Результаты представлены в таблице 3.4.

Таблица 3.4. Анализ метода идентификации нетиповых сценариев и сокращения объема анализируемой вручную информации

№	Общ. объем (симв.)	Не типовой сценарий	Объем облака тегов (слов)	Детальный анализ	Сокращение объема данных по выборкам %	Сокращение объема данных (общее) %
1	72985	+	100	-	85	92
2	34067	+		-		
3	127922	+		-		
4	48895	+		+		
5	34096	+		-		
6	55087	-		-	100	
7	14322	-		-		
8	39061	-		-		
9	64055	-		-		
10	99186	-		-		

Для эксперимента №4 было сформировано облако ключевых тематик (рис. 3.10), так как метод определил сценарий использования мобильного устройства как нетиповой.

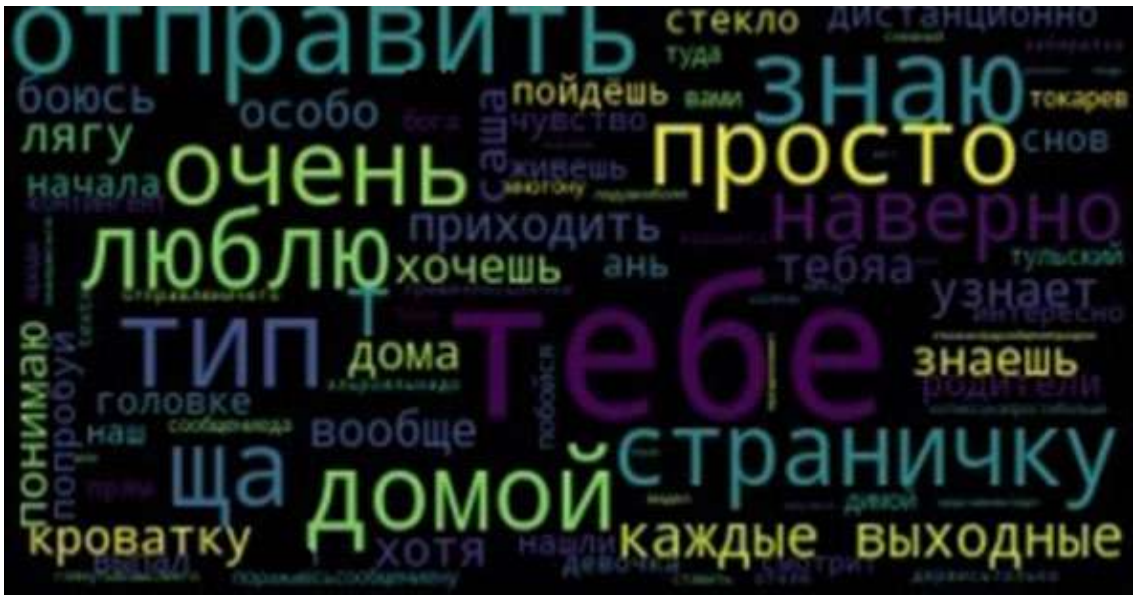


Рис. 3.10 – Облако тегов (эксперимент №4)

Для своевременного ознакомления с типом деятельности пользователя (эксперимент №8), использующего мобильное устройство в типовом сценарии, также было сформировано облако тегов при запросе эксперта (рис. 3.11).

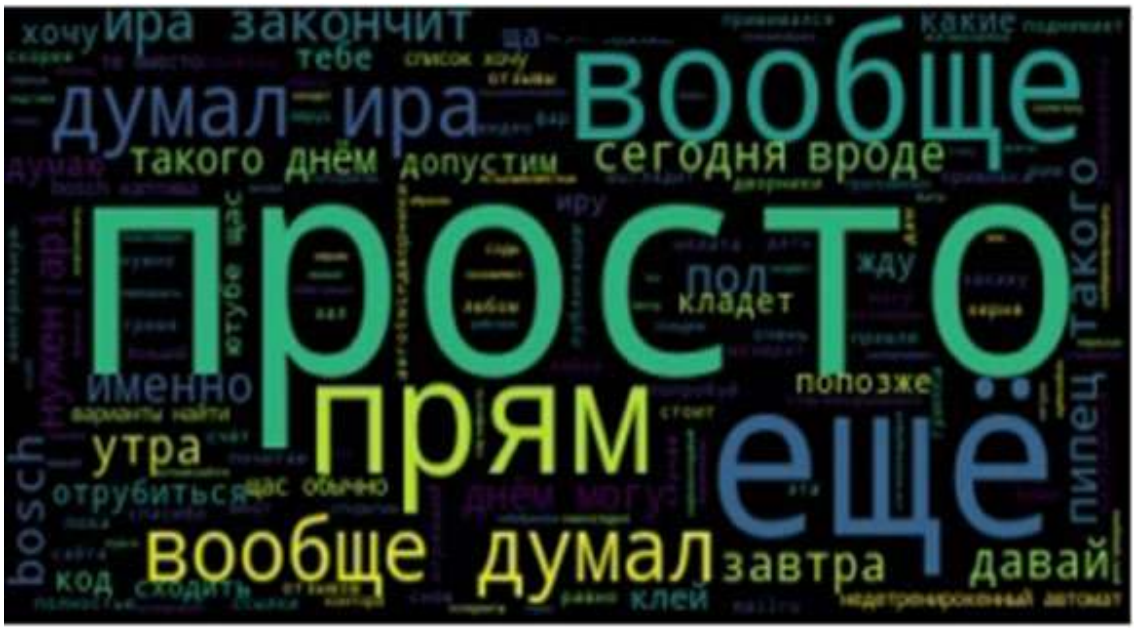


Рис. 3.11 – Облако тегов, сформированное по пользовательскому тексту в эксперименте №8

По результатам экспериментального исследования, разработанного метода идентификации нетиповых сценариев использования мобильных

устройств пользователями, удалось сократить объем анализируемой экспертами вручную информации до 92%, а также сформировать облака тегов по пользовательским наборам текстовых данных.

3.7 Выводы

В результате разработки метода идентификации нетиповых сценариев использования мобильных устройств пользователей, по их наборам коротких текстовых данных, были сформулированы следующие выводы:

1. Впервые был собран набор текстовых данных, состоящий из 4 953 300 реальных сообщений;
2. Установлено, что для идентификации нетиповых сценариев использования мобильных устройств пользователями требуется постоянная актуализация поведенческих данных (эталонного поведенческого профиля);
3. Было получено, что для анализа требуется формировать векторные представления на основе последовательно выбранных наборов текстов, имеющих одинаковый временной интервал выборки;
4. Было проведено экспериментальное исследование, по результатам которого установлено, что временными диапазонами выборки могут являться диапазоны 7, 14, 28 дней. Диапазоны до 7 дней не могут быть проанализированы ввиду присутствия в них большого количества информационного шума относительно общего объема выборки, интервалы в 14 или 28 дней показывают наиболее высокое сходство в длинах пользовательских текстов. Выявлено, что формирование выборок более 28 дней не рационально ввиду несвоевременности получения экспертами информации;
5. Было выбрано 2 метрики для идентификации нетиповых сценариев использования мобильных устройств, это косинусная мера сходства и евклидово расстояние;
6. Установлено, что анализ частоты использования устройства лишь по исходным длинам полученных выборок не является корректным, ввиду

невозможности оценки поведенческих характеристик пользователя. Для определения изменений частоты использования устройства пользователем, с учетом его поведенческих характеристик, в данной работе применяется Евклидова метрика;

7. Был создан метод идентификации нетиповых сценариев использования мобильных устройств пользователей;

8. В результате разработки метода идентификации нетиповых сценариев удалось сократить объем анализируемой экспертами вручную информации до 92%. Сформированы облака тегов с целевыми тематиками общения, для предоставления экспертам;

9. Разработанный метод идентификации нетиповых сценариев использования мобильных устройств позволяет сократить объем обрабатываемой экспертами информации за счет акцентирования их внимания на временных промежутках, в которых была обнаружена нетиповая активность, сформировав при этом облако ключевых тематик общения в пределах данных интервалов.

4 РЕАЛИЗАЦИЯ ПРОГРАММНОГО КОМПЛЕКСА СБОРА И АНАЛИЗА ТЕКСТОВЫХ ДАННЫХ

Данный раздел посвящен разработке и реализации экспериментального образца программного комплекса сбора текстовых данных пользователей и идентификации нетиповых сценариев использования мобильных устройств [86-88].

Раздел включает в себя описание и решение следующих задач:

- Описание сценариев использования экспериментального образца программного комплекса;
- Программная реализация;
- Экспериментальная проверка;
- Апробация программного комплекса.

4.1 Описание сценариев использования

Разрабатываемая система идентификации нетиповых сценариев использования мобильных устройств имеет следующие сценарии использования:

- Сбор наборов текстовых данных;
- Формирование векторных представлений, содержащих информацию о взаимодействии пользователя с мобильным устройством;
- Идентификация нетиповых сценариев использования мобильных устройств.

Далее будет представлено описание вариантов взаимодействия с модулями системы, а именно:

- Установка и первичная настройка мобильного приложения агента;
- Использование мобильного устройства с предустановленным приложением агентом и сбор текстовых данных;

- Использование Web интерфейса для управления устройствами пользователей и сбора данных;
- Использование Web интерфейса для идентификации нетиповых сценариев использования мобильных устройств и построения облака тегов.

4.1.1 Установка и первичная настройка мобильного приложения агента

Для корректной работы мобильного приложения – агента на пользовательском устройстве требуется его предварительная настройка, включающая в себя выдачу различных разрешений, активацию системных сервисов, настройки оптимизаторов операционной системы и оболочки производителя. Для снижения количества ошибок, допускаемых при настройке мобильного приложения агента, администратором системы, мобильный агент имеет следующие функциональные особенности:

- Последовательная настройка. Интерфейс системы построен таким образом, что переход к следующему пункту настройки доступен только после успешной активации предыдущего;
- Цветовая индикация. В случае корректного завершения настроек по выбранному пункту, напротив него отобразиться зеленый индикатор, информирующий о успешном завершении действия. При возникновении ошибки цвет индикатора будет изменен на красный;
- Информационные сообщения об ошибках. В случае возникновения ошибки при настройке выбранного пункта, представленного на интерфейсе, система автоматически предложит наиболее подходящий вариант ее решения, информируя пользователя о дальнейших действиях, исходя из возможных блокирующих факторов (версии операционной системы и графической оболочки);
- Запрет запуска мобильного приложения агента. Если пользователь пропустил настройку определенных пунктов или их конфигурирование завершилось с ошибкой, либо разрешения были выданы не полностью, при нажатии на кнопку «Запуск в фоне» система осуществит информирование в

виде всплывающего окна, содержащего инструкции для выполнения дальнейших действий;

- Индикатор работы сервиса. После успешного завершения настройки в верхней части экрана мобильного устройства будет отображено уведомление, информирующее о том, что приложение агент запущено и работает в фоне;
- Блокировка входа в приложение. После запуска приложения в фоне, в дальнейшем, открыть интерфейс настройки невозможно без специального сервисного кода, задаваемого администратором. Данный сервисной код требуется ввести в номеронабирателе и осуществить вызов (звонок), после чего мобильное приложение агент приостановит работу и интерфейс конфигурирования будет открыт и доступен для проведения дальнейших операций.

Интерфейс настроек мобильного приложения агента представлен на рис. 4.1.

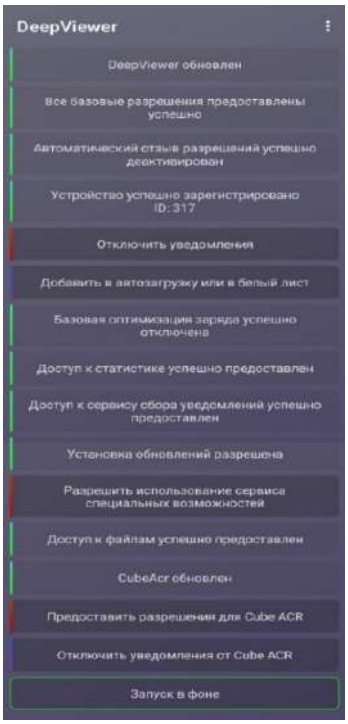


Рис. 4.1 – Интерфейс настроек мобильного приложения агента

Ввиду наличия определенных особенностей у различных системных оболочек, помимо зеленой и красной индикации, находящейся напротив

определенного действия, имеется синяя, сообщающая о невозможности проверки статуса выполнения действия или корректности завершении настройки по указанному пункту.

Первичным действием после загрузки пакетного файла формата «apk» является его запуск для последующей установки (рис. 4.2).

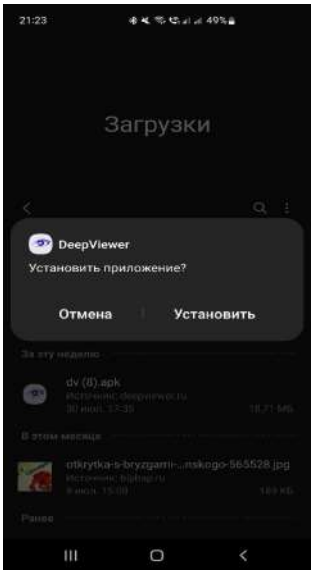


Рисунок 4.2 – Первичная установка мобильного приложения агента

После успешной установки мобильного приложения агента открывается интерфейс конфигурации.

Рисунок 4.3 – Интерфейс конфигурации мобильного приложения агента

Для дальнейшей настройки мобильного приложения агента требуется осуществить следующие действия:

- предоставить базовые разрешения;
- деактивировать автоматический отзыв разрешений;
- зарегистрировать мобильное устройство в системе;
- отключить уведомления;
- добавить в автозагрузку и белый лист;
- отключить базовую оптимизацию заряда;
- предоставить доступ к статистике;
- предоставить доступ к сервису сбора уведомлений;
- разрешить обновления;
- разрешить использование сервиса специальных возможностей;
- предоставить доступ к файлам.

Последовательное выполнение пунктов конфигуратора администратором системы представлено на рис. 4.4 – 4.6.

