DOI 10.36622/1682-7813.2024.27.2.002
УДК 519.766.4:004.032.26

ПРОГРАММНО-АЛГОРИТМИЧЕСКИЙ КОМПЛЕКС ПОИСКА ДЕСТРУКТИВНОЙ ИНФОРМАЦИИ В СИСТЕМАХ МГНОВЕННОГО ОБМЕНА СООБЩЕНИЯМИ

В.А. Минаев, А.В. Симонов

Описываются проблемы и этапы создания программно-алгоритмического комплекса обработки текстов в системах мгновенного обмена сообщений с целью выявления деструктивной информации. Осуществлен обзор актуальных научных работ в этой области. Раскрывается понятие и содержание деструктивной информации. Обоснован выбор решения задачи классификации текстов. Представлена процедура поиска и формирования обучающего контента. Представлена архитектура программного комплекса, даны функциональные характеристики его модулей. Описана модель данных, в соответствии с которой разработана связь сущностей в СУБД, а также выбор языка программирования и используемых библиотек. Обсуждается подход к унификации системы поиска деструктивной информации в социальных медиа, базирующийся на решении задачи классификации текстовых массивов любого размера.

Ключевые слова: программно-алгоритмическая реализация, социальные медиа, мессенджер, деструктивный контент, нейронная сеть, экстремизм, трансформер BERT.

Введение

Распространение деструктивного и опасного контента в Сети – современная проблема, все более угрожающая обществу, государству и его гражданам. Об этом со всей пронзительностью свидетельствует трагедия в Крокус Сити Холле, для реализации которой исполнители и организаторы теракта для своих транзакций до и после своего черного дела использовали мессенджер Telegram.

Под деструктивным контентом будем понимать информацию (текстовую, аудио, визуальную) с призывами к совершению массовых убийств и самоубийств, экстремистские материалы, признанные таковыми по решению суда, а также пропагандирующие противоправные методы и способы поведения и действий. Информация относится к противоправной, распространение которой запрещено законодательством Российской Федерации.

Важную роль в пропаганде противоправных действий играют экстремистские материалы (ЭМ) по следующим направлениям:

1. Призывы к совершению террористических актов.
2. Пропаганда радикального национализма, шовинизма.
3. Распространение материалов нацистского характера, а также оправдывающих военные преступления времен второй мировой войны.

Наиболее распространенным типом деструктивного контента стоит считать текстовый по причине простоты передачи, генерации и распространения.

Рассматриваемая проблема в связи со своей актуальностью включена в современную Доктрину информационной безопасности Российской Федерации в виде одной из основных угроз. Поэтому важнейшим направлением обеспечения информационной безопасности российского государства является защита от распространения информации, пропагандирующей экстремистскую идеологию.

Борьбой и выявлением противоправного контента занимаются как госорганы (ФСБ, МВД, Роскомнадзор), так и общественные и региональные организации (Лига безопасного интернета, Национальный центр информационного противодействия терроризму и экстремизму в образовательной среде и сети Интернет).

Поиск и выявление ЭМ может осуществляться ручными, автоматизированными и гибридными методами. При этом эффективными автоматизированными средствами поиска и

анализа противоправного контента обладают только государственные органы. Общественным организациям приходится либо самостоятельно разрабатывать компьютерные системы, которые помогают в принятии решений, либо ориентироваться на волонтеров и кибер-дружины, которые осуществляют поиск вручную.

Однако ручной анализ контента на наличие деструктивной составляющей является в настоящее время малоэффективным в связи с существенным ростом публикационной активности в социальных медиа.

Возникает необходимость разработки и реализации комплексных систем оценки деструктивности текстовых массивов, которые наполняют социальные медиа (СМ).

Целью статьи является создание программно-алгоритмического комплекса оценки деструктивности текстовых массивов (ПАК ОДТМ) с применением спектра передовых методов обработки информации произвольного объема, включая контент СМ Telegram.

