ПРОГРАММИРОВАНИЕ, 2019, № 3, с. 18–37

ИНФОРМАЦИОННЫЙ ПОИСК

УДК 004.8+004.9

МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ЗАДАЧИ ОБНАРУЖЕНИЯ И МОНИТОРИНГА ЭКСТРЕМИСТСКОЙ ИНФОРМАЦИИ В СЕТИ ИНТЕРНЕТ[1]

© 2019 г. И. В. Машечкин[a,*], М. И. Петровский[a,], Д. В. Царев[a,***], М. Н. Чикунов[a,****]**

[a] Московский государственный университет имени М.В. Ломоносова
Факультет вычислительной математики и кибернетики (ВМК)
119899 Москва, Ленинские горы, д. 1, стр. 8, Россия

** E-mail: mash@cs.msu.su*

*** E-mail: michael@cs.msu.su*

**** E-mail: tsarev@cs.msu.su*

***** E-mail: chikunovmn@mail.ru*

Поступила в редакцию 15.01.2019 г.
После доработки 15.01.2019 г.
Принята к публикации 17.01.2019 г.

Данная статья посвящена применению методов машинного обучения для решения задачи обеспечения безопасности, в части – противодействия терроризму и экстремизму с использованием информации из сети Интернет. Эти задачи включают поиск электронных сообщений, документов и web ресурсов, содержащих потенциально террористическую и экстремистскую информацию, выявление структуры групп пользователей и Интернет сообществ, распространяющих такую информацию, осуществление мониторинга и тематического моделирования потоков информации, циркулирующих в таких сообществах, оценку угроз и прогнозирование рисков на основе полученных результатов мониторинга. В работе предлагаются оригинальные языково-независимые алгоритмы для информационного поиска по образцу, тематического моделирования и прогнозирования характеристик потоков текстовых сообщений, оценки и прогнозирования риска, исходящего от членов Интернет сообщества с использованием данных о структуре связей в сообществе, что позволяет находить потенциально опасных пользователей, даже не имея полного доступа к контенту, который они распространяют, например, через закрытые каналы и чаты.

DOI: 10.1134/S0132347419030063

1. ВВЕДЕНИЕ

В современном мире растет число террористических актов, осуществляемых экстремистскими группами или отдельными людьми, находящимися под влиянием экстремистских идей. Это могут быть спланированные атаки, организованные крупным террористическим сообществом, такие как “11 сентября”, и атаки террористов одиночек, таких как Брейвик или братья Царнаевы. При этом за последние 20 лет ситуация значительно изменилась. В 90-х годах фактически существовала общая достаточно статичная картина [1] того, какие террористические группы существуют, как и кем финансируются, какого типа теракты и против каких государств или структур могут осуществить, кто является руководителем или организатором конкретных акций. В настоящее время все чаще появляются террористы одиночки, ничем, кроме Интернет общения, не связанные с вдохновителями теракта. Зачастую они сами являются организаторами атаки, находясь под воздействием пропаганды или экстремистской идеологии, также распространяемой через Интернет. Кроме того, даже крупные террористические и экстремистские организации в целях повышения живучести переходят от вертикально-иерархической к горизонтально-сетевой или слабо связанной структуре. Поэтому роль Интернет как средства обмена информацией и распространения пропаганды в рамках террористических и экстремистских сообществ многократно возрастает [2]. Основными средствами общения организаторов и исполнителей терактов, в зависимости от задач, служат электронная почта, социальные сети обмена короткими сообщениями (в том числе пуб-

[1] Исследование выполнено при поддержке гранта РФФИ 16-29-09555 офи_м.

личными), программы-мессенджеры, публичные форумы и просто веб ресурсы, которые служат в основном для ведения пропаганды и вербовки новых сторонников. Поэтому данные в сети Интернет могут оказаться бесценным источником информации для обнаружения, мониторинга активности, выявления структуры и оценки угрозы террористических и экстремистских сообществ с целью предотвращения террористических атак. Это подтверждает и анализ "постфактум" публичных данных из социальных сетей, оставленных исполнителями террористических актов, который показывает, что была техническая возможность заранее выявить склонность исполнителя к радикализму, а значит, была возможность предотвратить теракт. Например, исполнитель каталонского теракта Мусса Укабир (https://www.bbc.com/russian/news-40982274) до исполнения атаки выражал в социальных сетях желание "убить всех неверных".