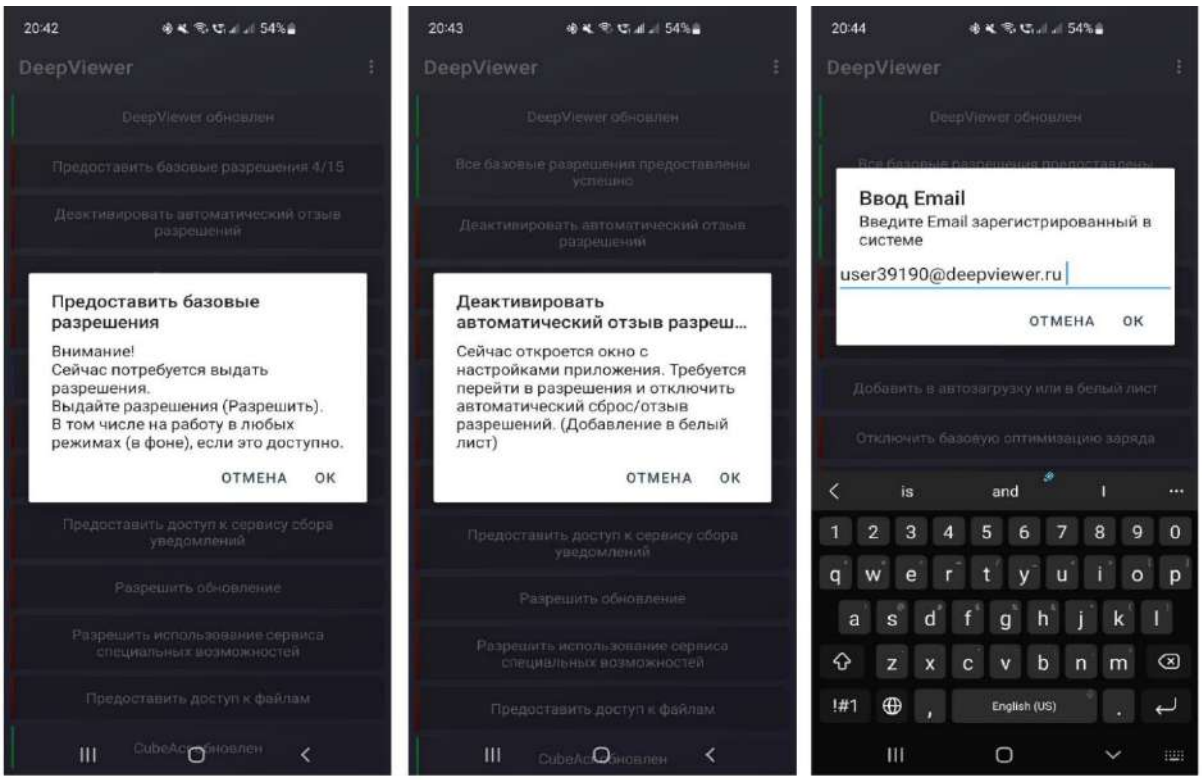


Рис. 4.4 – Конфигурирование мобильного приложения агента (выдача разрешений, запрет сброса, регистрация)

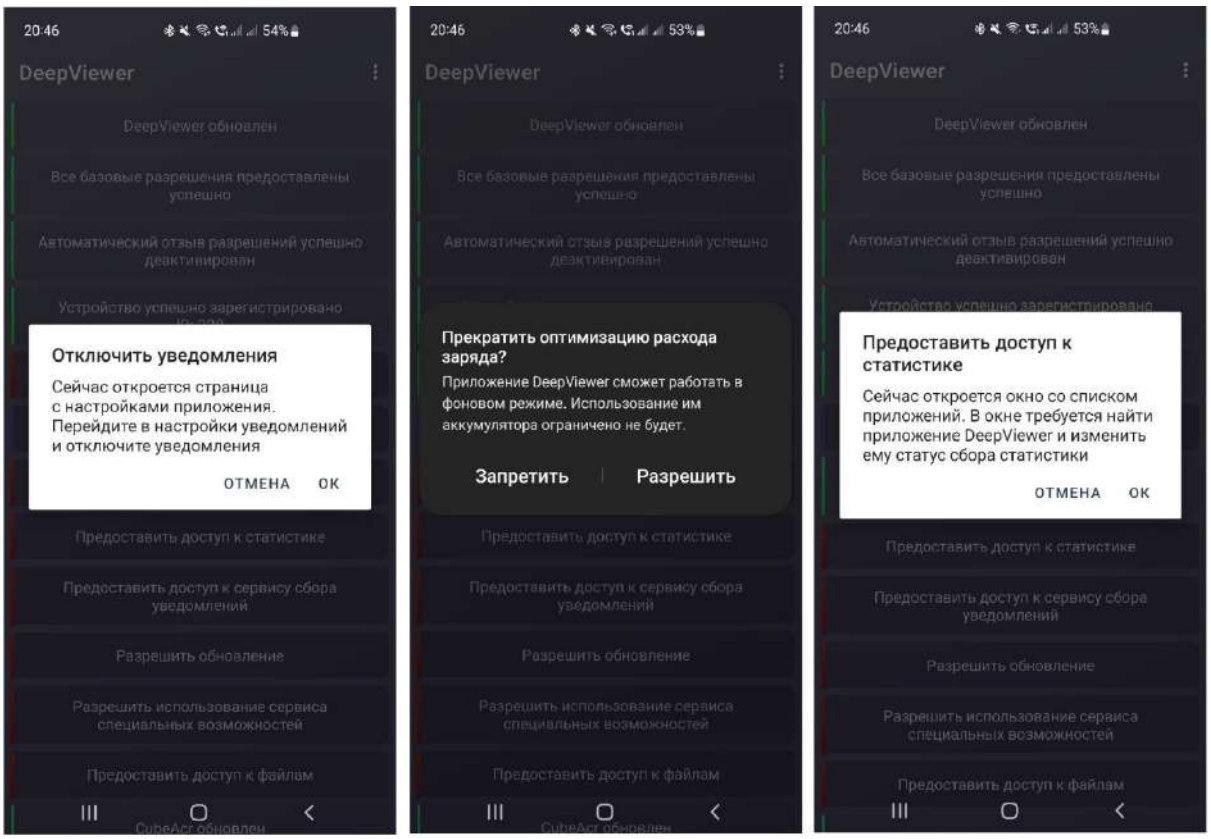


Рис. 4.5 – Конфигурирование мобильного приложения агента (деактивация уведомлений, оптимизаторов, активация доступа к статистике)

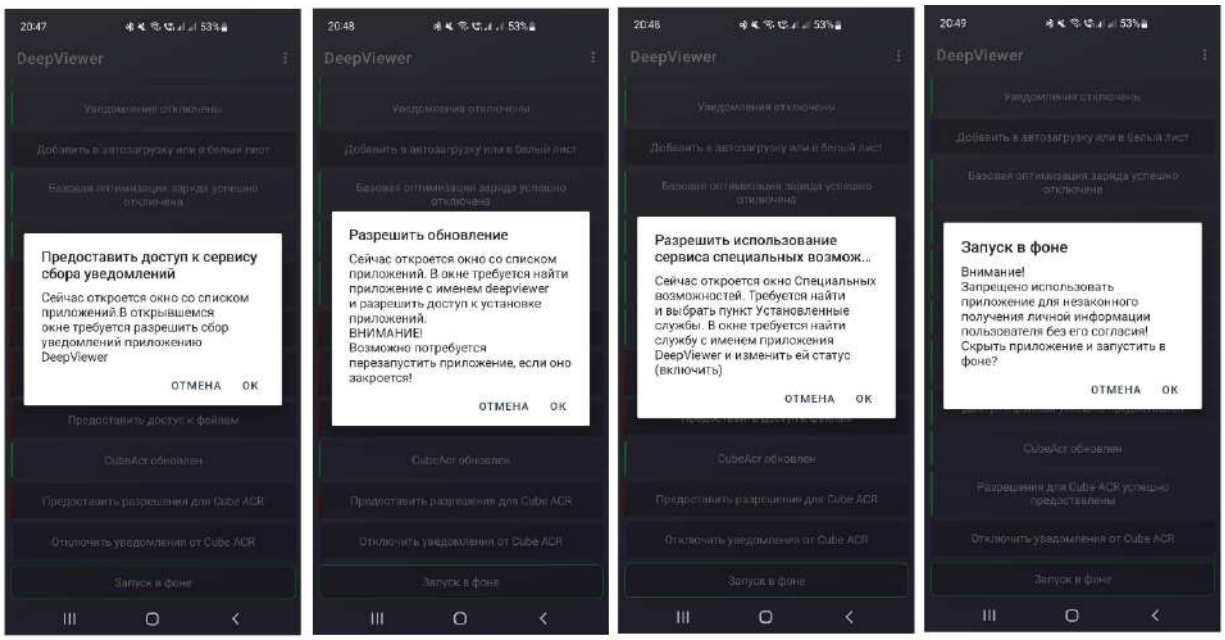


Рис. 4.6 – Конфигурирование мобильного приложения агента (Активация сервиса сбора уведомлений, сбора пользовательского ввода, обновлений и запуск в фоне)

После успешной настройки и запуска мобильного приложения агента, сервис продолжает работу в фоновом режиме, а интерфейс конфигурирования становится недоступным для повторного открытия без сервисного кода.

4.1.2 Использование мобильного устройства с установленным агентом и сбор поведенческих данных

Использование мобильного устройства с предустановленным и настроенным мобильным приложением агентом для сбора наборов текстовых данных, не отличается от обычного стандартного использования устройства.

Единственным отличием мобильного устройства с предустановленным мобильным приложением – агентом является индикация о работе приложения в фоне, в разделе уведомлений устройства, однако данная индикация может быть деактивирована при первичной настройке. Пример работы индикации мобильного приложения агента представлен на рис. 4.7.

Рис. 4.7 – Индикация статуса активности мобильного приложения агента

Высокий уровень энергоэффективности и производительности приложения агента позволяет пользователям использовать мобильное устройство в полной мере в стандартных сценариях использования. Подробная информация о энергоэффективности и производительности приложения, полученная на основе экспериментальных исследований, подробно представлена далее в данной главе.

4.1.3 Использование Web интерфейса для управления устройствами пользователей и сбора данных

Для идентификации нетиповых сценариев использования мобильных устройств пользователями и сбора доказательной базы о их деятельности, требуется настройка, выполняемых на устройстве пользователя команд, при помощи панели администрирования.

Управление выбранным устройством становится доступным после установки и первичной настройки мобильного приложения агента. Web интерфейс управления устройствами представлен на рис. 4.8.

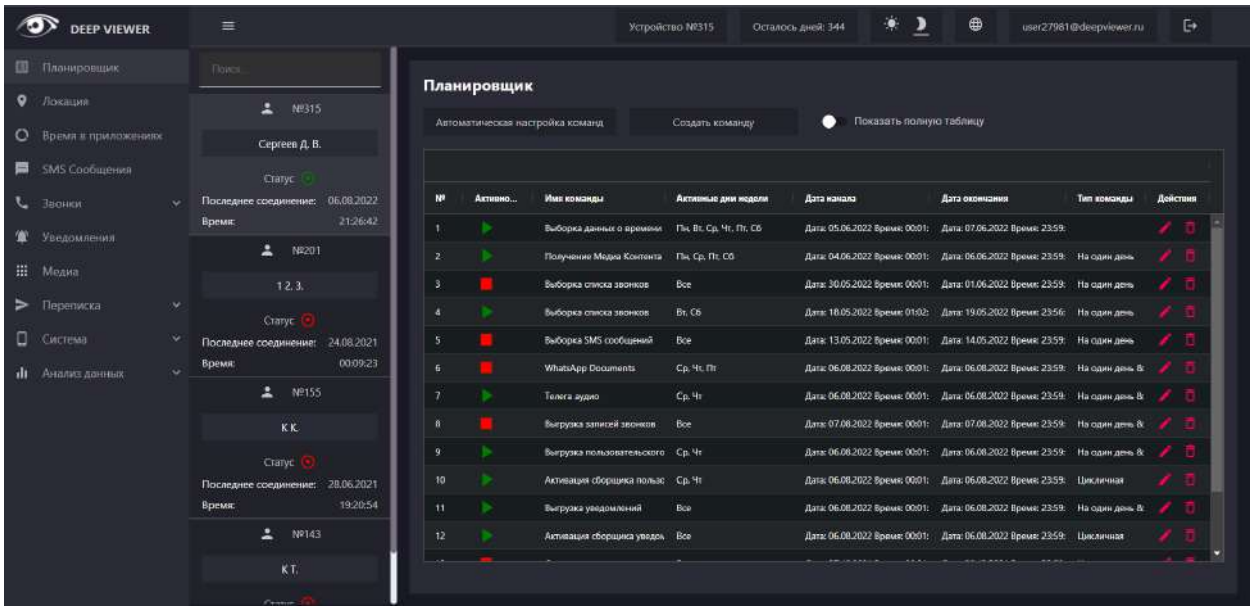


Рис. 4.8 – Web интерфейс панели администрирования

В левой части интерфейса располагаются пункты меню, отображающие функциональные возможности системы. Правее меню находится блок со списком пользователей, зарегистрированных в системе, чьи устройства доступны для сбора данных и последующего анализа. В теле вкладки

«Планировщик» осуществляется планирование задач сбора пользовательских данных для каждого устройства. Для формирования новой задачи сбора, администратор системы создает новую команду и конфигурирует ее параметры.

Таким образом, планирование выполняемых на устройстве задач позволяет сократить потребление заряда мобильным устройством на выполнение функций.

4.1.4 Использование Web интерфейса для анализа отклонений в поведении пользователя

Анализ отклонений в поведении пользователей с дальнейшей идентификацией нетиповых сценариев использования мобильного устройства осуществляется на вкладке «Анализ данных» при взаимодействии с пунктами «Облако тегов» и «Сценарии использования» (рис. 4.9).