Обзор существующих решений

Для достижения поставленной цели ПАК должен обеспечивать функционал:

- *комплексной оценки* каналов с расчетом потенциала информационно-психологического воздействия на их пользователей;
- *парсинга* – формирования систематизированного текстового контента из групп, чатов социальных сетей и каналов в СМ Telegram;
- *определения степени деструктивности* систематизированного текстового контента.

Проведенный анализ показал, что идентификация текстов деструктивной направленности сегодня решается с использованием трех подходов:

- поиска по ключевым словам и фразам;
- тематического моделирования;
- классификации текстов.

Дадим краткий обзор указанных подходов.

В работе [1] описана реализация программного комплекса NetEpidemic, который предназначен для мониторинга СМ «ВКонтакте». Отнесение анализируемого контента к деструктивному производится на основе наличия в тексте так называемых противоправных лемм, словарь которых готовится специалистами вручную. Недостатком такого подхода является высокий уровень ошибок первого и второго рода по причине отнесения к деструктивному тексту только при наличии в нем одного слова из заранее заданного словаря.

Работа [2] посвящена поиску деструктивной информации в сети Интернет. Под деструктивным в работе понимается любой контент, содержащий нецензурную лексику.

Работа [3] направлена на мониторинг наркоситуации в социальных сетях. При мониторинге происходит поиск и идентификация сообщений, относящихся к противоправным, по набору словоформ с учетом их возможной обфускации, в частности – намеренного допущения ошибок в словоформах.

Недостатком поиска по ключевым словам при идентификации контента деструктивного содержания является невозможность применения данного метода при поиске ЭМ по причине отсутствия необходимых словарных маркеров в них, невозможности однозначной идентификации ЭМ по одному слову или словоформе.

Работа [4] описывает методы обнаружения экстремистской информации в Интернет с использованием машинного обучения. Для этого производится предварительная обработка документов из набора данных KavkazChat, который создан на основе собранной информации с форумов, посвященных вопросам Северного Кавказа. Из указанных документов производится выделение ключевых слов, которые далее используются при осуществлении информационного поиска. Недостатком методов тематического моделирования, в рамках которого подготовлена статья, является отсутствие возможности учета семантики документа или предложения в связи с тем, что не учитывается порядок слов в текстах.

Зарубежное исследование [5] посвящено обнаружению экстремизма в тексте с

акцентированием на религиозной и политической составляющей.

Обучающая выборка составлена из открытых наборов публикаций в СМ Twitter от пользователей – сторонников ИГИЛ и неофашистских организаций США. Данные, поступающие в классификатор, проходили этапы предварительной обработки и векторизации с использованием метода TF-IDF [6]. Использовался составной классификатор, который сначала определял отнесение текста к деструктивному, а далее определялось конкретное направление (неонацизм, терроризм). На тестовой выборке показано, что наилучшие результаты показал алгоритм случайного леса.

Работа [7] представляет описание технологии обнаружения текстов экстремистской направленности, относящейся к пропаганде джихада на Индийском полуострове. Датасет формировался группой экспертов. В качестве векторизатора использовалась искусственная нейронная сеть (ИНС) LSTM, классификация осуществлялась однослойной сетью прямого распространения (FFNN).

Из указанных трех подходов предпочтительным для решения поставленной задачи является классификация текста. Классификация текста с использованием ИНС позволяет учитывать контекст содержимого, прослеживать зависимости и расположение слов в тексте, а также однозначно относить анализируемый текст к искомой тематике.

В то же время, к недостаткам классификации текста относится использование не самых лучших практик, отсутствие в свободном доступе обученных моделей на экстремистских текстах и соответствующих обучающих выборок.

Поэтому для создания ПАК ОДТМ необходимо собрать обучающую выборку с экстремистскими материалами, обучить модель классификатора, разработать ПАК, который в автоматизированном режиме будет собирать и оценивать на деструктивность текстовый контент в СМ.