Таким образом, для решения задач противодействия терроризму и экстремизму с использованием информации из сети Интернет является актуальным разработка программных систем, позволяющих:

1. Выявлять группы пользователей, сообщества и ресурсы в сети Интернет, где циркулирует информация террористического или экстремистского содержания.

2. Осуществлять мониторинг, получать и прогнозировать характеристики потоков сообщений и документов, распространяемых в таких группах.

3. Оценивать опасность и прогнозировать риски, которые несут члены таких сообществ.

При этом работа с информацией террористического и экстремистского содержания имеет ряд важных особенностей и ограничений, которые необходимо учитывать при разработке методов машинного обучения для решения обозначенных выше задач. Анализируемые текстовые сообщения можно представить в виде совокупности текста и набора нетекстовых атрибутов, характеризующих сообщение. В общем случае текст сообщения может быть документом произвольного объема с обязательным атрибутом – временной меткой регистрации сообщения или публикации документа. Возможны дополнительные атрибуты, например, отправитель и получатель или автор и читатель, которые могут использоваться для построения топологии группы пользователей или сетевого сообщества. Сообщения могут быть короткими, состоящими из нескольких слов, и очень большими текстовыми документами. Также присутствует проблема ссылок и хэштегов. Зачастую все сообщение может состоять только из них, поэтому для представления содержания такого сообщения необходимо выгружать и анализировать контент ресурсов или пользователей, на которые ссылается исходное сообщение. Важным фактором является язык написания сообщения. Причем особенностью текстов экстремистского и террористического содержания является использование нескольких различных языков в одном документе, а также наличие опечаток и грамматических ошибок, в том числе преднамеренных, с целью "замаскировать" ключевые слова, чтобы осложнить автоматический поиск по ним. Также в ресурсах террористического и экстремистского содержания может использоваться сленг или жаргон, употребляемый только в узком кругу пользователей, могут использоваться специальные кодовые слова или обозначения для замены ключевых слов, по которым обычно осуществляется поиск. Все эти особенности делают крайне трудоемким и малоэффективным применение в обозначенных задачах традиционных методов, основанных на лингвистике и NLP (Natural Language Processing), подразумевающих создание тезаурусов специфической лексики и ключевых слов экспертами-лингвистами. Поэтому в настоящем исследовании предлагается делать акцент на языково-независимые методы анализа текстов, преимущественно статистические, с выделением признаков текстов на основе n-грамм и латентно-семантического анализа.

При анализе структуры сетевых сообществ, в которых циркулирует информация террористического и экстремистского содержания, возникает проблема установления связей между членами сообщества. В каких-то случаях наличие связи между пользователями очевидно, например, при "подписке" или добавлении в "друзья", при прямых "репостах" сообщений. Но иногда, особенно в случае форума, а не социальной сети, наличие связи установить сложнее, например, то, что один пользователь ответил в ветке, созданной другим пользователем, не значит, что он прочитал всю ветку до корня, хотя это вероятно. В случае, когда структура графа общения пользователей все-таки получена, остается проблема, связанная с доступом к закрытой переписке между пользователями. Это приводит к необходимости решать задачу следующего вида: дан граф взаимодействия пользователей, про часть из них известно, что они генерируют или распространяют экстремистскую информацию, про часть известно, что нет, а про часть не известно, что именно они пишут или читают. Задача состоит в том, чтобы методами машинного обучения, с использованием информации только о структуре графа, спрогнозировать какие из пользователей, чья переписка и публикация не доступны, являются опасными, а какие – нет.

Структура дальнейшей части работы имеет следующий вид. В разделе 2 дается краткий обзор основных опубликованных подходов в области применения методов машинного обучения для

задачи обнаружения и мониторинга экстремистской информации в сети Интернет. Раздел 3 содержит описание предлагаемого языково-независимого подхода для поиска по образцу экстремистской информации в сети Интернет, а в разделе 4 предложены методы мониторинга, латентно-семантического анализа, оценки и прогнозирования характеристик потоков сообщений и документов, циркулирующих в подозрительных Интернет сообществах. В разделе 5 рассматривается задача восстановления топологии сетевых сообществ, а также задача оценки угрозы и прогнозирования рисков, исходящих от отдельных пользователей сообщества без учета генерируемого ими контента. В разделе 6 представлены выводы.

