DOI 10.36622/VSTU.2023.26.1.003
УДК 519.766.4:004.032.26

«ПРОСЕИВАНИЕ» ТЕЛЕГРАМ-КАНАЛОВ ПРИ ПОИСКЕ КОНТЕНТА ЭКСТРЕМИСТСКОГО ХАРАКТЕРА

В.А. Минаев, А.В. Симонов

Предложено решение задачи поиска и обнаружения каналов экстремистской направленности в наиболее популярном мессенджере Telegram. Разработан метод, основанный на использовании глубоких искусственных нейронных сетей BERT в качестве классификатора текстов. На его основе разработана программа, позволяющая в автоматизированном режиме осуществлять анализ телеграмм-каналов и выявлять материалы экстремистского характера. Описаны пять этапов разработки программы «просеивания» каналов: формирование текстовых корпусов для обучения BERT; выбор архитектуры BERT и обучение классификатора; выгрузка сообщений из анализируемых каналов; оценка наличия экстремизма в сообщениях; оценка каналов на экстремизм. Для апробации метода и программного продукта проведена серия экспериментов, в ходе которых осуществлен мониторинг 68 каналов по 8 категориям с целью их «просеивания» на наличие экстремистского контента. По результатам экспериментов ранжированы телеграмм-каналы, распространяющие экстремистские материалы. Оценена доля сообщений экстремистского характера в каждой категории каналов. Сделан вывод, что представленный подход к «просеиванию» телеграмм-каналов на наличие экстремистских сообщений целесообразно использовать в работе государственных структур, занимающихся выявлением и мониторингом распространения противоправной информации.

Ключевые слова: мессенджер Telegram, деструктивный контент, нейронная сеть, экстремизм, трансформер BERT.

Введение

Современное общество невозможно представить без цифровых технологий, в том числе – в сфере межличностного взаимодействия. Это наблюдается на примере повсеместного использования социальных медиа (СМ) – технологий обмена различной информацией, как для делового, так и для личного общения. К СМ относят не только социальные сети (ВКонтакте, Одноклассники), видеохостинги (TikTok, YouTube), но сервисы обмена мгновенными сообщениями (WhatsApp, Telegram). СМ развиваются и, благодаря высокому охвату аудитории, стремительно изменяют информационную модель общества.

В связи с общедоступностью и простой реализацией анонимности, СМ все в большей мере используют в незаконных целях – для распространения экстремистского контента (ЭК). С правовой точки зрения к ЭК относятся частичные или полные версии материалов, отнесенных к экстремистским в соответствии с решениями суда, и находящихся в Федеральном списке экстремистских материалов Минюста России (ФСЭМ).

Согласно Доктрине информационной безопасности Российской Федерации, утвержденной Указом Президента Российской Федерации № 646 от 05.12.2016 [1], распространение ЭК является одной из информационных угроз. Таким образом, проблема носит государственный характер.

Распространением ЭК в целях подрыва общественной безопасности и снижения доверия к структурам законной власти могут заниматься как отдельные адепты-идеологи, организованные радикальные группы (террористические формирования, объединения националистов), так и недружественные государства.

Функции по выявлению ЭК и борьбе с его распространением возложены на государственные органы Российской Федерации: ФСБ, МВД, Генеральную Прокуратуру, Роскомнадзор.

Согласно 149-ФЗ [2] владельцы СМ обязаны самостоятельно осуществлять мониторинг своих ресурсов на наличие в них информации противоправного характера. Вместе с тем, владельцы до сих пор не

торопятся осуществлять полноценный мониторинг своих СМ. Это обстоятельство ещё больше усугубляет проблему контроля распространения ЭК.

В соответствии с данными Генеральной Прокуратуры Российской Федерации [3], показанными на рис. 1, с 2019 года после определенного снижения наблюдается стремительный рост преступлений экстремистской направленности. Из анализа приведенных данных следует вывод о возрастании проблемы контроля за распространением ЭК в СМ.