В качестве СМ для апробации модели выбран Telegram, как наиболее популярный в настоящее время у молодежи и населения среднего возраста ресурс.

Архитектура программно-алгоритмического комплекса

Создание обучающей выборки для классификации связано с формированием массива текстов, состоящий из нескольких корпусов:

1. *Корпус экстремистского контента (ЭК)* состоит из материалов из Федерального списка экстремистских материалов, содержащий тексты трех направлений: радикального ислама (террористический), трудов лидеров нацисткой партии и третьего рейха (нацистский), антисемитских ксенофобских книг и материалов (антисемитский). Корпус формировался в сети Darknet, на форумах и площадках, доступ к которым скрыт из общих поисковых систем. Объем корпуса составил более 31 млн символов.

2. *Корпус нейтрального контента.* Состоит из данных пользовательских сообщений [8] и новостных сводок, также общим объемом 31 млн символов.

3. *Шумовой корпус.* Как показали исследования [9] для повышения помехоустойчивости и снижения ошибок первого и второго рода при решении задачи обнаружения деструктивного контента с использованием решения задачи классификации, целесообразно использование в обучающей выборки шумового корпуса, который содержать в себе сущности, по количеству приближенные к целевому корпусу, но отличающиеся от них семантическим содержанием. Корпус собран из исторических учебников и книг еврейской национальной литературы, текстов об исламской культуре. Общий объем составил около 39 млн символов.

Согласно работе авторов [9] наиболее высокие результаты показывает бинарная классификация. В связи с этим нейтральный и шумовой контент объединяются в один класс, которому противопоставляется корпус ЭК.

Для подготовки обучающей выборки проведена предварительная обработка текста: заменена буква «ё» на «е», ссылки, начинающиеся с «http://» заменены на слово «сайт», начинающиеся со знака «@» заменены на слово «никнейм». Лемматизация и стемминг не проводился.

В качестве преобразователя текстовых массивов в вектор выбрана глубокая искусственная нейронная сеть (ГИНС) ruBERT [10], предварительно обученная на корпусе русскоязычных текстов.

С ГИНС BERT проведена доменная адаптация – дообучение на целевом корпусе с целью более качественного распознавания исследуемых текстов.

В качестве головы классификатора использовалась нейронная сеть прямого распространения входной размерностью.

Подробное описание реализации указанного классификатора представлено в [11].

После реализации основного функционального модуля разработана программная помодульная архитектура ПАК ОДТМ. Диаграмма модулей и компонентов представлена на рис. 1.

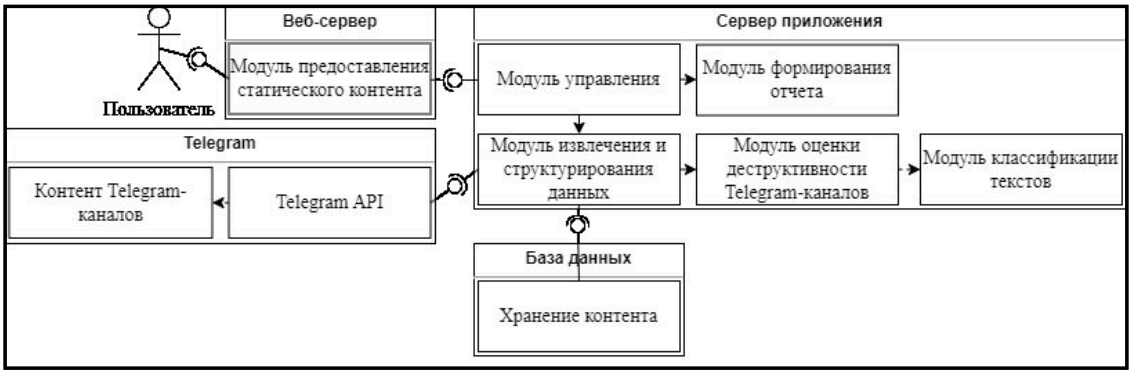

Рис. 1. Взаимодействие модулей ПАК ОДКМ

В качестве реализации выбрана трехзвенная архитектура приложения с веб-интерфейсом. Это позволило сократить зависимость от используемых платформ и не потребовало разработок операционной системы под каждый тип.