2. СУЩЕСТВУЮЩИЕ ПОДХОДЫ

Важность анализа Интернет информации при решении задач противодействия терроризму в настоящее время понимается на самых разных уровнях. Исследованиями в этой области активно занимаются государственные агентства, частные компании [1], академические исследователи [2–6], причем некоторые университеты, например, Duke University (США), проводят отдельные междисциплинарные учебные и исследовательские проекты по этому направлению и даже организуют межуниверситетские аналитические центры по анализу и противодействию терроризму, такие как Triangle Center on Terrorism and Homeland Security (TCTHS) (http://sites.duke.edu/tcths/). Проводятся конференции, симпозиумы и рабочие группы, посвященные, в том числе, задачам выявления тематик циркулирующей в Интернет информации и структуры террористических и экстремистских Интернет-сообществ [7].

Из опубликованных материалов можно выделить три основных направления исследований в области применения методов машинного обучения для обнаружения, мониторинга и прогнозирования активности террористического и экстремистского характера в сети Интернет.

1. Сбор анализируемых данных. В качестве анализируемых данных в сети Интернет используют следующие типы источников информации: web-сайты террористической или экстремистской направленности, новостные web-сайты, страницы пользователей социальной сети твиттер, корпоративные сообщения электронной почты. Помимо текстовых данных широко используются открытые международные базы событий террористических атак, такие как Rand Database of Worldwide Terrorism Incidents (RDWTI), Global Terrorism Database (GTD), World Incident Tracking System (WITS), Terrorism in Western Europe: Events Data (TWEED) и другие. В большинстве из них в хронологическом порядке перечислены факты совершенных террористических атак с указанием времени, места, целей, числа жертв, используемых средств, ответственной организации или группы, текстовым резюме и другими аналогичными признаками. На основе этих баз многие исследователи строят модели прогнозирования типов и характеристик будущих терактов, распознавания террористических групп, совершивших теракт и т.д. Но результаты, полученные с использованием таких баз в качестве единственного источника информации, заслуженно критикуются [4] по ряду важных причин. А именно, в базах содержится информация только о совершенных (не предотвращенных) актах, нет информации о предпринятых в тот временной период действиях компетентных служб по защите, предотвращению или уменьшению нанесенного ущерба. Помимо этого, зачастую целью террористической атаки является не максимизация жертв как таковых и не нанесение ущерба какой-либо инфраструктуре, а ее пропагандистская значимость, то, насколько она повлияла на общественное мнение. Этой информации также нет в открытых базах событий террористических атак.

В последнее время появилось довольно много работ, в которых тестирование алгоритмов анализа текстов экстремистского содержания проводится на данных проекта Dark Web. Эти данные были собраны сотрудниками Аризонского университета (The University of Arizona) с различных форумов и сайтов выявленных террористических организаций [8–10, 17]. Появление Dark Web дало импульс к проведению большого числа разнообразных исследований, основанных на тематическом анализе. Эти данные содержат несколько терабайт текстовых сообщений, преимущественно из исламистских форумов и чатов, которые были классифицированы в лаборатории Искусственного интеллекта университета Аризоны, США, как потенциально террористические. В рамках проекта DarkWeb есть примеры таких материалов (от нескольких тысяч до нескольких сотен тысяч сообщений в каждом) на английском, арабском и французском языках. Наиболее часто используемый англоязычный набор Ansar1 считается “полностью экстремистским”, поскольку содержит материалы закрытого джихадистского форума. Отдельно следует отметить набор данных под названием KavkazChat из проекта DarkWeb. Этот набор содержит информацию, собранную на форумах, преимущественно посвященных проблемам и жизни российского Северного Кавказа, где были выявлены сообщения экстремистского и террористического содержания. Объем текстовых данных достаточно велик, весь набор содержит более 600 гигабайт текстовых данных, включая сообщения на русском языке – на кириллице и в транслите, на арабском языке, на национальных языках Северного Кавказа в кириллической