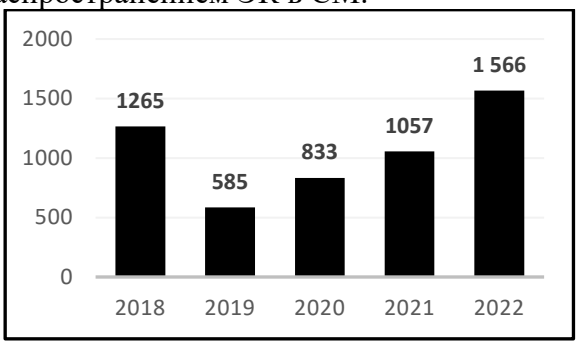

Рис.1. Динамика преступлений экстремистской направленности (данные Генеральной Прокуратуры России)

Одним из наиболее употребляемых форматов контента остается текстовый. Контент такого формата легче другого (видео, аудио, фотоизображения) создать, разместить и, что самое главное – модифицировать.

Сегодня по количеству распространяемого текстового контента лидирует Telegram. Необходимо отметить, что Роскомнадзор уже принимал меры в отношении Telegram из-за несоблюдения им порядка работы с запрещенной к распространению в России деструктивной информации.

Ежемесячно в русскоязычном сегменте Telegram создается более 700 миллионов сообщений. Такой информационный трафик невозможно анализировать на предмет деструктивности исключительно силами экспертов, необходимо применять автоматизированные методы, основанные на алгоритмах искусственного интеллекта.

Описание метода «просеивания»

В ряде научных работ уже сделаны попытки решить задачу обнаружения экстремистского контента в Telegram.

Например, в работе [4] осуществляется анализ деструктивного контента Telegram-каналов (ТГ-каналы). В основе реализованного в ней подхода лежит формирование словарей, где к деструктивной категории причисляются те сообщения, которые используют терминологию из них. Такой подход имеет недостатки, связанные с так называемыми шумовыми текстами, синтактически похожими, но семантически отличными от экстремистских. Примером может быть использование слова «фюрер» в различных контекстах.

В работе [5] используются методы машинного обучения для мониторинга информации экстремистской направленности. Поскольку для определения тематики контента применяют ключевые слова, недостатки метода аналогичны вышеупомянутой работе.

В качестве программного продукта (ПП) для обнаружения деструктивного контента может использоваться комплекс «Терьер» [6].

Однако у рассмотренных работ [4-6] есть общий недостаток. Как упомянуто ранее, к экстремистскому контенту, с правовой точки зрения закона, относятся только материалы, которые признаны таковыми по решению суда. В связи с этим формировать терминологические словари и обучать модели распознавания необходимо только с учетом данных материалов, чего не сделано в рассматриваемых работах.

Решение задачи «просеивания» связано с разработкой и применением такого ПП, который реализует методы обработки, основанные на классификации текстов на естественном языке.

В качестве основного преобразователя текста в вектор, позволяющего учитывать семантическую составляющую, используем глубокую нейронную сеть BERT, эффективность применения которой показана в работе [7]. Согласно работе [8] для борьбы с шумовыми текстами используем их в обучающей выборке. Таким образом, предлагаемый подход позволяет избежать недостатков методов, рассмотренных выше.

Для работы с текстовыми данными, методами машинного обучения и взаимодействием с web-сервисами, как показал опыт, хорошо подходит язык программирования Python, на котором

разработан ПП.

Разработка ПП по «просеиванию» ТГ-каналов состоит из следующих этапов:

1) формирование текстовых корпусов для обучения BERT,

2) выбор архитектуры BERT. Обучение классификатора,

3) выгрузка сообщений из анализируемых ТГ-каналов,

4) оценка наличия экстремизма в сообщениях,

5) оценка ТГ-каналов на экстремизм.

Опишем каждый этап подробнее.

Этап 1. Для корректной работы ПП необходимо обучение искусственной нейронной сети (ИНС) BERT на релевантных корпусах текстов. В разрабатываемом ПП используется бинарный классификатор, то есть корпуса текстов подбираются только двух классов: нейтрального и экстремистского.