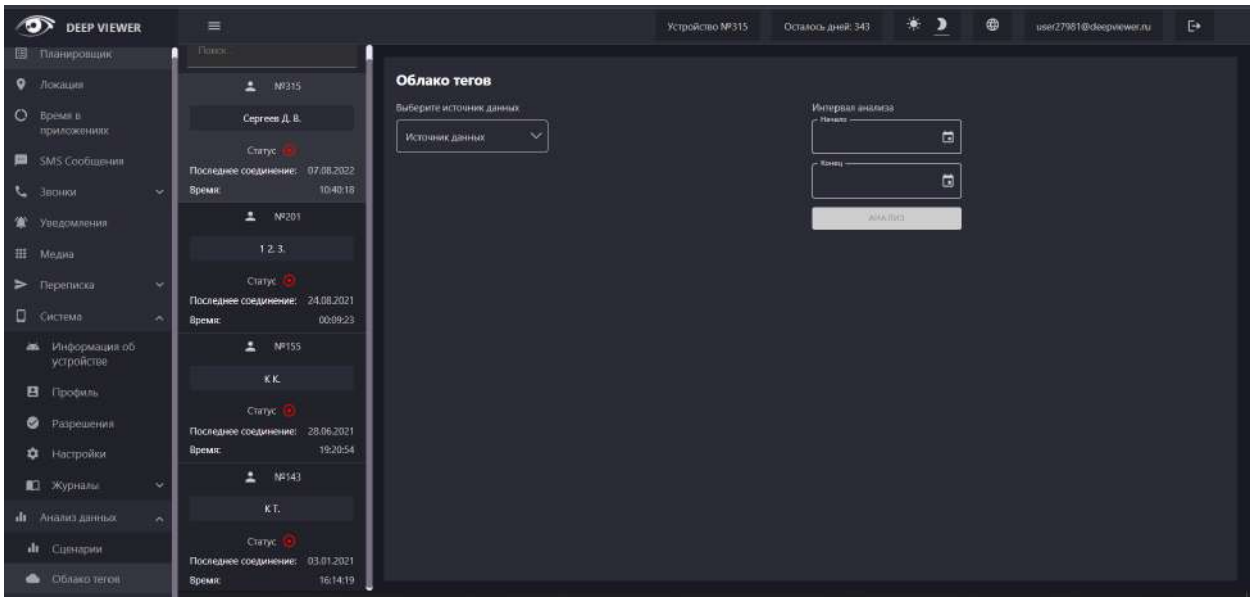


Рис. 4.9 – Содержимое пункта меню «Анализ данных»

Для своевременного ознакомления с деятельностью выбранного пользователя, администратор системы (эксперт) имеет возможность сформировать облако тегов за выбранный временной интервал. Результат формирования облака тегов представлен на рис. 4.10.

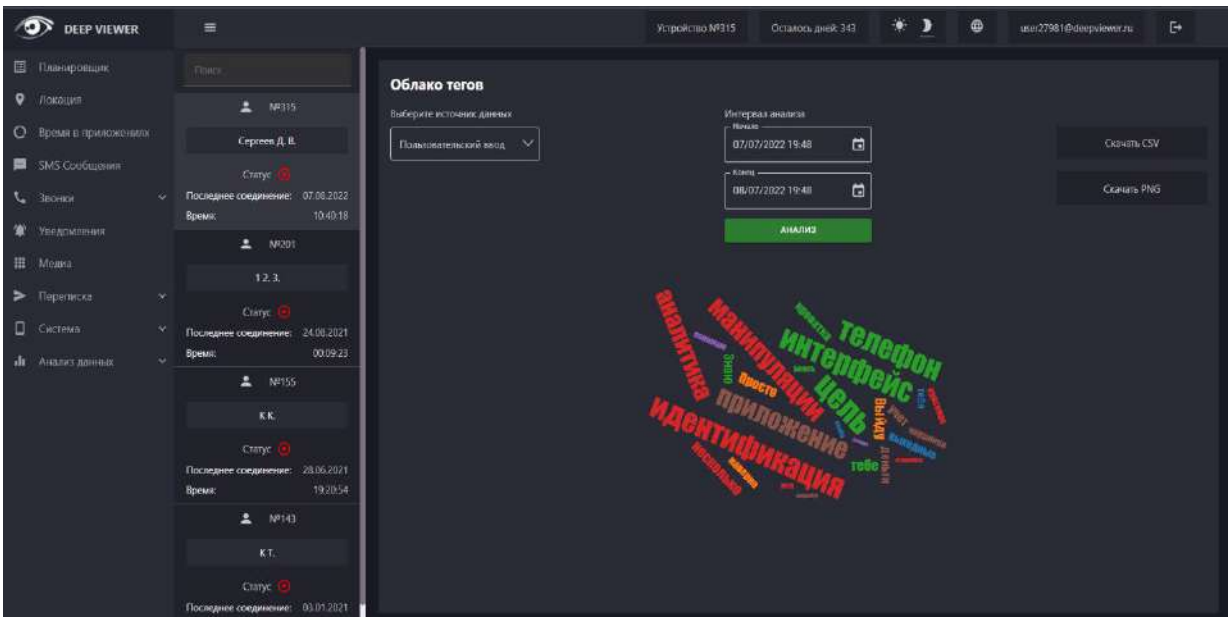


Рисунок 4.10 – Облако тегов с месячным интервалом для пользователя №315

Облако тегов формируется на основе наиболее часто употребляемых пользователем слов и отображается в графическом виде. Полученный результат может быть выгружен в формате документа csv и графическом png.

Идентификация нетиповых сценариев использования мобильного устройства пользователем осуществляется на вкладке «Сценарии» раздела анализ данных (рис. 4.11).

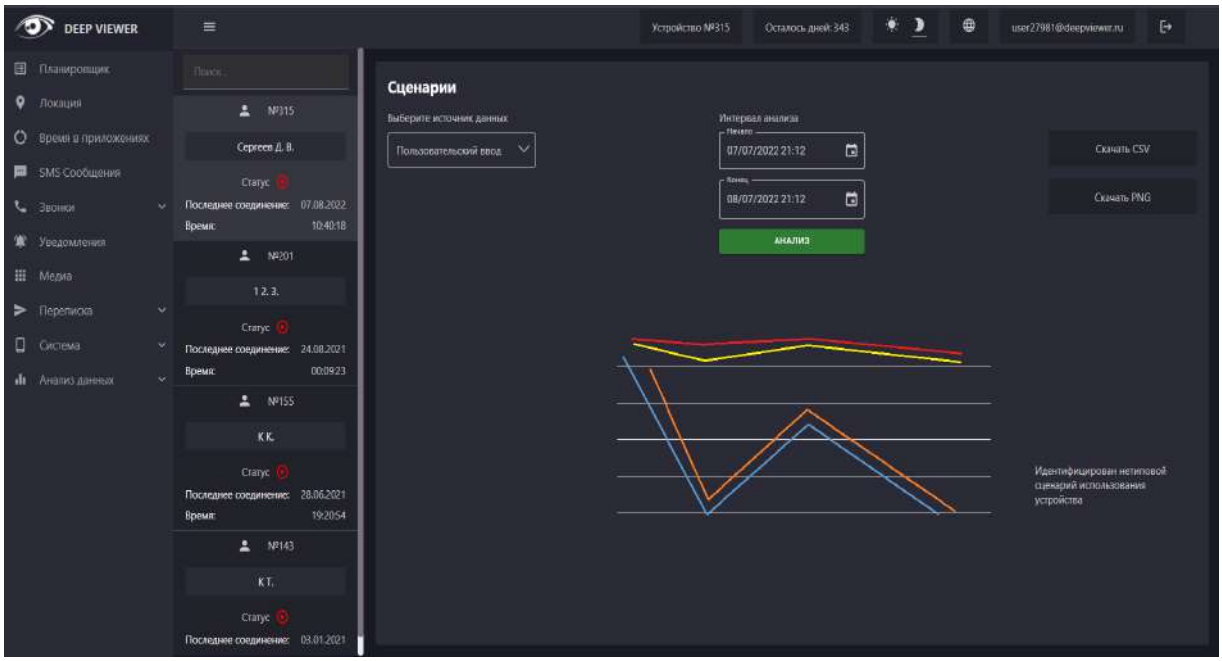


Рис. 4.11 – Идентификация нетиповых сценариев использования мобильного устройства

Для идентификации типа сценария использования мобильного устройства администратору системы (эксперту) требуется установить один интервал анализа. Остальные интервалы, требующиеся для анализа, система сформирует автоматически. После проведения анализа в центре экрана будет отображен график изменений в поведении пользователя и сформировано результирующее сообщение о сценарии использования устройства.

4.2 Программная реализация

4.2.1 Проектирование архитектуры программного комплекса

Поддержание высокого уровня отказоустойчивости и стабильности работы модулей, а также возможность масштабирования системы является одной из первоприоритетных задач. Для ее достижения требуется осуществить деление системы на программные модули, осуществляющие различные логические задачи такие как [89-92]:

- *Мобильное приложение - агент.* Модуль - агент сбора данных;
- *API модуль клиент-серверного взаимодействия.* Осуществляет взаимодействие с клиентскими модулями – агентами;
- *Интеллектуальный сервер поведенческого анализа.* Осуществляет интеллектуальную обработку уникальных наборов текстовых данных пользователей, идентифицирует нетиповые сценарии использования мобильного устройства и формирует облако тегов по ключевым темам собранных текстов пользователя;
- *Web модуль администрирования системы Dashboard (Web панель)*. Предоставляет администратору системы доступ к управлению системой, анализу и просмотру данных;
- *Web модуль администрирования системы Dashboard (Web API).* Осуществляет взаимодействие с Web модулем администрирования системы.

Графическое представление спроектированной отказоустойчивой и масштабируемой архитектуры экспериментального образца программного комплекса представлено на рис. 4.12.

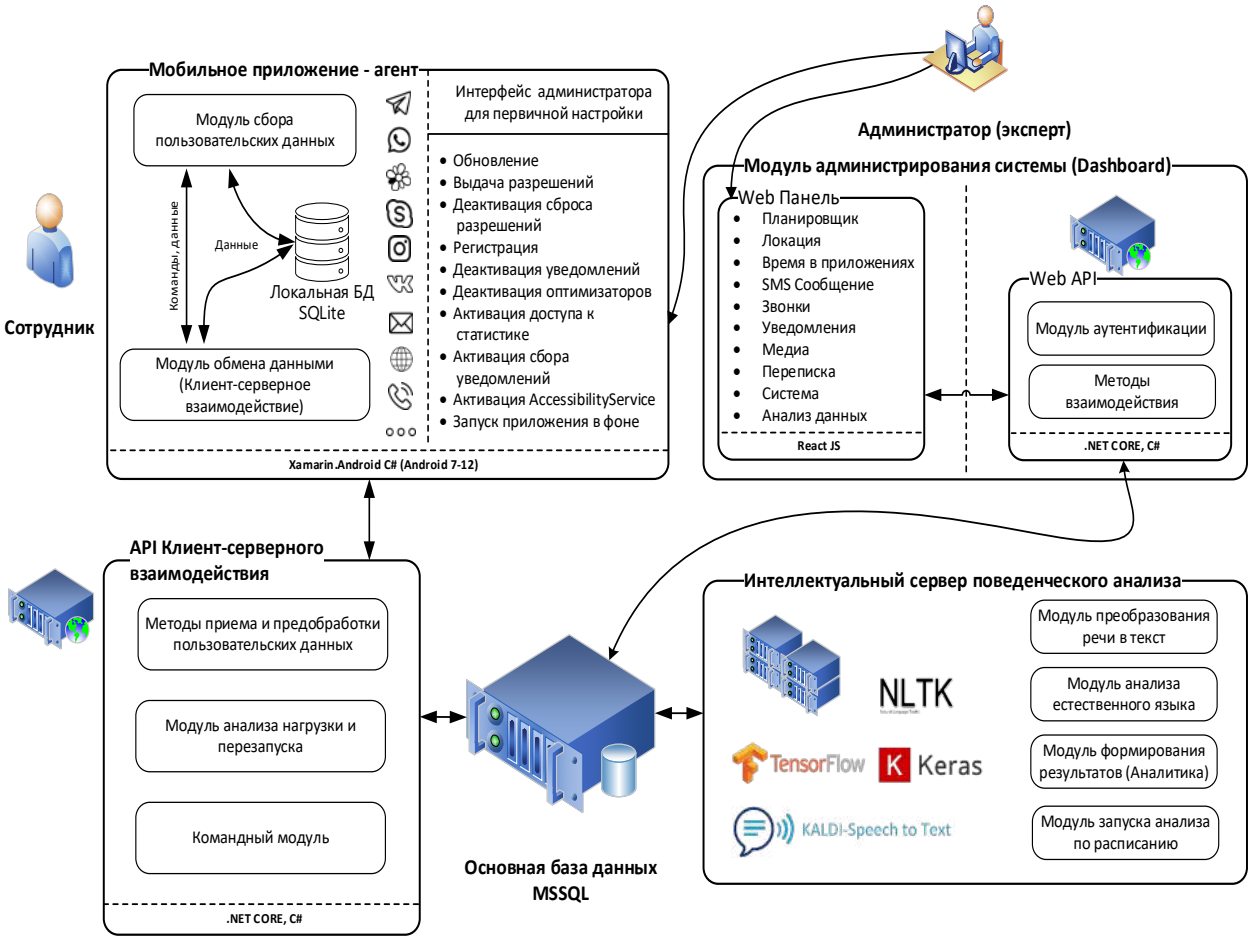


Рис. 4.12 – Архитектура экспериментального образца программного обеспечения

В верхнем левом углу графического представления архитектуры, представлены клиентские устройства с предустановленными мобильными агентами сбора (программный модуль - «Мобильное приложение агент»). Мобильные устройства подключаются к модулю «API клиент-серверного взаимодействия». Данный модуль осуществляет обмен информацией с мобильными устройствами пользователей системы. После получения модулем клиент-серверного взаимодействия информации от мобильных клиентов данные обрабатываются и записываются в базу данных.

Взаимодействие администратора с системой осуществляется через «Web модуль администрирования системы». Данный интерфейс аналогичным образом интегрирован с «WebAPI», который получает запросы от Web интерфейса системы и осуществляет их обработку с дальнейшей выборкой или получением информации из базы данных о выбранном пользователе.

Модуль «Интеллектуальный сервер поведенческого анализа» активирует методы машинного обучения и анализа естественного языка по расписанию, путем запроса данных о предстоящем анализе из базы данных и дальнейшего размещения в ней результатов.

К системе возможно подключение нескольких модулей клиент серверного взаимодействия и серверов управления Web интерфейсом. При отказе одного из модулей, оставшиеся продолжают функционировать в полном объеме.

Выгрузка пользовательских данных и результатов анализа возможна лишь через серверный модуль «*Web API»,* благодаря чему осуществляется изоляция данных и невозможность их получения через другие внешние модули системы. Для повышения уровня безопасности пользовательских данных, Web модуль администрирования системы и WebAPI интерфейса, могут быть изолированы от внешней сети, что так же повысит уровень сохранности конфиденциальных данных пользователей системы. Отдельно реализованный логический модуль клиент-серверного взаимодействия может быть развернут как во внутренней сети организации, так и во внешней независимо от способа размещения остальных модулей системы, что позволяет осуществлять сбор данных пользователей как внутри контура организации, посредствам внутренней сети, так и при помощи сети интернет.