На веб-сервере располагается модуль предоставления статического контента, который отвечает за формирование UI пользователя, отображение и возможность взаимодействия с сервером приложения. Пользователь попадает в интерфейс только после прохождения процедуры аутентификации и авторизации в телеграмм-боте со своим номером телефона. Номер телефона, одноразовые пароли и ключи от Telegram API подлежат маскированию в логах.

Сервер приложений содержит в себе несколько модулей.

Модуль управления – отвечает за генерацию контента для интерфейса пользователя, подготовку сканируемых каналов, взаимодействие всех модулей и веб-интерфейса.

Модуль извлечения и структурирования данных, после соответствующей команды от модуля управления, производит запрос для сбора текстового контента с выбранных каналов. Производит предварительную обработку текстов, извлекает идентификаторы каналов и публикаций, отправляет данные в модуль оценки деструктивности каналов.

Модуль оценки деструктивности каналов направляет в модуль классификации все имеющиеся в исследуемом канале текстовые публикации, формируя на выходе оценку степени деструктивности канала, которая рассчитывается по формулам:

$$F_{ext} = V_{extmess} \cdot Prev_{extmess}, \quad (1)$$

$$V_{extmess} = P_{ext} \cdot L_{mess} \quad (2)$$

$$Prev_{extmess} = N_{view} + N_{rep} \quad (3)$$

где $V_{extmess}$ – уровень противоправности публикации, являющийся произведением количества символов в публикации L_{mess} на вероятность отнесения публикации к экстремистской P_{ext};

$Prev_{extmess}$ – распространенность публикации, которая является суммой

количества ее просмотров N_{view} на количество репостов N_{rep}.

Модуль классификации текстов – это ГИНС BERT, на вход которой поступают текстовые данные, а на выходе формируется значение P_{ext}.

За хранение результатов сканирования отвечает *модуль хранения* контента в отдельно размещаемой базе данных.

Модуль формирования отчета отвечает за создание excel файла с выгружаемыми постами и оценками работы ПАК ОДТМ, производит парсинг проанализированных данных и сохранение их в файл вывода.

Пользовательский путь выглядит следующим образом.

1. Пользователь аутентифицируется в ПАК ОДТМ.
2. Выбирает Telegram-каналы для сканирования.
3. Запускает сканирование.
4. Получает оценку деструктивности каналов и сообщений.

Модель данных

Построение архитектуры ПАК ОДТМ показало, что для его эффективной работы необходимо реализовать унифицированное хранение данных.

Разработанная модель данных в виде ER-диаграммы представлена на рис. 2. Модель данных состоит из трех сущностей:

1. Пользователи (Users).
2. Каналы (Channels).
3. Сообщения (Messages).