Обучающая выборка нейтрального класса состоит из следующих пяти тематических корпусов.

1. *Пользовательский* – применен набор пользовательских твитов из работы [9].

2. *Новостной* – использован набор новостных сводок Интернет-издания «Лента».

3. *Исторический* – включен ряд книг об истории второй мировой войны.

4. *Еврейской культуры* – создан на основе еврейской культурной и исторической литературы.

5. *Исламской культуры* – создан на основе религиозной литературы, рекомендованной муфтиятами России.

Корпусы 1-2 используются для моделирования традиционного наполнения ТГ- каналов, а 3-5 в качестве шумовых корпусов с целью снижения ложных срабатываний классификатора.

Выборка экстремистского класса создана из материалов ФСЭМ и включает в себя следующие корпуса.

1. *Террористический* – содержащий запрещенные к распространению материалы радикального ислама, пропагандирующие терроризм.

2. *Националистический* – собранный из различных шовинистических и антисемитских материалов, пропагандирующих превосходство на основе расовой или национальной принадлежности.

3. *Нацистский* – состоящий из литературы и мемуаров лидеров Третьего рейха, а также литературы, в которой реабилитируются преступления нацистского режима Германии.

Объем выборок текстовых корпусов представлен в табл. 1.

Таблица 1

Объемы текстовых корпусов

Корпус	Объем (в символах)
Пользовательский	17 016 237
Новостной	14 248 399
Исторический	13 913 148
Еврейской культуры	13 707 565
Исламской культуры	11 985 455
Террористический	11 049 519
Националистический	10 228 206
Нацистский	10 187 667

Этап 2. Для использования обучающей выборки, сформированной на этапе 1, необходимо проведение ее предварительной обработки. Из массива текстов удаляются специальные символы за исключением знаков препинания, ссылки и сайты преобразуются к специальному виду.

В качестве входных в BERT подаются только токенизированные данные. Токенизация – это процесс разделения слов на единицы (токены), которым в соответствии со специальным словарем имеется однозначное числовое представление. В Telegram максимально возможная длина публикации – 4096 символов, в связи с чем все данные корпусов текстов разбиваются на части, равные этому числу. Экспериментально получено, что среднее количество символов в каждом токене равно 5,4. Из этого следует, что минимальная длина последовательности, подаваемой на вход в BERT, равна 745. Такому входному вектору соответствует модель rubert-tiny2 (768 токенов), доступ к которой возможен с репозитория моделей HuggingFace [10].

Процесс обучения BERT можно разделить на следующие три подзадачи.

Предварительное обучение – позволяет «научить» модель конкретному языку, его семантическим и синтаксическим особенностям. Данная задача решена разработчиком rubert-tiny2, в связи с чем

обучать модель с нуля нет необходимости.

Доменная адаптация – обучение BERT на специфических корпусах текстов, редко встречающихся в обычной литературе. В нашем случае к ним относятся корпуса экстремистского контента, поэтому доменная адаптация проводилась.

Тонкая настройка – процесс обучения BERT для решения конкретной задачи – классификация текстов, извлечение сущностей, поиск похожих предложений и т.д. Для решения задачи «просеивания» произведена тонкая настройка классификатора BERT, в роли которого использовалась полносвязная однослойная нейронная сеть.

В качестве обучающей выборки использовались 85% всех данных, остальные 15% использовались для оценки работы классификатора.

Качество классификации оценивались F-мерой [11], представляющей гармоническое среднее между точностью и полнотой. Эта мера стремится к нулю, если точность или полнота стремится к нулю.

Под точностью классификации понимается доля сообщений в ТГ-канале, действительно принадлежащих данному классу относительно всех сообщений, которые классификатор отнес к этому классу.

Полнота – это доля найденных классификатором сообщений в ТГ-канале, принадлежащих классу относительно всех сообщений этого класса в тестовой выборке.

Популярность *F*-меры обусловлена тем, что она учитывает как полноту, так и точность работы классификатора, что очень важно при неравномерных выборках.