4.2.2 Мобильный агент сбора поведенческой информации

Сбор поведенческих данных осуществляет отдельный модуль программного комплекса, именуемый как «Мобильное приложение - агент». Данный модуль реализован в виде приложения агента, работающего в

фоновом режиме мобильного устройства на базе операционной системы Android.

После установки мобильного приложения – агента требуется его первичная настройка администратором системы. После ручного конфигурирования приложение работает незаметно, не отвлекая пользователя от целевых задач. Работая в фоне, мобильный агент запрашивает список сформированного для него набора команд у модуля клиент-серверного взаимодействия раз в *N* минут. Использование временного интервала *N = 2* позволяет достигнуть стабильного отклика от мобильного устройства пользователя, что в свою очередь положительно влияет на актуальность получаемых администратором системы пользовательских данных.

Мобильное приложение - агент использует интерфейсы взаимодействия с системными оптимизаторами устройства для его поддержания в активном режиме и экономии заряда батареи мобильного устройства пользователя.

Мобильное приложение - агент имеет два основных режима работы:

- *Сбор с моментальной отправкой*. Доступен при активации данного режима администратором системы и при наличии соединения с модулем клиент-серверного взаимодействия;
- *Сбор с архивацией.* Доступен при активации данного режима администратором или при включении режима сбора с моментальной отправкой и отсутствия соединения с модулем клиент-серверного взаимодействия.

При отсутствии связи с модулем клиент серверного взаимодействия, мобильный клиент продолжает выполнение полученных ранее команд и сохранение результатов в локальную базу данных SQLite [93]. При восстановлении соединения с сервером и наличии соответствующей команды на выгрузку, сохраненные в локальную базу данные отправляются в модуль клиент серверного взаимодействия и в дальнейшем могут быть доступны администратору системы для последующего просмотра после процедуры системной обработки.

Мобильное приложение агент состоит из следующих основных логических модулей:

- *Модуль сбора*. Осуществляет сбор и базовую фильтрацию набираемых пользовательских текстов для дальнейшей передачи на сервер или сохранения в локальную базу данных;
- *Локальная БД*. Хранит персональные пользовательские текстовые наборы данных на мобильном устройстве для их последующей выгрузки;
- *Модуль клиент-серверного взаимодействия.* Осуществляет процедуру обмена данными между пользовательским устройством и сервером, а также прием команд, предназначенных для выполнения.

Общая структура мобильного приложения представлена на рис.4.13.

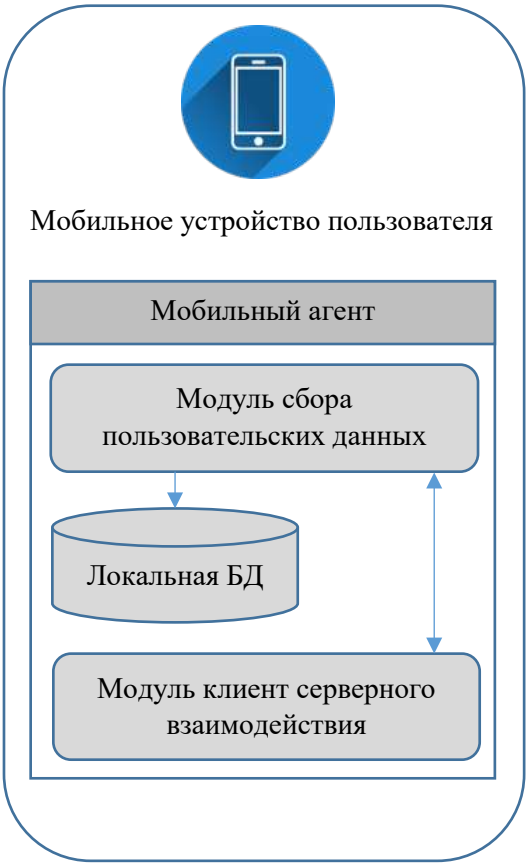


Рис. 4.13 – Архитектура экспериментального образца программного обеспечения

Сбор текстовых данных осуществляется путем перехвата пользовательского ввода при помощи использования возможностей системного сервиса (Accessibility Service) операционной системы Android. По умолчанию мобильный агент собирает пользовательский ввод из всех приложений, однако администратор системы может указать фильтры используя Web интерфейс управления.

4.2.3 Модуль поведенческого анализа

Серверный модуль поведенческого анализа состоит из двух логических частей:

- Серверный модуль;
- Анализирующие скрипты.

Серверный модуль реализован на базе платформы .NET CORE [94] на языке C#, и предоставляет API (Application programming interface) для доступа к методам анализа и дальнейшей возможности взаимодействия с сторонними решениями [95-96].

Обеспечение взаимодействия со сторонними модулями, в разработанной системе, реализовано посредствам выборки заранее сформированных значений расписания из базы данных. Анализ отклонений в поведении пользователей осуществляется при помощи внутреннего вызова скрипта, передачи в него наборов коротких пользовательских текстов и получении результатов после обработки текстовых массивов пользовательских данных. Реализация скрипта выполнена на языке Python 3 [97] в среде Jupyter Notebook [98]. Данный скрипт использует методы естественного анализа языка и соответствующие модели (BOW, TF-IDF, BERT, Word2Vec, GloVe), предобработка текстов осуществляется при помощи методов библиотеки NLTK [99]. Отображение целевых тематик общения, за выбранный анализируемый период, реализовано при помощи библиотеки Matplotlib [100] и WordCloud [101].

4.2.4 Серверные модули обработки информации

Разработанный экспериментальный образец программного обеспечения включает в себя следующие серверные компоненты:

- WebAPI модуль администрирования системы (Dashboard);
- API модуль клиент серверного взаимодействия.

WebAPI модуль администрирования системы представлен в виде API сервиса, осуществляющего взаимодействие между клиентским интерфейсом и базой данных. Реализация данного серверного модуля выполнена при помощи технологии .NET Core 3 на языке C#. Методы API сервиса осуществляют выборку запрашиваемых данных администратором, инициируемые через Web интерфейс системы.

Данный модуль имеет следующую структуру:

- *Методы приема Web запросов.* Осуществляют прием запросов, отправляемых интерфейсом системы, вызывают соответствующий обрабатывающий метод и возвращают результат на Web интерфейс администратору;
- *Методы обработки данных.* Получают на вход сформированные данные в требуемом виде, осуществляют их обработку с последующей записью в базу данных, формируют результирующий ответ и возвращают его.

Реализация модуля клиент-серверного взаимодействия для обработки получаемых от мобильных устройств пользовательских текстовых наборов, и их дальнейшей записи в базу данных так же выполнена в виде API сервиса на базе .NET CORE 3. Клиент-серверный модуль отправляет клиентскому устройству информацию, содержащую статус выполнения команды. В случае успешной записи полученных значений в базу данных, мобильное приложение агент, получает соответствующий ответный статус и удаляет текстовые данные, накопленные ранее на мобильном устройстве. При формировании сервером значения ошибки в результате выполнения запроса на запись полученных данных или проблем, связанных со стабильностью сети у

целевого мобильного устройства, накопленные поведенческие данные, а именно наборы коротких пользовательских текстов, не удалятся с клиентского устройства мобильным агентом, и будут отправлены повторно при последующем запросе на выборку данных. Выбранный подход позволяет сохранить целостность данных и сократить возможность потери уникальных пользовательских текстовых наборов из-за программных или аппаратных сбоев, связанных с передачей или конечной записью в базу данных.

Получение и выборка поведенческих данных посредствам использования данного модуля невозможна. Благодаря применению выбранного подхода сохраняется возможность сокрытия основных обрабатывающих модулей системы, оперирующих с пользовательскими данными, внутри инфраструктуры организации, что позволяет избежать инцидентов, связанных с получением доступа к пользовательским данным третьими лицами.

Клиент-серверный обмен осуществляется в виде приема и передачи пакетов данных в формате JSON [101] в зашифрованном виде по протоколу HTTPS, что так же позволяет избежать получения доступа к данным третьими лицами.

4.3 Экспериментальная проверка показателей производительности

В рамках диссертационного исследования впервые были получены показатели производительности следующих разработанных модулей экспериментального образца программного обеспечения:

- Мобильное приложение – агент;
- API модуль клиент-серверного взаимодействия;
- Серверный модуль управления Web интерфейса;
- Интеллектуальный сервер поведенческого анализа.

4.3.1 Показатели производительности мобильного приложения - агента

Важнейшими показателями качества работы мобильного приложения агента являются его стабильность, производительность и энергоэффективность на протяжении всего времени его активности. Стабильность функционирования мобильного приложения влияет на непрерывность и качество сбора текстовых данных пользователей. В случае низкого уровня стабильности модулей мобильного приложения агента, связанного с возникновением критических ошибок, конфликтов с оптимизаторами операционной системы и оболочкой устанавливаемой производителем устройства, время получения текстовых наборов пользовательских данных может быть не стабильным и сильно больше значений, установленных администратором системы. При блокировке активности мобильного приложения агента оптимизаторами системы и средствами оболочки производителя, взаимодействие с устройством может быть временно приостановлено или остановлено полностью.

Для обеспечения высокого уровня стабильности работы мобильного приложения агента на различных мобильных устройствах пользователей, были решены задачи для осуществления взаимодействия с операционной системой, а именно:

- Определение оптимального интервала запроса команд у модуля клиент серверного взаимодействия;
- Реализация взаимодействия с системным оптимизатором «DOZE» в ОС Android [102];
- Реализация взаимодействия с оптимизаторами оболочек различных производителей мобильных устройств;
- Реализация алгоритмов автоматического перезапуска мобильного приложения агента в случае критической ошибки или его закрытия.

Сравнительные результаты работы модулей мобильного приложения агента с использованием, представленных ранее средств оптимизации представлены в таблице 4.1.

Таблица 4.1. Сравнительный анализ времени стабильной работы мобильного приложения агента с интеграций с оптимизаторами и без нее

№	Кол-во устройств	Среднее время активной работы (часов в сутки) / количество устройств со сбоями	Результирующие значения во временном интервале		
			7 дней	28 дней	180 дней
Без использования решений взаимодействия с оптимизаторами					
1	10	Время	11,5	7,2	2,4
		Кол-во	6	7	9
2	50	Время	11,4	7,1	2
		Кол-во	32	36	47
3	100	Время	11,4	5,3	1,7
		Кол-во	65	77	96
С использованием решений взаимодействия с оптимизаторами					
4	10	Время	23,1	22,6	22,1
		Количество	1	1	2
5	50	Время	23,8	23,6	22,6
		Количество	4	5	7
6	100	Время	23,8	23,7	22,8
		Количество	8	9	11

Для тестирования использовались мобильные устройства с ОС Android 7 и выше. Для получения результирующих значений стабильности работы мобильного приложения агента во времени анализировались данные об ошибках получаемые при помощи платформы Google «Firebase Crashlytics», а также собранных значений истории отклика мобильных устройств.

По значениям, представленным в таблице 4.1 получено, что использование решений для взаимодействия с системными оптимизаторами сильно повышает стабильность работы приложения агента на мобильных устройствах конечных пользователей системы. По значениям, представленным в экспериментах №4, 5, 6 наблюдается значительное увеличение времени стабильной работы мобильного приложения агента. По значениям экспериментов № 1, 2, 3 видно, что средства оптимизации со временем блокируют активность приложения на большинстве устройств, что

не позволяет использовать мобильный агент без его интеграции с системами оптимизации ввиду потери контроля над управляемыми устройствами и, следовательно, дальнейшей невозможностью сбора наборов пользовательских текстов.

Производительность работы мобильного приложения агента прямолинейно коррелирует с энергозатратами, требуемыми на ее поддержание. При высокой и длительной нагрузке на ЦП, связанной с локальной обработкой данных, возрастает энергопотребление, и снижается время автономной работы мобильного устройства, что в свою очередь затрудняет его полноценное использование целевыми пользователями, а также и вносит изменения в стандартный сценарий его эксплуатации.

Экспериментально полученные значения производительности мобильного приложения агента выглядят следующим образом. Пиковые значения нагрузки, зафиксированные при одновременном сборе пользовательских текстов и отложенной выгрузке ранее полученных наборов данных, составляют 11-18% в зависимости от модели мобильного устройства, а длительность составляет до 30 секунд в зависимости от скорости выгрузки данных на сервер. В режиме сбора и сохранения данных в локальную базу SQLite значение загрузки ЦП задачами мобильного агента не превышает 2%.

При таких режимах работы достигается достаточно высокий уровень энергоэффективности, что позволяет длительно эксплуатировать мобильное устройство в полной мере без изменения сценария его использования. Экспериментально получены следующие значения энергопотребления мобильного приложения агента в пределах полного цикла заряда мобильного устройства:

- Работа агента мониторинга в фоне без выполнения задач сбора и выгрузки (ожидание получения команды) - до 1%;
- Работа агента мониторинга в режиме сбора и сохранении в локальную базу данных – до 5%;

- Работа агента мониторинга в режиме сбора текстовых данных и их постоянной отправки – до 9%.

Полученные значения энергопотребления не являются значительными относительно общего объема заряда, что позволяет пользователям использовать мобильное устройство в стандартных сценариях.

4.3.2 Показатели производительности серверных модулей

Ввиду ограниченного объема машинного времени и его высокой стоимости в кластерных системах, особое внимание производительности уделяется решениям, в которых осуществляются длительные вычисления, а также Web- приложениям, где время формирования информации на конечной странице, для конечного пользователя, критично и зависит от серверных мощностей.

Под понятием производительности понимается его реактивность и продуктивность:

- продуктивность – объем информации, обрабатываемой системой за единицу времени;
- реактивность – время между предъявлением системе входных данных и появлением соответствующей выходной информации.

Очевидной и логичной задачей, является увеличение производительности, которая формально являющейся задачей оптимизации. Критерием оптимизации является некоторая функция:

$$y = \varphi(x_1, x_2, \ldots, x_n) \tag{9}$$

где:

- y – время обработки информации;
- x_1, x_2, x_n - факторы влияющие прямым или косвенным способом на производительность системы;

- $x_i \in [a_i, b_i]$ - область определения i-го фактора, являющаяся ограничением задачи.