и латинской транскрипции. Причем многие сообщения содержат текст сразу на нескольких языках. В наборе данных содержится 16 тысяч веток обсуждения разной тематической направленности, в которых участвуют несколько тысяч пользователей. Объем веток обсуждения варьируется от одного килобайта до 5 мегабайт. Далеко не все ветки содержат информацию потенциально экстремистского содержания. Много сообщений посвящено обсуждению религиозных тем – правил поведения в исламском обществе, взаимоотношений между мужчинами и женщинами в нем; встречаются также бытовые темы (кулинария, спорт, автомобили); много сообщений посвящено обсуждению политических событий в мире, так или иначе связанных с Россией, Кавказом и Ближним Востоком. Следует отметить, что простой "ручной" поиск по ключевым словам для такого типа данных дает крайне низкую точность выявления экстремистской информации. В ветках, посвященных обсуждению политических событий, используется близкая лексика, при этом зачастую грань между обычным комментарием и потенциально экстремистским может быть очень тонкой. Например, к вполне нейтральному новостному сообщению о событии в горячей точке может быть добавлен комментарий, использующий словосочетание "русские оккупанты" или "американские террористы", что делает ветку подозрительной с точки зрения потенциального содержания экстремистской информации.

2. Методы анализа текстовой информации [3–6, 13]. Применяются традиционные подходы классификации, такие как деревья решений, логическая регрессия, наивный байесовский классификатор, метод опорных векторов, и другие. Используются методы выявления структурированной информации из неструктурированных или слабо структурированных данных, такие как распознавание именованных сущностей (Named Entity Recognition, NER). Для формирования признакового пространства описания текстовых сообщений используются традиционные признаки – ключевые слова и часто употребляемые словосочетания (фразы). Описанные методы применяются для классификации отдельных пользователей и Интернет-сообществ, порождающих или читающих контент террористического и экстремистского характера. Разработка языково-независимых моделей представления данных для потоков текстовых и гипертекстовых сообщений, с учетом их размера, наличия информационного шума, сленга, жаргона и "маскирующих" кодовых слов, ссылочной структуры, дополнительных атрибутов, таких как время создания, автор и получатель (читатель) материала и других, является в настоящее время важным открытым исследовательским направлением в области анализа Интернет-информации, в том числе для задач противодействия экстремизму и терроризму. Для тематического моделирования наиболее популярным подходом является использование вероятностных моделей представления на основе LDA (с использованием скрытого распределения Дирихле) и других аналогичных моделей, а методы матричного разложения критикуются за неиспользование при моделировании вероятностной природы процессов порождения текстов. Но результаты [11, 12] показывают, что такая критика не заслужена, и методы на основе матричных разложений не уступают, а зачастую и превосходят вероятностные модели.

3. Исследование топологии Интернет-сообществ. Это направление включает выявление ключевых узлов, расчет их метрик (связность, мощность и другие), построение моделей поведения пользователей для оценки влияния отдельных узлов на сообщество в целом. Изначально эти методы применялись для решения вполне "гражданских" задач, таких как маркетинг, исследование игровых или потребительских сообществ, но многие из этих методов успешно нашли свое применение и в области противодействия терроризму [5, 14, 15]. Также разрабатывались специальные модели и методы, ориентированные на контртеррористическую тематику, среди которых можно выделить следующие. В работе [16] по данным записей в Twitter решается задача выявления пользователей-экстремистов, а также оценивается, будет ли обычный пользователь выбирать экстремистские материалы и будут ли пользователи отвечать на контакты, инициированные экстремистами. В работе [17] предлагается подход, комбинирующий традиционные методы сетевого анализа для выявления перекрывающихся сообществ со средствами текстового анализа тематических моделей. Для выявления тематик в работе применяется LDA, который в комбинации с алгоритмом "все предыдущие ответы" (all-previous-reply) позволяет построить сеть взаимосвязей участников форума по набору тематик. Работы [18, 19] исследуют возможность идентификации вербовочной активности экстремистских групп на сайтах социальных сетей и предлагают методы прогнозирования уровня ежедневной активности кибер-вербовки. Для идентификации вербовочных постов используется модель на основе SVM. Текстовое содержание анализируется с помощью LDA. Результаты анализа подаются в различные модели временных рядов для прогнозирования активности вербовки. Количественный анализ показывает, что использование основанных на LDA тематик в качестве предикторов в моделях временных рядов уменьшает ошибку прогнозирования по сравнению с другими методами. Схожий подход предлагается в работе [20], посвященной решению задачи выявления ключевых членов сообщества на основе тематик, для чего

Рис 1. Схема работы поиска по образцу.

комбинируются инструменты интеллектуального анализа текстов и анализа социальных сетей.