Результаты расчета точности на оценочной выборке представлены в табл. 2, где НК – нейтральный класс, а ЭК – экстремистский класс.

Таблица 2

Результаты работы классификатора на оценочной выборке

Истинный класс	ЭК	**0,992**	**0,007**
	НК	**0,013**	**0,986**
		НК	ЭК
		Предсказанный класс	

Этап 3. Для выгрузки необходимой информацией использовалась библиотека telethon. Для того, чтобы получить информацию с определенных каналов, необходимо создать пользователя, с которым посредством библиотеки telethon будет происходить взаимодействие.

Для сбора необходимых данных используется модуль, на вход которого подаются сообщения в каждом Telegram-канале. Выходными данными модуля являются: название и ID канала, ссылка на сообщение, текст и длина сообщения, количество просмотров сообщения, количество репостов сообщения).

Этап 4. В рамках оценки наличия экстремизма в сообщениях, перед подачей текстов на вход классификатора происходит их первичная фильтрация по объему символов. Это необходимо для снижения большого количества ложных срабатываний. Все сообщения и публикации, в которых объем символов меньше 300, не анализируются.

После этого сообщения, прошедшие первичную фильтрацию, подвергаются обработке, описанной на этапе 2.

Этап 5. Для каждого канала рассчитываются следующие показатели:

N – количество всех проанализированных сообщений;

N_{ME} – количество экстремистских сообщений;

N_{MN} – количество нейтральных сообщений;

$\omega_{ME} = \frac{N_{ME}}{N}$ – доля экстремистских сообщений;

$\omega_{MN} = \frac{N_{MN}}{N}$ – доля нейтральных сообщений.

Приведенные показатели позволяют ранжировать каналы по критерию экстремистской направленности, обеспечивая поддержку принятия решений операторам для дальнейшего точечного анализа каналов и сообщений.

Результаты экспериментов

Для оценки метода и реализованной на его основе программы проведен отбор 68 телеграмм-каналов, материалы которых отражали каждый из текстовых корпусов, представленных в табл. 1.

На рис. 2 представлена общая статистика по всем проанализированным каналам, сгруппированным по категориям, представленным в табл. 1. В качестве столбцов на гистограмме представлена доля экстремистских сообщений в их общем количестве.

При тщательном экспертном анализе выявлено, что в указанных каналах присутствуют цитаты и выдержки из экстремистских материалов нацистского, националистического или террористического содержания.

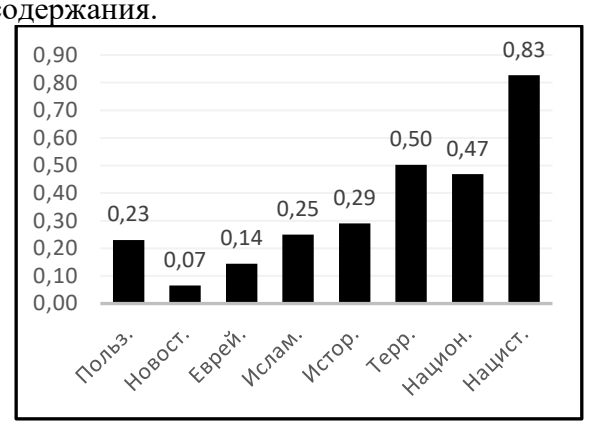

Рис. 2. Гистограмма доли сообщений экстремистского характера в каналах по категориям

Из рис. 2 видно, что наибольшее количество экстремистского контента выявлено в каналах нацистской категории. Экстремистские сообщения в каналах террористической категории и категории, представляющей радикальный национализм, составили около половины.

Выводы

В статье описан метод «просеивания» ТГ-каналов при поиске контента экстремистского характера с использованием глубокой искусственной нейронной сети BERT.

На основе разработанного метода создан программный продукт, выполняющий автоматический сбор, анализ и классификацию сообщений в ТГ-каналах, направленную на поддержку принятия решений по углубленному экспертному анализу конкретных материалов и каналов.