Выбор факторов, влияющих на производительность программы, является не тривиальной задачей. Как правило, современные программные комплексы имеют большое число связей и зависимостей так как любая программа функционирует в операционной системе и взаимодействует с иными сервисами. По данной причине требуется выбрать основные факторы прямым образом, влияющие на конечную производительность программного комплекса. Для решения поставленной задачи предлагается использовать аппарат, разработанный в теории математического планирования эксперимента (МПЭ) [103].

Одной из основных идей планирования эксперимента состоит в использовании для исследуемого объекта кибернетической абстракции черного ящика [104] (рис. 4.14).

Рис. 4.14 – Абстракция черного ящика

Данная абстракция предполагает менее детальное рассмотрение системы и внутренних процессов анализируемого объекта из-за их большой сложности. Эксперимент сводится к анализу воздействующих на объект входных параметров, и результирующих выходных значениях. Влияющие на производительность факторы и их важность будут рассмотрены далее в данной главе.

Количество пользователей в экспериментальной выборке составило 100 человек. Экспериментальные исследования проводились с использованием следующего аппаратного обеспечения:

- ЦП – Intel XEON E5 - 2678v3;

- Тех.процесс: 22 нм;
- Ядер: 12;
- Потоков: 24;
- Базовая частота: 2500 МГц;
- Максимальная частота в режиме Turbo Boost: 3300 МГц;
- Кэш: 30 Mb.

- ОЗУ – 64 Гб, DDR3 (1833 МГц);
- HDD RAID 1 (2x4Тб)
  - SATA 3;
  - Кэш память 256 Мб;
  - Максимальная скорость передачи данных 255 Мб/сек.
- Сетевая карта 1 Гбит.

4.3.2.1 API модуль клиент-серверного взаимодействия

Основным узлом системы, являющимся непосредственным посредником между пользовательскими устройствами и системными модулями является API модуль клиент серверного взаимодействия, чья производительность напрямую зависит от следующих основных факторов:

- Количество подключенных устройств (мобильных агентов);
- Режим работы подключенных устройств;
- Объем выгружаемых данных;
- Частота клиент серверного взаимодействия.

Данные факторы являются весомыми и отражают основные характеристики производительности анализируемого модуля клиент-серверного взаимодействия.

Формирование значения интервала (частоты) обмена данными между клиентом и сервером производилось с целью одновременного поддержания высокого уровня производительности системы и сохранения стабильности клиент серверного взаимодействия. Экспериментально получено, что при выборе интервала более 5 минут сильно ухудшается возможность

взаимодействия с мобильными агентами ввиду сложения интервала обновления команд, накладных сетевых и временных расходов, особенностей оптимизаторов операционной системы и оболочки. При выборе значения интервала обновления команд менее двух минут, частота запросов становится излишней, что сказывается на снижении производительности мобильного агента и модуля клиент-серверного взаимодействия, повышении объемов используемого трафика, при отсутствии повышения качества взаимодействия клиентских модулей агентов с модулем клиент-серверного взаимодействия.

Для оценки производительности системы при выбранном интервале клиент серверного обмена равного 2 минутам были проведены экспериментальные исследования обработки данных устройств (10, 50 и 100) в различных режимах, результирующие значения представлены в таблице 4.2.

Таблица 4.2. Результирующие значения производительности модуля клиент-серверного взаимодействия в зависимости от режимов работы клиентского приложения агента

№	Режим работы мобильного клиента	Кол-во мобильных устройств	Загрузка ЦП сервисом (%)		Загрузка ОЗУ сервисом (Мб)		Загрузка диска Сервисом (%)	
			Ср.	Макс.	Ср.	Макс.	Ср.	Макс.
1	Ожидание команд	10	2	6	34	52	1	2
		50	3	7	40	55	1	2
		100	4	9	52	70	1	2
2	Сбор текста в локальную БД	10	2	6	35	53	1	2
		50	3	7	39	54	1	2
		100	4	9	52	68	1	2
3	Сбор текста и отправка	10	4	8	55	62	2	4
		50	6	9	66	98	4	9
		100	10	13	71	89	7	13
4	Единовременная выгрузка сохраненных ранее текстовых наборов	10	8	12	76	92	11	16
		50	27	35	142	179	17	35
		100	38	44	326	425	34	52

Ожидание команд и сбор текстовых данных в локальную базу мобильного приложения агента, производят практически идентичную нагрузку на модуль клиент-серверного взаимодействия ввиду высокой схожести процессов взаимодействия. Максимальная (пиковая) и средняя

нагрузка в данных режимах минимальна и не является критической, даже в результатах экспериментов, полученных при тестировании с использованием 100 пользовательских мобильных устройств.

Более ресурсоемкой операцией является сбор текста и его отправка на сервер (активная фаза работы мобильного агента). При тестировании на 100 пользователях средняя нагрузка на аппаратное обеспечения, представленная в эксперименте 3, превышает значения экспериментов 1 и 2, однако не является критической.

По результатам эксперимента 4 наблюдаются достаточно высокие средние и максимальные (пиковые) значения нагрузки на аппаратное обеспечение при работе мобильных устройств в режиме единовременной выгрузки, собранной за выбранный период информации.

По результатам проведенных экспериментов установлено, что целесообразно использовать мобильный агент в режиме ожидания команд и режиме сбора текста и его моментальной выгрузки, так как данные режимы предоставляют администратору системы полную информацию в актуальном виде и минимально нагружают модуль клиент серверного взаимодействия, что повышает его стабильность и снижает риск возникновения отказоустойчивых ситуаций.

Режимы «Сбор текста в локальную БД» и «Единовременная выгрузка сохраненных ранее текстовых наборов» целесообразно использовать при наличии лимитного подключения к сети у мобильного устройства для экономии трафика, либо в случае если известно, что устройство не имеет возможности выгрузки через мобильную сеть, и обмен данными осуществляется только через корпоративную точку доступа Wi-Fi.

4.3.2.2 Серверный модуль управления Web интерфейса

От уровня производительности серверного модуля Web интерфейса системы напрямую зависит время формирования информации на конечной странице, отображаемой администратору системы. Проблемы с доступом к

Web - интерфейсу могут осложнить управление мобильными устройствами и затруднить своевременное получение актуальной информации из базы данных о пользователях в выбранном временном промежутке. По данной причине серверный модуль управления Web интерфейса реализован как отдельное решение, взаимодействующее с базой данных и формирующее отображаемые результирующие значения. При высоком уровне нагрузки на модуль клиент-серверного взаимодействия и возможном возникновении его отказа от обслуживания, администратор системы, работающий в Web интерфейсе, не теряет управление над мобильными устройствами и имеет возможность редактирования его режимов работы и получения данных о выбранных пользователях. Благодаря изолированности его функциональных возможностей, от остальных модулей системы, возможно размещение (сокрытие) панели управления внутри контура и осуществление к ней доступа непосредственно только в сети организации, при одновременной к модулю клиент серверного взаимодействия мобильными устройствами из внешней сети. При использовании данного подхода повышается стабильность работы обоих серверных модулей, так как они независимы друг от друга.

Экспериментально получены значения времени загрузки страниц при одновременном многопользовательском доступе к Web – интерфейсу системы, результаты представлены в таблице 4.3.

Таблица 4.3. Результирующие значения производительности модуля клиент-серверного взаимодействия в зависимости от режимов работы клиентского приложения агента

№	Количество одновременных запросов обновления страницы	Время загрузки страницы (секунд)	
		Среднее	Максимальное
1	1	1,2	1,91
2	10	1,2	1,93
3	100	1,7	2,11
4	1000	2,73	3,15
5	10000	4,82	6,19

Для определения средних и максимальных значений времени загрузки Web – интерфейса системы была проведена серия проверок из 100 повторов для каждой позиции эксперимента, графическое представление результатов представлено на рис. 4.15.

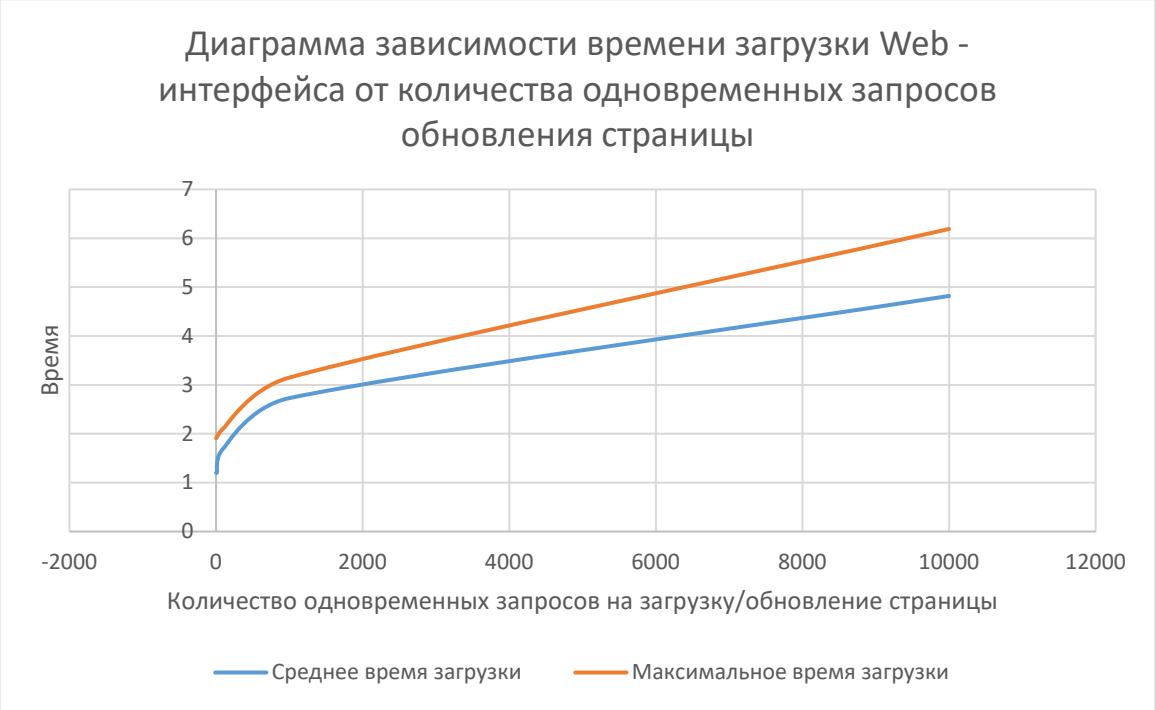


Рис. 4.15 – Диаграмма зависимости времени загрузки Web – интерфейса от количества одновременных запросов страницы

Проверка на высоких количественных значениях одновременных запросов к ресурсу производилась с целью визуализации уровня стабильности работы Web интерфейса системы.

По результатам проведенных экспериментов была получена количественная оценка производительности и стабильности работы Web сервиса. Полученные показатели скорости работы серверного модуля при средних и высоких нагрузках, позволяют администратору системы осуществлять быстрый доступ к интерфейсу и своевременно получать требуемую информации о деятельности выбранного пользователя в указанном временном промежутке, а также управлять мобильными агентами посредствам добавления выполняемых ими задач в планировщик.

4.3.3 Показатели производительности разработанного метода поиска аномального поведения

Анализ пользовательских данных с последующим формированием результирующих значений поведенческих отклонений осуществляется отдельным независимым модулем системы «Серверный модуль анализа отклонений». Для получения более детальной информации о производительности разработанного метода, рассмотрим его отдельно от серверного модуля. Анализ производился на реальных пользовательских текстовых выборках различной длины в диапазонах от 7 до 28 дней. Результаты эксперимента представлены в таблице 4.4.

Таблица 4.4. Значения производительности работы метода идентификации нетиповых сценариев использования устройства

№	ID	Интервалы дат	Длина текста в выборке (символов)	Время выполнения запроса (секунд)	Загрузка ЦП сервисом (%)	Объем занятой сервисом ОЗУ (Гб)
Выборка двух интервалов по 7 дней						
1	144	1.01.2021-7.01.2021; 8.01.2021-14.01.2021.	30618; 22866	89	6-8	2
2		14.01.2021-21.01.2021; 21.01.2021-28.01.2021.	8105; 1418	88	6-8	1,5
3		1.02.2021-7.02.2021; 7.02.2021-14.02.2021.	20858; 20703	95	7-12	1,8
4		14.02.2021-21.02.2021; 21.02.2021-28.02.2021.	63591 2700	91	5-6	2,1
Выборка двух интервалов по 14 дней						
5	155	1.01.2021-14.01.2021; 14.01.2021-28.01.2021.	6459; 4287	82	9-13	1,4
6		1.02.2021-14.02.2021; 14.02.2021-28.02.2021.	19646; 53203	97	6-13	2,3
7		1.03.2021-14.03.2021; 14.03.2021-28.03.2021.	8552; 18639;	86	6-9	2,3
8		1.04.2021-14.04.2021; 14.04.2021-28.04.2021.	21843; 15485;	93	6-9	2,3
Выборка двух интервалов по 28 дней						
9	175	1.02.2021-28.02.2021; 1.03.2021-28.03.2021.	13761; 28803;	99	7-15	2,6
10		1.04.2021-28.04.2021; 1.05.2021-28.05.2021.	13336; 18304	91	7-12	2,3
11	186	1.03.2021-28.03.2021; 1.04.2021-28.04.2021.	34891; 63422;	107	9-16	2,6
12		1.05.2021-28.05.2021; 1.06.2021-28.06.2021	57223; 38090	105	9-15	2,6

Разница во времени обрабатываемых наборов текстовых данных пользователей обусловлена длиной текстовой выборки и ее содержимым. Таким образом минимальное время анализа было зафиксировано в эксперименте №5 и составило 82 секунды, максимальное в эксперименте №11 (107 секунд).

Для повышения скорости обработки наборов коротких пользовательских текстов, на используемом аппаратном обеспечении, с целью поведенческого анализа, возможно применение параллельного запуска анализирующего метода. Параллельный анализ пользовательских текстовых наборов позволяет сократить время обработки данных.

Для распределения нагрузки и сохранения стабильности работы серверного модуля анализа отклонений применяется запуск анализирующего метода по расписанию, установленному администратором системы. Рекомендуемое время запуска анализа - 01:00, так как в данное время пользовательская активность минимальна, а получаемые результирующие данные полностью охватывают пользовательскую активность, осуществляемую за предыдущий день. Запуск модуля по расписанию производится автоматически, однако возможен ручной запуск анализа для выбранных администратором системы пользователей.

Максимальный объем используемой ОЗУ и ЦПУ зафиксирован при анализе выборки двух интервалов по 28 дней в эксперименте №11 и составляет 2,6 Гб, ЦПУ – 16%, минимальный - в эксперименте №2 при анализе выборки двух интервалов по 7 дней.