Отдельно хотелось бы отметить подход к комплексному выявлению потенциально опасных людей и сообществ на основе технологий Больших данных [1]. Это перспективное направление, подразумевающее, помимо анализа Интернет-данных, сбор огромного количества информации из различных источников, таких как история мобильных звонков, билеты на транспортные средства, таможенные декларации, факты пересечения границы, аренда автомобилей, криминальные сводки по региону и многие другие. Основным математическим инструментом, используемым при обнаружении фактов подозрительной активности в такой постановке, являются методы интеллектуального анализа данных для моделирования типовых сценариев поведения людей и поиска исключений – фактов, кардинально отличающихся от типового поведения. Далее эти отдельные факты анализируются более пристально с привлечением экспертов на предмет наличия террористической угрозы. Стоит отметить, что такие проекты можно реализовывать только с поддержкой компетентных государственных структур, имея законный доступ к конфиденциальной информации из указанных типов источников.

3. ПРЕДЛАГАЕМЫЙ ПОДХОД ДЛЯ ПОИСКА И МОНИТОРИНГА ЭКСТРЕМИСТСКОЙ ИНФОРМАЦИИ В СЕТИ ИНТЕРНЕТ

Большинство существующих и разрабатываемых подходов поиска информации террористического и экстремистского содержания в сети Интернет носят языково-зависимый характер. Специалистами формируются тезаурусы экстремисткой лексики, в том числе с учетом “горячих” регионов и грамматики национальных языков в этих регионах. Далее эти тезаурусы используются в системах информационного поиска для обнаружения текстов, содержащих найденные лексические конструкции. Безусловно, такой подход дает достаточно точные поисковые результаты по сравнению с рассматриваемым в настоящей статье языково-независимым, также он более прост с точки зрения применения исполнителем, осуществляющим поиск, поскольку не требует обучения и высокой квалификации. Но в то же время такой подход обладает рядом критических недостатков. Он является экстенсивным и из-за этого весьма трудоемким в плане настройки. Требуется значительное время для создания тезаурусов и настройки поисковых систем. При добавлении новых языков и регионов необходимо привлекать лингвистов и вносить существенные изменения в алгоритмы работы поисковых систем. Есть проблемы с нечетким поиском для обнаружения искаженных, замаскированных или неправильно написанных слов и терминов. Причем даже в рамках одного языка и региона лексика в области экстремизма и терроризма постоянно меняется, перестают использоваться одни термины и появляются новые термины, упоминания новых людей и географических мест, обычные слова приобретают экстремистское значение (например, “ватник”, “укроп” и т.д.). Также при этом подходе велика вероятность ложно отрицательной ошибки, т.е. пропуска экстремистского текста, если он использует нестандартную лексику. Все это приводит к тому, что существующие системы информационного поиска террористической и экстремистской информации всегда “на шаг позади”, т.е. они ищут информацию актуальную “вчера”, а не сегодня, и требуют больших трудозатрат и высокой квалификации экспертов (в том числе лингвистов) для поддержания актуального состояния своих баз.

3.1. Двухэтапный поиск по образцу

Для решения перечисленных выше проблем предлагается новый подход на основе двухэтапного поиска по образцу [22]. В этом подходе вместо традиционного поискового запроса используется документ-образец, а цель поискового процесса – найти документы и сообщения релевантные этому образцу. В качестве документа-образца удобно использовать примеры сообщений и публикаций,

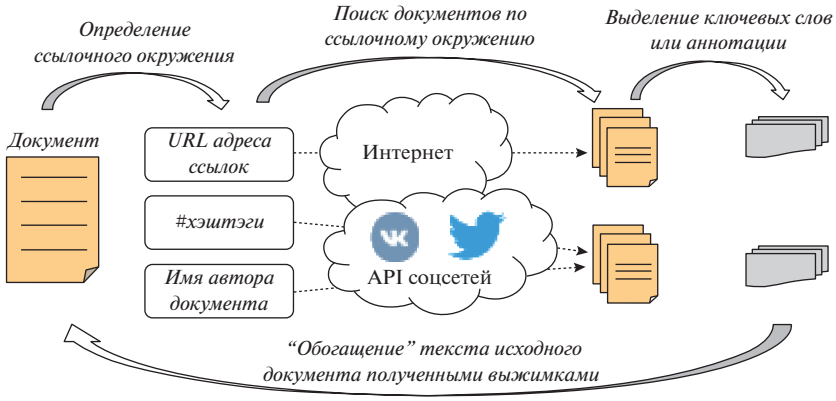

Рис 2. Схема процедуры "обогащения" анализируемого документа за счет его ссылочного окружения.