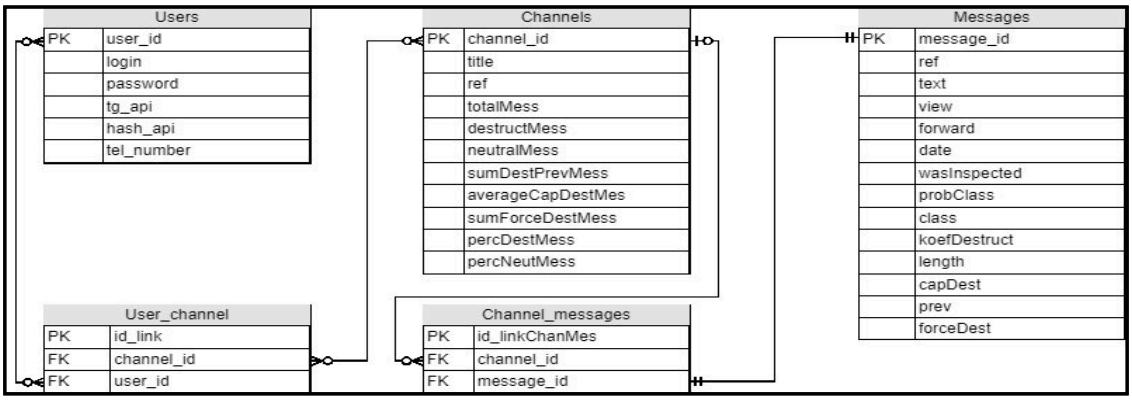

Рис. 2. ER-диаграмма модели данных ПАК ОДТМ

В сущности Users хранятся идентификационные и аутентификационные данные пользователей, с помощью которых происходит подключение к ПАК ОДКМ.

Таблица Channels хранит основные данные о канале: наименование (title), ссылка на канал (ref), общее количество сообщений (totalMess). После сканирования заполняются поля destructMess (деструктивные сообщения), neutralMess (нейтральные сообщения).

После анализа всех сообщений в канале рассчитывается суммарная распространенность деструктивных публикаций (sumDestPrevMess), средний уровень деструктивности канала (averageCapDestMess), суммарная степень деструктивности (sumForceDestMess) и др.

Таблица Messages хранит информацию о сущности сообщений. Для каждого из них рассчитывается уровень, распространенность и степень деструктивности по формулам (1) – (3).

User_channel и Channel_message – таблицы, выполняющие роль связки между сущностями. Использование связующих таблиц позволяет строить валидные SQL запросы в базу данных для построения и выгрузки отчетов.

Реализация системы

В целях повышения горизонтальной и вертикальной масштабируемости, простоты расширения и внесения изменений использована технология контейнеризации Docker. Контейнеры используются в состоянии Multistage – сначала создается и собирается образ со всеми зависимыми и необходимыми библиотеками, происходит компиляция исходного кода, а далее

скомпилированный код «перекладывается» в отдельный чистый контейнер операционной системы (ОС) Alpine Linux. Это позволяет снизить размеры контейнеров, их программные уязвимости, которые имеют необходимые для работы библиотеки. Для управления несколькими контейнерами выбрана система Docker Compose.

В качестве хостовой ОС для размещения контейнеров выбран Debian 12 – ОС с открытым исходным кодом.

ПАК ОДТМ делится на 3 инфраструктурные компоненты: веб-сервер, сервер приложений, база данных.

Веб-сервер реализуется на ПО Nginx, избранном потому, что показывает лучшую производительность в сравнении с другими популярными веб-серверами. С точки зрения удобства и дружелюбности настроек Nginx также выигрывает: конфигурационные файлы просты в понимании и настройке, в наличии полная техническая документация.

Для выбора системы управления базами данных (СУБД) необходимо определиться с типом баз данных (БД). В настоящее время выделяются два типа баз: реляционные (РБД) и нереляционные (НРБД).

В плане скорости записи данных и скорости выполнения запросов и хранимых процедур выигрывают РБД. Сущность Messages больше подходит для хранения в НРБД, но для одной таблицы создавать отдельную СУБД нецелесообразно. С учетом указанных соображений выбрана СУБД типа РБД с открытым исходным кодом PostgreSQL.

Реализация логической части приложения связана с использованием языка программирования Python, удобного и потому, что ГИНС Bert создавалась и обучалась с использованием библиотеки Huggingface.

Для взаимодействия с Telegram API используется библиотека telethon, что позволило не создавать собственные запросы для вызова API.

Работа с данными, составление отчетов связаны с применением библиотек Pandas и NumPy.

При предварительной обработке текстов использовалась библиотека Re, позволяющая программировать необходимые регулярные выражения.

Для генерации статического контента и связи с фронт частью использовалось ПО Figma.