Для компьютерных экспериментов по «просеиванию» ТГ-каналов сформирован список каналов по восьми категориям, подвергшихся дальнейшему анализу. Приведенные эксперименты показали эффективность разработанного метода.

Эксперименты показали, что ряд исследованных каналов публикует экстремистские материалы. Это доказано с помощью автоматизированного сопоставления содержания публикуемых материалов с текстами, содержащимися в ФСЭМ. Наибольшее количество экстремистских материалов содержатся в каналах националистической направленности.

Представленный подход к «просеиванию» ТГ-каналов на наличие экстремистских сообщений целесообразно использовать в работе государственных структур, занимающихся выявлением и мониторингом распространения противоправной информации.

Список литературы

1. Указ Президента Российской Федерации "Об утверждении Доктрины информационной безопасности Российской Федерации" от 05.12.2016 № 646 // Собрание законодательства Российской Федерации. 2016 г. № 50. Ст. 7074.

2. Федеральный закон Российской Федерации "Об информации, информационных технологиях и защите информации" № 149 от 14.07.2006 // Собрание законодательства Российской Федерации. 2006. № 31. Ст. 3448.

3. Показатели преступности России // Генеральная Прокуратура Российской Федерации, портал правовой статистики. URL: http://crimestat.ru/offenses_chart (дата обращения 16.02.2022).

4. Углова А. Б., Низомутдинов Б. А. Анализ деструктивного контента телеграмм-каналов как фактора развития саморазрушающего поведения // International Journal of Open Information Technologies. 2022. Т.10. № 11. С. 81-86.

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

6. Программный комплекс «Терьер» // ФНПЦ АО «НПО «Марс». URL: http://www.npomars.com/ru/products/sys_upr_all/terer/ (Дата обращения: 16.02.2022).

7. Минаев В. А., Симонов А. В. Сравнение моделей-трансформеров BERT при выявлении деструктивного контента в социальных медиа // Информация и безопасность. 2022. Т. 25. № 3. С. 341-348.

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

9. Рубцова Ю. В. Автоматическое построение и анализ корпуса коротких текстов (постов микроблогов) для задачи разработки и тренировки тонового классификатора // Инженерия знаний и технологии семантического веба. 2012. Т. 1. С. 109-116. URL: https://huggingface.co/cointegrated/rubert-tiny2 (дата обращения: 16.02.2022).

10. Forman G. An Extensive Empirical Study of Feature Selection Metrics for Text Classification // Journal of Machine Learning Research. 2003. No 3 (Mar.). Pp. 1289-1305.

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 20.01.23

Информация об авторах

Минаев Владимир Александрович – д-р техн. наук, 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

SCREENING OF TELEGRAM CHANNELS WHEN SEARCHING FOR EXTREMISM CONTENT

V.A. Minaev, A.V. Simonov

A solution to the problem of searching and detecting extremist channels in the most popular Telegram messenger is proposed. A method based on the use of deep artificial neural networks BERT as a classifier of texts has been developed. Based on it, a program has been developed that allows automated analysis of telegram channels and identification of extremist materials. Five stages of the development of the channel "sifting" program are described: the formation of text corpora for BERT training; the choice of the BERT architecture and the training of the classifier; unloading messages from analyzed channels; assessing the presence of extremism in messages; evaluating channels for extremism. To test the method and the software product, a series of experiments were conducted, during which 68 channels were monitored in 8 categories in order to "sift" them for the presence of extremist content. According to the results of the experiments, telegram channels disseminating extremist materials were ranked. The share of extremist messages in each category of channels is estimated. It is concluded that the presented approach to "sifting" telegram channels for the presence of extremist messages is advisable to use in the work of state structures engaged in identifying and monitoring the dissemination of illegal information.

Keywords: Telegram messenger, destructive content, extremism, neural network, transformer BERT.

Submitted 20.01.23

Information about the authors

Vladimir A. Minaev – Dr. Sc. (Technical), Professor, Professor of the Department of Special Information Technologies, 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