В общем случае значения используемых машинных ресурсов, полученные в экспериментах 1-12, не является критическими. В результате проводимых экспериментов не было зафиксировано отказов системы и ошибок при анализе данных.

4.4 Апробация программного комплекса

Разработанный экспериментальный образец программного комплекса сбора текстовых данных пользователей и идентификации нетиповых сценариев использования мобильных устройств апробирован в рамках выполнения следующих работ:

- Грант ФСИ №13121ГУ/2018 «Разработка автоматизированной системы анализа и контроля деятельности сотрудников»;
- Грант РФФИ 19-37-90111 «Использование методов и алгоритмов анализа данных и машинного обучения в информационных системах»;
- Грант ФСИ 4ГАИИС13-D7/72316 «Исследование применимости методов машинного обучения для поиска аномальной активности в поведении пользователей и разработка программного обеспечения для обнаружения поведенческих отклонений».

Разработанные программные модули были апробированы в следующих организациях:

- ООО «ГУД ФУД»;
- ООО «Кадастр Экспресс».

4.5 Выводы

Разработан экспериментальный образец программного обеспечения идентификации нетиповых сценариев использования мобильных устройств, реализующий сбор наборов пользовательских текстов с мобильных устройств пользователей, формирование на их основе индивидуальных поведенческих моделей и их анализ.

Представлено описание сценариев использования экспериментального образца программного комплекса, включающее в себя установку и первичную настройку мобильного приложения, использование мобильного устройства с установленным агентом, сбор поведенческих данных, использование Web

интерфейса для управления мобильными устройствами и сбора данных, а также анализ нетиповых сценариев использования мобильных устройств.

Представлена программная реализация экспериментального образца программного комплекса, включающая в себя проектирование архитектуры программного комплекса, состоящего из мобильного агента сбора поведенческой информации, модуля поведенческого анализа и серверных модулей обработки информации.

Проведен эксперимент и получены показатели производительности мобильного приложения агента и серверных модулей, а именно модуля клиент серверного взаимодействия и API модуля управления веб интерфейса, а также метода идентификации нетиповых сценариев использования мобильных устройств.

Представлена информация о апробации программного комплекса в трех грантовых программах и внедрении в двух организациях.

5 ЗАКЛЮЧЕНИЕ

В ходе выполнения диссертационной работы, получены следующие основные результаты:

1. Осуществлен обзор решений и методов анализа данных для идентификации изменений в поведении пользователей и их применение в современных интеллектуальных системах машинного обучения;

2. Разработан метод предварительной обработки накапливаемой текстовой информации из коротких выборок, обеспечивающий уменьшение информационного шума, отличающийся предварительным формированием оптимальной длинны коротких последовательностей пригодных для дальнейшего анализа длиной от 7 до 100 символов;

3. Разработан метод идентификации нетиповых сценариев использования мобильных устройств пользователями по наборам коротких текстовых данных, отличающийся применением методов машинного обучения, анализа естественного языка (bag-of-words, TF-IDF, Word2Vec, GloVe, BERT) и метрик сходства, обеспечивающий сокращение объемов анализируемой экспертами вручную текстовой информации, собираемой с мобильных устройств пользователей, а так же экономное потребление вычислительных ресурсов;

4. Разработана архитектура программного комплекса идентификации нетиповых сценариев использования мобильных устройств пользователями, отличающаяся расширяемой модульной структурой, обеспечивающая сбор биометрических данных, содержащих пользовательские поведенческие характеристики, и идентификацию нетиповых сценариев использования мобильного устройства на их основе;

5. Разработан программный комплекс, реализующий предложенные методы предварительной обработки накапливаемой текстовой информации из коротких выборок и идентификации нетиповых сценариев использования мобильных устройств, обеспечивающий сбор и анализ биометрических

данных, уменьшение информационного шума и экономное потребление вычислительных ресурсов;

6. Проведено экспериментальное исследование разработанных методов и программного комплекса идентификации нетиповых сценариев использования мобильных устройств и анализ полученных результатов.

Разработанный экспериментальный образец программного комплекса сбора текстовых данных пользователей и идентификации нетиповых сценариев использования мобильных устройств апробирован в рамках выполнения грантовых работ ФСИ №13121ГУ/2018 «Разработка автоматизированной системы анализа и контроля деятельности сотрудников», РФФИ 19-37-90111 «Использование методов и алгоритмов анализа данных и машинного обучения в информационных системах», ФСИ 4ГАИИС13-D7/72316 «Исследование применимости методов машинного обучения для поиска аномальной активности в поведении пользователей и разработка программного обеспечения для обнаружения поведенческих отклонений».

Также разработанные модули программного комплекса были апробированы и успешно внедрены в деятельность в следующих организаций ООО «ГУД ФУД» и ООО «Кадастр Экспресс».

Полученные в ходе выполнения диссертационной работы результаты могут послужить основой как для построения новых перспективных систем поведенческого анализа, так и внедрения разработанных методов и алгоритмов в уже существующие.

Направлением дальнейших исследований является: исследование применимости разработанного метода идентификации нетиповых сценариев использования мобильных устройств со сторонними наборами данных, извлекаемыми из различных информационных источников DLP, SIEM, UBA систем, с целью своевременного получения краткой агрегированной информации о деятельности пользователей и идентификации сценариев использования анализируемых устройств.

6 СПИСОК ЛИТЕРАТУРЫ

1. Котенко И. В. и др. Выявление инсайдеров в корпоративной сети: подход на базе UBA и UEBA //Защита информации. Инсайд. – 2019. – №. 5. – С. 26-35.
2. Ушаков И. А. и др. АГРЕГАЦИЯ И ПРЕДОБРАБОТКА БОЛЬШИХ ДАННЫХ С ЦЕЛЬЮ ОПРЕДЕЛЕНИЯ ИНСАЙДЕРОВ В КОРПОРАТИВНОЙ СЕТИ //МОЛОДЕЖНАЯ НАУКА КАК ФАКТОР И РЕСУРС ИННОВАЦИОННОГО РАЗВИТИЯ. – 2020. – С. 10-19.
3. Поляничко М. А. Методика обнаружения аномального взаимодействия пользователей с информационными активами для выявления инсайдерской деятельности //Труды учебных заведений связи. – 2020. – Т. 6. – №. 1. – С. 94-98.
4. Поляничко М. А. Моделирование действий инсайдеров на основе аппарата информатики поведения //Естественные и технические науки. – 2018. – №. 12. – С. 441-443.
5. Корниенко А. А., Поляничко М. А. Метод обнаружения инсайдерской деятельности в организации //Программные системы и вычислительные методы. – 2019. – №. 1. – С. 30-41.
6. Машечкин И. В., Петровский М. И., Царёв Д. В. Методы машинного обучения для анализа поведения пользователей при работе с текстовыми данными в задачах информационной безопасности //Вестник Московского университета. Серия 15. Вычислительная математика и кибернетика. – 2016. – №. 4.
7. Машечкин И. В., Петровский М. И., Трошин С. В. Мониторинг и анализ поведения пользователей компьютерных систем. – 2008.
8. Королёв В. Ю. и др. Применение временных рядов в задаче фоновой идентификации пользователей на основе анализа их работы с текстовыми данными //Труды Института системного программирования РАН. – 2015. – Т. 27. – №. 1. – С. 151-172.

9. Машечкин И. В., Петровский М. И. СИСТЕМА МОНИТОРИНГА РАБОТЫ ПОЛЬЗОВАТЕЛЕЙ С ИНФОРМАЦИОННЫМИ РЕСУРСАМИ КОРПОРАТИВНОЙ КОМПЬЮТЕРНОЙ СЕТИ НА ОСНОВЕ МОДЕЛИРОВАНИЯ ПОВЕДЕНИЯ ПОЛЬЗОВАТЕЛЕЙ С ЦЕЛЬЮ ПОИСКА АНОМАЛИЙ И ИЗМЕНЕНИЙ В РАБОТЕ. – 2011.
10. Казачук М. А. и др. Методы поиска исключений в потоках сложноструктурированных данных //Вестник Московского университета. Серия 15. Вычислительная математика и кибернетика. – 2019. – №. 3. – С. 17-28.
11. Отраднов К. К., Жуков Д. О., Новикова О. А. Модель кластеризации слабоструктурированных текстовых данных //Современные информационные технологии и ИТ-образование. – 2017. – Т. 13. – №. 3. – С. 100-115.
12. Отраднов К. К., Раев В. К. Экспериментальное исследование эффективности методик векторизации текстовых документов и алгоритмов их кластеризации //Вестник Рязанского государственного радиотехнического университета. – 2018. – №. 64. – С. 73-84.
13. Сигов А. С. и др. Психолингвистический анализ русскоязычных текстовых сообщений на основе их фоносемантических статистических характеристик //Информатика и её применения. – 2017. – Т. 11. – №. 3. – С. 80-89.
14. Liu H., Lang B. Machine learning and deep learning methods for intrusion detection systems: A survey //applied sciences. – 2019. – Т. 9. – №. 20. – С. 4396.
15. Liu H. et al. CNN and RNN based payload classification methods for attack detection //Knowledge-Based Systems. – 2019. – Т. 163. – С. 332-341.
16. Alharbi A. S. M., de Doncker E. Twitter sentiment analysis with a deep neural network: An enhanced approach using user behavioral information //Cognitive Systems Research. – 2019. – Т. 54. – С. 50-61.

17. Alharbi A. S. M., de Doncker E. Emotional Awareness based Classification Model for Twitter Sentiment Analysis using a Deep Neural Network //Proceedings on the International Conference on Artificial Intelligence (ICAI). – The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2019. – С. 142-145.
18. Лёвин Б. А., Цветков В. Я. Информационные процессы в пространстве «больших данных» //Мир транспорта. – 2017. – Т. 15. – №. 6. – С. 20-30.
19. Кузнецов С. Д., Велихов П. Е., Фу Ц. Аналитика в реальном времени, гибридная транзакционная/аналитическая обработка, управление данными в основной памяти и энергонезависимая память //Труды института системного программирования РАН. – 2021. – Т. 33. – №. 3. – С. 171-198.
20. Ghani N. A. et al. Social media big data analytics: A survey //Computers in Human Behavior. – 2019. – Т. 101. – С. 417-428.
21. Usman N. et al. Intelligent dynamic malware detection using machine learning in IP reputation for forensics data analytics //Future Generation Computer Systems. – 2021. – Т. 118. – С. 124-141.
22. Waheed N. et al. Security and privacy in IoT using machine learning and blockchain: Threats and countermeasures //ACM Computing Surveys (CSUR). – 2020. – Т. 53. – №. 6. – С. 1-37.
23. Li W. et al. A comprehensive survey on machine learning-based big data analytics for IoT-enabled smart healthcare system //Mobile Networks and Applications. – 2021. – Т. 26. – №. 1. – С. 234-252.
24. Babar M., Tariq M. U., Jan M. A. Secure and resilient demand side management engine using machine learning for IoT-enabled smart grid //Sustainable Cities and Society. – 2020. – Т. 62. – С. 102370.
25. Биктимиров М. Р., Елизаров А. М., Щербаков А. Ю. Тенденции развития технологий обработки больших данных и инструментария хранения разноформатных данных и аналитики //Электронные библиотеки. – 2016. – Т. 19. – №. 5. – С. 390-407.

26. Смирнов Д. В. и др. Система сбора и анализа информации из различных источников в условиях Big Data //International Journal of Open Information Technologies. – 2021. – Т. 9. – №. 4. – С. 64-71.
27. Банокин П. И. Система классификаторов на основе нейронной сети для идентификации аномального поведения пользователей корпоративного программного обеспечения //Молодежь и современные информационные технологии: сборник трудов XIV Международной научно-практической конференции студентов, аспирантов и молодых ученых, г. Томск, 7-11 ноября 2016 г. Т. 2.—Томск, 2016. – Изд-во ТПУ, 2016. – Т. 2. – С. 120-121.
28. Возмитель И. Г. Информационные технологии в менеджменте: экскурс в прошлое и будущее //Современные информационные технологии и ИТ-образование. – 2015. – Т. 1. – №. 11. – С. 585-592.
29. Орлова Ю. А., Репина И. Б., Чуднова О. А. ЦИФРОВАЯ ТРАНСФОРМАЦИЯ МЕТОДОВ И СРЕДСТВ КОНТРОЛЯ КАЧЕСТВА //Компетентность. – 2022. – №. 4. – С. 22-25.
30. LaueA T. et al. A SIEM Architecture for Advanced Anomaly Detection. – 2022.
31. Мир А.В., Кумар К.Р.Р. Расширенная реализация системы управления безопасностью (SSMS) с использованием UEBA в SCADA-системах на основе интеллектуальных сетей // Применение машинного интеллекта в инженерии. – CRC Press, 2022. – С. 1-11.
32. Martín A. G. et al. Combining user behavioural information at the feature level to enhance continuous authentication systems //Knowledge-Based Systems. – 2022. – Т. 244. – С. 108544.
33. LaueA T. et al. A SIEM Architecture for Advanced Anomaly Detection. – 2022.
34. Обзор решений SIEM (Security information and event management) // Kickidler. Обзор и сравнение DLP систем 2022 года URL: https://www.kickidler.com/ru/info/obzor-i-sravnenie-luchshix-besplatnyix-open-source-dlp-sistem.html (дата обращения: 25.02.2021).

35. Обзор решений SIEM (Security information and event management) // Habr URL: https://habr.com/ru/company/roi4cio/blog/528770/ (дата обращения: 20.05.2021).
36. Самые популярные СЭД/ECM-системы // TAdviser URL: https://www.tadviser.ru/index.php/Статья:Самые_популярные_СЭД/ECM-системы (дата обращения: 22.03.2021).
37. Обзор решений класса Security Orchestration, Automation and Response (SOAR) // Anti-Malware URL: https://www.anti-malware.ru/analytics/Market_Analysis/Security-Orchestration-Automation-and-Response-SOAR-Solution-Overview (дата обращения: 16.04.2021).
38. Обзор UBA систем // Anti-Malware URL: https://www.tadviser.ru/index.php/Статья:UBA_(User_Behavior_Analytics,_Анализ_поведения_в_сфере_систем_обеспечения_безопасности (дата обращения: 16.06.2021).
39. Савенков П. А., Трегубов П. С. Поиск поведенческих аномалий в деятельности сотрудников при помощи методов пространственной кластеризации, основанных на плотности //Известия Тульского государственного университета. Технические науки. – 2020. – №. 9. – С. 250-259.
40. Чернокнижный Г. М. АКТУАЛЬНЫЕ ПРОБЛЕМЫ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ И ПУТИ ИХ РЕШЕНИЯ //Инновационные технологии и вопросы обеспечения безопасности реальной экономики. – 2021. – С. 203-215.
41. Выдрина О. А. ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ В ОБЕСПЕЧЕНИИ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ: СИСТЕМЫ АНАЛИЗА ПОВЕДЕНИЯ ПОЛЬЗОВАТЕЛЕЙ //Молодежная научная школа кафедры" Защищенные системы связи". – 2021. – Т. 1. – №. 1. – С. 49-52.
42. Akutota T., Choudhury S. Big data security challenges: An overview and application of user behavior analytics //Int. Res. J. Eng. Technol. – 2017. – Т. 4. – С. 1544-1548.