Заключение

Разработанный ПАК ОДТМ является прототипом развивающихся систем оценки деструктивности публикаций. Одним из ограничений прототипа является то, что максимальный размер текста, подвергаемый анализу и классификации, ограничен 4096 символами (ограничение публикации в Telegram).

Сегодня же существует необходимость в системе, которая осуществляет анализ текстового контента и на других типах ресурсов СМ, где отсутствует указанное ограничение. Эта необходимость связана с тем, что деструктивная составляющая может содержаться только в некоторых фрагментах текста. В связи с этим предлагается подход к поиску деструктивной информации в тексте любого размера.

Суть его заключается в том, чтобы подвергать большие тексты фрагментированному анализу размером в 4096 символов. Далее дается суммарная оценка деструктивности всего текста, которая складывается из суммы конкретных частей текста, которые идентифицированы как деструктивные. Это позволяет более точно производить экспертизу фрагментов контента.

Алгоритм оценки деструктивности текстовых массивов произвольного размера в СМ представлен на рис. 3.

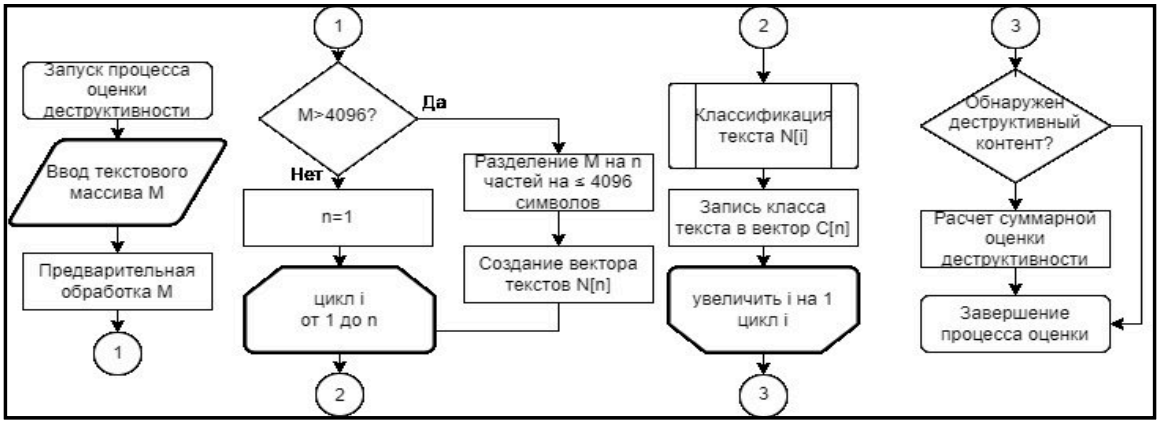

Рис. 3. Алгоритм анализа деструктивности текстового контента произвольного размера

Обсуждение и выводы

В статье представлена архитектура ПАК ОДТМ, целью которого является идентификация деструктивной информации в СМ Telegram с использованием методов классификации, основанных на ГИНС BERT.

Приведена программная архитектура комплекса, разбитая на функциональные модули. Дано описание каждого модуля. Разработана модель данных, на основе которой организовано хранение данных ПАК ОДТМ. Обоснован выбор каждой системной компоненты.

Описан подход к анализу данных на деструктивность любого размера.

Результаты работы ПАК ОДТМ представлены в [11].

Список литературы

1. Остапенко А. Г. и др. Программное обеспечение для мониторинга процессов восприятия и распространения деструктивных контентов в социальных сетях //Информация и безопасность. 2019. Т. 22. № 2. С. 188-205.

2. Давидюк Н. В., Гостюнина В. А., Байдулова Д. Р. Интеллектуальный алгоритм идентификации деструктивной информации в тексте // Вестник Астраханского государственного технического университета. Серия: Управление, вычислительная техника и информатика. 2019. № 2. С. 29-39.

3. Давыдова Ю. В. Модель ошибок для нечеткого текстового поиска в задаче мониторинга виртуальных социальных сетей для обеспечения информационно-психологической безопасности личности // Современные информационные технологии и ИТ-образование. 2017. Т. 13. № 3. С. 72-82.

4. Методы машинного обучения для задачи обнаружения и мониторинга экстремистской информации в сети Интернет / И. В. Машечкин, М. И. Петровский, Д. В. Царев, М. Н. Чикунов // Программирование. 2019. № 3. С. 18-37.

5. Berhoum A. et al. An Intelligent Approach Based on Cleaning up of Inutile Contents for Extremism Detection and Classification in Social Networks //ACM Transactions on Asian and Low-Resource Language Information Processing. 2023. Т. 22. №. 5. С. 1-20.

6. Sparck Jones K. A statistical interpretation of term specificity and its application in retrieval //Journal of documentation. 1972. Т. 28. №. 1. С. 11-21.

7. Kaur A., Saini J. K., Bansal D. Detecting radical text over online media using deep learning //arXiv preprint arXiv:1907.12368. 2019.

8. Рубцова Ю. В. Построение корпуса текстов для настройки тонового классификатора //Программные продукты и системы. 2015. №. 1 (109). С. 72-78.

9. Минаев, В. А. Методы снижения шумовых факторов при выявлении контента экстремистского характера в социальных медиа / В. А. Минаев, Е. С. Поликарпов, А. В. Симонов // Информация и безопасность. 2022. Т. 25, № 2. С. 179-186.

10. Kuratov Y., Arkhipov M. Adaptation of deep bidirectional multilingual transformers for Russian language //arXiv preprint

arXiv:1905.07213. 2019.

11. Минаев В. А., Симонов А. В. "Просеивание" телеграмм-каналов при поиске контента экстремистского характера // Информация и безопасность. 2023. Т. 26. № 1. С. 25-30.

Московский университет МВД России им. В.Я. Кикотя
V. Ya. Kikot Moscow University of the Russian Internal Affairs Ministry

Московский государственный технический университет им. Н. Э. Баумана
Bauman Moscow State Technical University

Поступила в редакцию 29.04.24

Информация об авторах

Минаев Владимир Александрович – д-р техн. наук, профессор, профессор кафедры специальных информационных технологий, Московский университет МВД РФ им. В.Я. Кикотя, e-mail: m1va@yandex.ru

Симонов Александр Валерьевич – аспирант кафедры защиты информации, Московский государственный технический университет им. Н. Э. Баумана, e-mail: san.siman@yandex.ru

SOFTWARE AND ALGORITHMIC SEARCH COMPLEX OF DESTRUCTIVE INFORMATION IN INSTANT MESSAGING SYSTEMS

V.A. Minaev, A.V. Simonov

The problems and stages of creating a software-algorithmic complex for processing text arrays in order to identify destructive information in instant messaging systems are described. A review of current scientific works in this field has been carried out. The concept and content of destructive information are revealed. The choice of a solution to the problem of text classification is justified. The procedure for searching and generating training content is presented. The architecture of the software package is presented, the functional characteristics of its modules are given. The data model is described, according to which the relationship of entities in the DBMS is developed, as well as the choice of programming language and libraries used. An approach to unifying the system of searching for destructive infomation in social media is discussed, based on solving the problem of classifying text arrays of any size.

Keywords: software and algorithmic implementation, social media, messenger, destructive content, neural network, extremism, transformer BERT.

Submitted 29.04.24

Information about the authors

Vladimir A. Minaev – Dr. Sc. (Technical), Professor, Professor of the Special Information Technologies Department, V. Ya. Kikot Moscow University of the Russian Internal Affairs Ministry, e-mail: m1va@yandex.ru

Alexander V. Simonov – Post-graduate student of the Information Security Department, Bauman Moscow State Technical University, e-mail: san.siman@yandex.ru