43. Локтионова Е. А., Рагозина А. В. Особенности применения систем анализа больших данных в деятельности коммерческого банка //Baikal Research Journal. – 2017. – Т. 8. – №. 2. – С. 9.
44. Корганова О. Г., Панфилова И. Е. Модель управления информационными рисками социотехнической системы на основе поведенческих особенностей человека //Сборник научных трудов Новосибирского государственного технического университета. – 2020. – №. 1-2. – С. 89-98.
45. Kotenko I. et al. Attack detection in IoT critical infrastructures: a machine learning and big data processing approach //2019 27th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP). – IEEE, 2019. – С. 340-347.
46. Guo J. et al. Long text generation via adversarial training with leaked information //Proceedings of the AAAI conference on artificial intelligence. – 2018. – Т. 32. – №. 1.
47. Singh A. K., Shashi M. Vectorization of text documents for identifying unifiable news articles //International Journal of Advanced Computer Science and Applications. – 2019. – Т. 10. – №. 7.
48. Pustejovsky J., Stubbs A. Natural Language Annotation for Machine Learning: A guide to corpus-building for applications. – " O'Reilly Media, Inc.", 2012.
49. Фадеева А. А., Сиякина В. В., Салахутдинов Э. Р. ПРИМЕНЕНИЕ ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ ДЛЯ АНАЛИЗА ТОНАЛЬНОСТИ ТЕКСТА //Социально-экономические и технические системы: исследование, проектирование, оптимизация. – 2020. – №. 1. – С. 164-171.
50. When to Use One-Hot Encoding in Deep Learning // Analyticsindiamag URL: https://analyticsindiamag.com/when-to-use-one-hot-encoding-in-deep-learning/ (дата обращения: 25.12.2021).
51. Oxford // oxfordlearnersdictionaries URL: https://www.oxfordlearnersdictionaries.com/wordlist/ (дата обращения: 25.12.2021).

52. Oxford // Google URL: https://developers.google.com/machine-learning/crash-course/embeddings/video-lecture (дата обращения: 1.12.2021).
53. Pennington J., Socher R., Manning C. D. Glove: Global vectors for word representation //Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). – 2014. – С. 1532-1543.
54. Ma L., Zhang Y. Using Word2Vec to process big text data //2015 IEEE International Conference on Big Data (Big Data). – IEEE, 2015. – С. 2895-2897.
55. Sethy A., Ramabhadran B. Bag-of-word normalized n-gram models //Ninth Annual Conference of the International Speech Communication Association. – 2008.
56. Yun-tao Z., Ling G., Yong-cheng W. An improved TF-IDF approach for text classification //Journal of Zhejiang University-Science A. – 2005. – Т. 6. – №. 1. – С. 49-55.
57. Мэн Ц. Анализ методов классификации информации в интернете при решении задач информационного поиска //Моделирование, оптимизация и информационные технологии. – 2016. – №. 2. – С. 19-19.
58. Батура Т. В. Методы автоматической классификации текстов //Программные продукты и системы. – 2017. – Т. 30. – №. 1. – С. 85-89.
59. Гнидко К. О., Макаров С. А., Сабиров Т. Р. Разработка нейросетевого классификатора для формирования обучающих выборок вредоносного графического контента в сети интернет //Региональная информатика и информационная безопасность. – 2019. – С. 242-245.
60. Астапов Р. Л., Мухамадеева Р. М. АВТОМАТИЗИРОВАННАЯ ПРЕДОБРАБОТКА ТЕКСТА ДЛЯ ОПРЕДЕЛЕНИЯ ЭМОЦИОНАЛЬНОЙ ОКРАСКИ ТЕКСТА //Актуальные научные исследования в современном мире. – 2021. – №. 5-2. – С. 19-23.
61. Мартынов В. А., Плотникова Н. П. Нормализация и фильтрация текста для задачи кластеризации //XLVIII Огарёвские чтения. – 2020. – С. 448-452.

62. ПОЛОНСКИЙ Д. А., ФЕДОСОВА А. О. ПРЕДОБРАБОТКА ТЕКСТА ДЛЯ РЕШЕНИЯ NLP (NATURAL LANGUAGE PROCESSING) //МАВЛЮТОВСКИЕ ЧТЕНИЯ. – 2021. – С. 798-802.
63. Kannan S. et al. Preprocessing techniques for text mining //International Journal of Computer Science & Communication Networks. – 2014. – Т. 5. – №. 1. – С. 7-16.
64. Vijayarani S. et al. Preprocessing techniques for text mining-an overview //International Journal of Computer Science & Communication Networks. – 2015. – Т. 5. – №. 1. – С. 7-16.
65. G. Ganesan K., Subotin M. A general supervised approach to segmentation of clinical texts //2014 IEEE International Conference on Big Data (Big Data). – IEEE, 2014. – С. 33-40.
66. НАПРАСНИКОВА М. А. АНАЛИЗ ПРОЦЕССА ПРЕДОБРАБОТКИ ДАННЫХ ИЗ TWITTER ДЛЯ ИСПОЛЬЗОВАНИЯ МЕТОДОВ Data Mining //Председатель оргкомитета-Емельянов Сергей Геннадьевич, д. т. н. – 2016. – С. 257.
67. Савенков П. А., Ивутин А. Н. МЕТОДЫ АНАЛИЗА ЕСТЕСТВЕННОГО ЯЗЫКА В ЗАДАЧАХ ДЕТЕКТИРОВАНИЯ ПОВЕДЕНЧЕСКИХ АНОМАЛИЙ //Известия Тульского государственного университета. Технические науки. – 2022. – №. 3. – С. 358-366.
68. Ивутин А. Н., САВЕНКОВ П. А., ТРЕГУБОВ П. С. ДЕТЕКТИРОВАНИЕ АНОМАЛЬНОГО ПОВЕДЕНИЯ ПОЛЬЗОВАТЕЛЕЙ ПО НАБОРАМ ИХ ТЕКСТОВ.
69. Савенков П. А., Трегубов П. С. СОХРАНЕНИЕ ЦЕЛОСТНОСТИ ДАННЫХ ПРИ ПОМОЩИ АНАЛИЗА АНОМАЛИЙ В ПОВЕДЕНЧЕСКОЙ ДЕЯТЕЛЬНОСТИ ПОЛЬЗОВАТЕЛЕЙ //Известия Тульского государственного университета. Технические науки. – 2021. – №. 2. – С. 45-49.
70. Немчинова Е. А., Плотникова Н. П., Федосин С. А. Подготовка и обработка нормативно-справочной текстовой информации для классификации с

помощью искусственных нейронных сетей //Нелинейный мир. – 2019. – Т. 17. – №. 2. – С. 27-33.

71. Hakim A. A. et al. Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency (TF-IDF) approach //2014 6th international conference on information technology and electrical engineering (ICITEE). – IEEE, 2014. – С. 1-4.
72. Christian H., Agus M. P., Suhartono D. Single document automatic text summarization using term frequency-inverse document frequency (TF-IDF) //ComTech: Computer, Mathematics and Engineering Applications. – 2016. – Т. 7. – №. 4. – С. 285-294.
73. Lilleberg J., Zhu Y., Zhang Y. Support vector machines and word2vec for text classification with semantic features //2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC). – IEEE, 2015. – С. 136-140.
74. McCormick C. Word2vec tutorial-the skip-gram model //Apr-2016.[Online]. Available: http://mccormickml. com/2016/04/19/word2vec-tutorial-the-skip-gram-model. – 2016.
75. González-Carvajal S., Garrido-Merchán E. C. Comparing BERT against traditional machine learning text classification //arXiv preprint arXiv:2005.13012. – 2020.
76. Кочегурова Е. А., Мартынова Ю. А. Особенности непрерывной идентификации пользователей на основе свободных текстов в режиме скрытого мониторинга //Программирование. – 2020. – №. 1. – С. 15-28.
77. Savenkov P. A., Ivutin A. N. Methods and Algorithms of Data and Machine Learning usage in Management Decision Making Support Systems //2019 8th Mediterranean Conference on Embedded Computing (MECO). – IEEE, 2019. – С. 1-4;
78. Савенков П.А. Использование методов и алгоритмов машинного обучения в системах поддержки принятия управленческих решений // Известия

Тульского государственного университета. Технические науки, выпуск 2, 2019. – с.213-218;

79. Ivutin A. N., Savenkov P. A., Veselova A. V. Neural network for analysis of additional authentication behavioral biometrie characteristics //2018 7th Mediterranean Conference on Embedded Computing (MECO). – IEEE, 2018. – С. 1-3;

80. Валиев А. И., Лысенкова С. А. ПРИМЕНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ АВТОМАТИЗАЦИИ ПРОЦЕССА АНАЛИЗА СОДЕРЖАНИЯ ТЕКСТА //Вестник кибернетики. – 2021. – №. 4 (44). – С. 12-15.

81. Ермаков П. Д., Федянин Р. В. Исследование методов машинного обучения в задаче автоматического определения тональности текстов на естественном языке //Новые информационные технологии в автоматизированных системах. – 2015. – №. 18. – С. 600-615.

82. Савенков П.А. Использование методов и алгоритмов машинного обучения в системах поддержки принятия управленческих решений // Известия Тульского государственного университета. Технические науки, выпуск 2, 2019. – с.213-218;

83. Савенков П.А. Использование методов и алгоритмов анализа данных в мобильной UEBA/DSS – системе для решения задач информационной безопасности // Известия Тульского государственного университета. Технические науки, выпуск 12, 2019. – с 585-588;

84. Савенков П. А., Трегубов П. С. Использование методов и алгоритмов анализа данных и машинного обучения в UEBA/DSS для поддержки принятия управленческих решений // Моделирование, оптимизация и информационные технологии, выпуск 8(1), 2020;

85.. П.А. Савенков, П.С. Трегубов Сохранение целостности данных при помощи анализа аномалий в поведенческой деятельности пользователей // Известия Тульского государственного университета. Технические науки, выпуск 2, 2021. – с.45-49;

86. Савенков П.А. Сравнение методов кластеризации DBSCAN и модифицированного WrapDBSCAN для поиска аномальных перемещений пользователей в мобильной UBA системе // Моделирование, оптимизация и информационные технологии, выпуск 9(4), 2021;
87. И.Н. Набродова, П.А. Савенков, П.С. Трегубов Проблема избыточности и плотностная нормализация координат геолокации как этап подготовки данных к детектированию аномалий местоположения плотностным методом машинного обучения WrapDBSCAN // Известия Тульского государственного университета. Технические науки, выпуск 9, 2020. – с.119-127;
88. П.А. Савенков, П.С. Трегубов Поиск поведенческих аномалий в деятельности сотрудников при помощи методов пространственной кластеризации, основанных на плотности // Известия Тульского государственного университета. Технические науки, выпуск 9, 2020. – с 250-259;
89. Savenkov P. A., Ivutin A. N. Organizations Data Integrity Providing through Employee Behavioral Analysis Algorithms //2020 9th Mediterranean Conference on Embedded Computing (MECO). – IEEE, 2020. – С. 1-3;
90. Savenkov P. A., Ivutin A. N. Methods of Machine Learning in System Abnormal Behavior Detection //International Conference on Swarm Intelligence. – Springer, Cham, 2020. – С. 495-505;
91. Savenkov P. A., Ivutin A. N. DBScan and WrapDBScan methods applying for intellectual variance analysis in employee's moving //Procedia Computer Science. – 2021. – Т. 186. – С. 177-184;
92. Savenkov P.A., Ivutin A. N. Heterogeneous text data parallel processing to behavioral anomalies search using machine learning methods and algorithms //2021 10th Mediterranean Conference on Embedded Computing (MECO). – IEEE, 2021. – С. 1-3;

93. Obradovic N., Kelec A., Dujlovic I. Performance analysis on Android SQLite database //2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH). – IEEE, 2019. – С. 1-4.
94. Lock A. ASP. NET core in Action. – Simon and Schuster, 2021.
95. Nagel C. Professional C# 7 and. Net Core 2.0. – John Wiley & Sons, 2018.
96. Joshi B. ASP. NET Core Web API //Beginning Database Programming Using ASP. NET Core 3. – Apress, Berkeley, CA, 2019. – С. 175-226.
97. Прохоренок Н., Дронов В. Python 3. Самое необходимое, 2-е изд. – БХВ-Петербург, 2019.
98. Якимчик А. И. Jupyter Notebook: система интерактивных научных вычислений //Геофизический журнал. – 2019.
99. Schmitt X. et al. A replicable comparison study of NER software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate //2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS). – IEEE, 2019. – С. 338-343.
100. Lemenkova P. Python Libraries Matplotlib, Seaborn and Pandas for Visualization Geo-spatial Datasets Generated by QGIS //Analele stiintifice ale Universitatii" Alexandru Ioan Cuza" din Iasi-seria Geografie. – 2020. – Т. 64. – №. 1. – С. 13-32.
101. Bourhis P., Reutter J. L., Vrgoč D. JSON: Data model and query languages //Information Systems. – 2020. – Т. 89. – С. 101478.
102. González-Pérez A. et al. Using mobile devices as scientific measurement instruments: reliable android task scheduling //Pervasive and Mobile Computing. – 2022. – Т. 81. – С. 101550.
103. Юдин Ю. В., Майсурадзе М. В., Водолазский Ф. В. Организация и математическое планирование эксперимента: учебное пособие. – 2018.
104. Старолетов С. М., Ануреев И. С. На пути к модульному тестированию событийно-управляемых требований //Вычислительные технологии. – 2022. – Т. 27. – №. 1. – С. 88-100.

105. ГОСТ Р 50739-95 Средства вычислительной техники. Защита от несанкционированного доступа к информации. Общие технические требования.
106. ГОСТ 28195-89. Оценка качества программных средств. Общие положения.
107. ГОСТ 28195-89. Оценка качества программных средств. Общие положения.
108. Свидетельство о государственной регистрации программы для ЭВМ №2020612542 «Автоматизированная система анализа и контроля деятельности сотрудников (Клиент)». Номер и дата поступления заявки: №2020611000, 03.02.2020. Дата государственной регистрации в Реестре программ для ЭВМ – 26.02.2020 г;
109. Свидетельство о государственной регистрации программы для ЭВМ №2021613363 «DeepViewer Сервер панели администрирования». Номер и дата поступления заявки: 2021612125, 15.02.2021. Дата государственной регистрации в Реестре программ для ЭВМ – 05.03.2021 г;
110. Свидетельство о государственной регистрации программы для ЭВМ №2021613188 «DeepViewer Панель администрирования». Номер и дата поступления заявки: 2021612094, 15.02.2021. Дата государственной регистрации в Реестре программ для ЭВМ – 03.03.2021 г;
111. Свидетельство о государственной регистрации программы для ЭВМ №2021613047 «DeepViewer Мобильный агент». Номер и дата поступления заявки: 2021611628, 10.02.2021. Дата государственной регистрации в Реестре программ для ЭВМ – 01.03.2021 г.;
112. Свидетельство о государственной регистрации программы для ЭВМ №2021612980 «DeepViewer Сервер мобильного агента». Номер и дата поступления заявки: 2021611602, 10.02.2021. Дата государственной регистрации в Реестре программ для ЭВМ – 26.02.2021 г.;
113. Свидетельство о государственной регистрации программы для ЭВМ №2021614978 «Программный комплекс анализа поведенческих

биометрических характеристик пользователей». Номер и дата поступления заявки: 2021613888, 24.03.2021. Дата государственной регистрации в Реестре программ для ЭВМ – 01.04.2021 г.;

114. Свидетельство о государственной регистрации программы для ЭВМ №2021615006 «Программный комплекс сбора поведенческих биометрических характеристик пользователей». Номер и дата поступления заявки: 2021613889, 24.03.2021. Дата государственной регистрации в Реестре программ для ЭВМ – 01.04.2021 г.;
115. Свидетельство о государственной регистрации программы для ЭВМ №2021661392 «Детектирование аномалий местоположения пользователя». Номер и дата поступления заявки: 2021660224, 28.06.2021. Дата государственной регистрации в Реестре программ для ЭВМ – 09.07.2021 г.;
116. Свидетельство о государственной регистрации программы для ЭВМ №2021661489 «Детектирование аномального поведения пользователей по наборам их текстов». Номер и дата поступления заявки: 2021660269 28.06.2021. Дата государственной регистрации в Реестре программ для ЭВМ – 12.07.2021 г.
117. Свидетельство о государственной регистрации программы для ЭВМ №2021661489 «Скрипт анализа тональности текстовых наборов пользовательских данных» Номер и дата поступления заявки: 2022618739 13.05.2022. Дата государственной регистрации в Реестре программ для ЭВМ – 20.05.2022г.

ПРИЛОЖЕНИЯ

Приложение 1. Акты об использовании результатов диссертационной работы

ООО «Кадастр Экспресс»
Адрес 300041, Тульская область, город Тула, проспект Ленина, дом 35, офис 402
Тел: +7 (4872) 587-583
ОГРН 1177154024080/ ИНН 7107122999/ КПП 710701001

АКТ

о внедрении программного продукта

Настоящий акт о внедрении свидетельствует о том, что программный продукт, разработанный Савенковым Павлом Анатольевичем, имеющий следующие свидетельства о регистрации программы ЭВМ:

1. Свидетельство о государственной регистрации программы для ЭВМ №2021613047 «DeepViewer Мобильный агент» Номер и дата поступления заявки: №2021611628, 10.02.2021. Дата государственной регистрации в Реестре программ для ЭВМ – 01.03.2021 г.
2. Свидетельство о государственной регистрации программы для ЭВМ №2021612980 «DeepViewer Сервер мобильного агента» Номер и дата поступления заявки: №2021611602, 10.02.2021. Дата государственной регистрации в Реестре программ для ЭВМ – 26.02.2021 г.
3. Свидетельство о государственной регистрации программы для ЭВМ №2021613188 «DeepViewer Панель администрирования» Номер и дата поступления заявки: №2021612094, 15.02.2021. Дата государственной регистрации в Реестре программ для ЭВМ – 03.03.2021 г.
4. Свидетельство о государственной регистрации программы для ЭВМ №2021613363 «DeepViewer Сервер панели администрирования» Номер и дата поступления заявки: №2021612125, 15.02.2021. Дата государственной регистрации в Реестре программ для ЭВМ – 05.03.2021 г.

внедрен в эксплуатацию в ООО «Кадастр Экспресс».

В ходе эксплуатации программного продукта было подтверждено, что он обладает всеми заявленными возможностями.

Генеральный директор …….. _____ Е.Я. Соловьева

М.П.

Рисунок П.1.1. Акт об использовании результатов диссертационной работы в ООО «Кадастр Экспресс»

ООО «Гуд Фуд»

Юр адрес.300036, г. Тула ул. Привокзальная 25

ИНН 7104510601/КПП 710401001

р/сч 40702810800020000721 в

Филиал Центральный ПАО Банка

«ФК ОТКРЫТИЕ»

Тел.702-172

АКТ

о внедрении программного продукта

Настоящий акт о внедрении свидетельствует о том, что программный продукт, разработанный Савенковым Павлом Анатольевичем, имеющий следующие свидетельства о регистрации программы ЭВМ:

1. Свидетельство о государственной регистрации программы для ЭВМ №2021613047 «DeepViewer Мобильный агент» Номер и дата поступления заявки: №2021611628, 10.02.2021. Дата государственной регистрации в Реестре программ для ЭВМ – 01.03.2021 г.
2. Свидетельство о государственной регистрации программы для ЭВМ №2021612980 «DeepViewer Сервер мобильного агента» Номер и дата поступления заявки: №2021611602, 10.02.2021. Дата государственной регистрации в Реестре программ для ЭВМ – 26.02.2021 г.
3. Свидетельство о государственной регистрации программы для ЭВМ №2021613188 «DeepViewer Панель администрирования» Номер и дата поступления заявки: №2021612094, 15.02.2021. Дата государственной регистрации в Реестре программ для ЭВМ – 03.03.2021 г.
4. Свидетельство о государственной регистрации программы для ЭВМ №2021613363 «DeepViewer Сервер панели администрирования» Номер и дата поступления заявки: №2021612125, 15.02.2021. Дата государственной регистрации в Реестре программ для ЭВМ – 05.03.2021 г.

внедрен в эксплуатацию в ООО « Гуд Фуд».

В ходе эксплуатации программного продукта было подтверждено, что он обладает всеми заявленными возможностями.

директор ООО «Гуд Фуд» _____ А.Е. Иванушкин

М.П.

Рисунок П.1.2. Акт об использовании результатов диссертационной работы в ООО «Гуд Фуд»

Приложение 2. Свидетельства о государственной регистрации программы для ЭВМ

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2020612542

Автоматизированная система анализа и контроля деятельности сотрудников (Клиент)

Правообладатель: ***Савенков Павел Анатольевич (RU)***

Автор: ***Савенков Павел Анатольевич (RU)***

Заявка № **2020611000**
Дата поступления **03 февраля 2020 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***26 февраля 2020 г.***

Руководитель Федеральной службы по интеллектуальной собственности

Г.П. Ивлиев

Рисунок П 2.1. Свидетельство о государственной регистрации программы для ЭВМ №2020612542

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2021613047

DeepViewer Мобильный агент

Правообладатели: ***Савенков Павел Анатольевич (RU), Трегубов Павел Сергеевич (RU)***

Авторы: ***Савенков Павел Анатольевич (RU), Трегубов Павел Сергеевич (RU)***

Заявка № **2021611628**

Дата поступления **10 февраля 2021 г.**

Дата государственной регистрации в Реестре программ для ЭВМ ***01 марта 2021 г.***

Руководитель Федеральной службы по интеллектуальной собственности

Г.П. Ивлиев

Рисунок П 2.2. Свидетельство о государственной регистрации программы для ЭВМ №2021613047

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2021613188

DeepViewer Панель администрирования

Правообладатели: ***Савенков Павел Анатольевич (RU), Трегубов Павел Сергеевич (RU)***

Авторы: ***Савенков Павел Анатольевич (RU), Трегубов Павел Сергеевич (RU)***

Заявка № **2021612094**

Дата поступления **15 февраля 2021 г.**

Дата государственной регистрации в Реестре программ для ЭВМ ***03 марта 2021 г.***

Руководитель Федеральной службы по интеллектуальной собственности

Г.П. Ивлиев

Рисунок П 2.3. Свидетельство о государственной регистрации программы для ЭВМ №2021613188

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2021612980

DeepViewer Сервер мобильного агента

Правообладатели: ***Савенков Павел Анатольевич (RU), Трегубов Павел Сергеевич (RU)***

Авторы: ***Савенков Павел Анатольевич (RU), Трегубов Павел Сергеевич (RU)***

Заявка № **2021611602**

Дата поступления **10 февраля 2021 г.**

Дата государственной регистрации в Реестре программ для ЭВМ ***26 февраля 2021 г.***

Руководитель Федеральной службы по интеллектуальной собственности

Г.П. Ивлиев

Рисунок П 2.4. Свидетельство о государственной регистрации программы для ЭВМ №2021612980

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2021613363

DeepViewer Сервер панели администрирования

Правообладатели: ***Савенков Павел Анатольевич (RU), Трегубов Павел Сергеевич (RU)***

Авторы: ***Савенков Павел Анатольевич (RU), Трегубов Павел Сергеевич (RU)***

Заявка № **2021612125**
Дата поступления **15 февраля 2021 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***05 марта 2021 г.***

Руководитель Федеральной службы
по интеллектуальной собственности

Г.П. Ивлиев

Рисунок П 2.5. Свидетельство о государственной регистрации программы для ЭВМ №2021613363

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2021661392

«Детектирование аномалий местоположения пользователя»

Правообладатель: *Федеральное государственное бюджетное образовательное учреждение высшего образования «Тульский государственный университет» (ТулГУ) (RU)*

Авторы: *Ивутин Алексей Николаевич (RU), Савенков Павел Анатольевич (RU), Трегубов Павел Сергеевич (RU)*

Заявка № **2021660224**
Дата поступления **28 июня 2021 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***09 июля 2021 г.***

Руководитель Федеральной службы по интеллектуальной собственности

ДОКУМЕНТ ПОДПИСАН ЭЛЕКТРОННОЙ ПОДПИСЬЮ
Сертификат 0x02A5CFBCCCB1ACF59A40A2F080922E9A116
Владелец Ивлиев Григорий Петрович
Действителен с 15.01.2021 по 15.01.2035

Г.П. Ивлиев

Рисунок П 2.6. Свидетельство о государственной регистрации программы для ЭВМ №2021661392


РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2021661489

«Детектирование аномального поведения пользователей по наборам их текстов»

Правообладатель: *Федеральное государственное бюджетное образовательное учреждение высшего образования «Тульский государственный университет» (ТулГУ) (RU)*

Авторы: *Ивутин Алексей Николаевич (RU), Савенков Павел Анатольевич (RU), Трегубов Павел Сергеевич (RU)*

Заявка № **2021660269**
Дата поступления **28 июня 2021 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***12 июля 2021 г.***

Руководитель Федеральной службы по интеллектуальной собственности

ДОКУМЕНТ ПОДПИСАН ЭЛЕКТРОННОЙ ПОДПИСЬЮ
Сертификат 0x02A5CFBCCCB1ACF59A4CA2F08092E9A116
Владелец Ивлиев Григорий Петрович
Действителен с 15.01.2021 по 15.01.2035

Г.П. Ивлиев

Рисунок П 2.7. Свидетельство о государственной регистрации программы для ЭВМ №2021661489

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2021614978

«Программный комплекс анализа поведенческих биометрических характеристик пользователей»

Правообладатель: *Федеральное государственное бюджетное образовательное учреждение высшего образования «Тульский государственный университет» (ТулГУ) (RU)*

Авторы: *Ивутин Алексей Николаевич (RU), Савенков Павел Анатольевич (RU), Трегубов Павел Сергеевич (RU)*

Заявка № 2021613888
Дата поступления 24 марта 2021 г.
Дата государственной регистрации
в Реестре программ для ЭВМ *01 апреля 2021 г.*

Руководитель Федеральной службы по интеллектуальной собственности

Г.П. Ивлиев

Рисунок П 2.8. Свидетельство о государственной регистрации программы для ЭВМ №2021614978


РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2021615006

«Программный комплекс сбора поведенческих биометрических характеристик пользователей»

Правообладатель: *Федеральное государственное бюджетное образовательное учреждение высшего образования «Тульский государственный университет» (ТулГУ) (RU)*

Авторы: *Ивутин Алексей Николаевич (RU), Савенков Павел Анатольевич (RU), Трегубов Павел Сергеевич (RU)*

Заявка № 2021613889

Дата поступления 24 марта 2021 г.

Дата государственной регистрации в Реестре программ для ЭВМ *01 апреля 2021 г.*

Руководитель Федеральной службы по интеллектуальной собственности

Г.П. Ивлиев

Рисунок П 2.9. Свидетельство о государственной регистрации программы для ЭВМ № 2021615006

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2022619336

«Скрипт анализа тональности текстовых наборов пользовательских данных»

Правообладатель: ***Федеральное государственное бюджетное образовательное учреждение высшего образования «Тульский государственный университет» (ТулГУ) (RU)***

Авторы: ***Ивутин Алексей Николаевич (RU), Савенков Павел Анатольевич (RU)***

Заявка № **2022618739**
Дата поступления **13 мая 2022 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***20 мая 2022 г.***

Руководитель Федеральной службы по интеллектуальной собственности

ДОКУМЕНТ ПОДПИСАН ЭЛЕКТРОННОЙ ПОДПИСЬЮ
Сертификат 0x02A5CFBC0B1ACF59A40A2F08092E9A116
Владелец Ивлиев Григорий Петрович
Действителен с 15.01.2021 по 15.01.2035

Г.П. Ивлиев

Рисунок П 2.10. Свидетельство о государственной регистрации программы для ЭВМ №2022619336