полученных из заведомо террористического или экстремистского источника. При этом сам поиск в сети Интернет осуществляется с помощью стандартных поисковых машин или поисковых машин социальных сетей. Для этого на первом этапе из документа-образца выделяются ключевые слова, с использованием языково-независимого подхода на основе ортонормированной неотрицательной матричной факторизации [11] и представления слов в виде n-грамм. Найденные ключевые слова формируют поисковый запрос, который выполняется сторонней поисковой машиной, а возвращаемая выдача, содержащая много шума и ошибок, уже ранжируется предложенным оригинальным методом на втором этапе. Ранжирование осуществляется таким образом, чтобы более релевантными были те документы в выдаче, в которых веса скрытых тематик документа-образца максимальны. Общая схема работы предложенного подхода представлена на рис. 1.

Документ-образец перед поиском и документы, возвращаемые поисковой машиной, "обогащаются" за счет информации из их ссылочного окружения. Для этого выделяются все ссылки и хэштеги в документах, а специально реализованный краулер осуществляет поиск в сети Интернет информации по заданной ссылочной структуре в зависимости от типа ссылки. Для веб ссылки выполняется http запрос для получения содержимого страницы по ее адресу, по хештегу осуществляется выборка заданного числа последних сообщений, содержащих этот тег в социальной сети, где он используется, для имени пользователя осуществляется выборка заданного числа его последних сообщений из соответствующей социальной сети. Таким образом формируется документ, приблизительно описывающий информацию по каждой заданной ссылке. К этому документу применяется разработанный метод выделения ключевых слов или метод аннотирования (раздел 3.3). Получаемая информационная "выжимка" вместо ссылочной структуры добавляется в исходный анализируемый текст до того, как будет формироваться его матричное представление (раздел 3.2). Общая схема "обогащения" документа за счет информации из его ссылочного окружения представлена на рис. 2.

Это позволяет учесть информацию из ссылочного окружения анализируемого документа и, в том числе, анализировать короткие сообщения, содержащие в основном ссылки и хэштеги (очень распространены в культуре пользователей Twitter).

3.2. Матричное представление на основе n-грамм и тематической модели документа

Предлагается использовать представление документа в виде матрицы, для этого его текст разбивается на непересекающиеся фрагменты. В качестве фрагментов выбирают предложения текста или его параграфы (при обработке текста большого размера). Для представления полученной коллекции фрагментов применяется векторная модель [23, 24], в которой в качестве признаков фрагментов текста используются n-граммы. Это позволяет решить проблемы, связанные с многоязыковостью, наличием ошибок и опечаток. Для получения признаков на основе n-грамм для каждого слова в тексте берутся подряд идущие буквосочетания фиксированной длины n [3]. Данный метод является достаточно универсальным, так как он применим для многих современных языков. Таким образом, пространство признаков формируется из множества полученных различных n-грамм L_m. Основным достоинством n-граммного представления текстов является отсутствие необходимости дополнительной лингвистической обработки текста. Разбиение на n-граммы вычислительно проще, чем лингвистический стемминг, а из-за ограниченности алфавита во всех языках максимальное число различных признаков также ограничено и невелико (при небольших n) по

сравнению с числом слов. К недостаткам n-грамм можно отнести то, что они могут сильно увеличить количество "ненулевых" признаков каждого отдельного текста, особенно при небольших значениях n.

Набор текстовых фрагментов документа в такой модели представляется в виде числовой матрицы $A \in \mathbb{R}^{m \times n}$, строки которой соответствуют признакам, а столбцы – фрагментам. Каждый фрагмент A_j $(1 \le j \le n)$ представляется в виде числового вектора $A_j = [a_{1,j}, a_{2,j}, \ldots, a_{m,j}]^T$ фиксированной размерности m, равной числу уникальных n-грамм, $a_{i,j}$ – i-я $(1 \le i \le m)$ компонента вектора A_j определяет вес i-го признака в j-м фрагменте. Вес $a_{i,j}$ вычисляется как произведение трех составляющих: $a_{i,j} = L_{i,j} \cdot G_i \cdot N_j$, где $L_{i,j}$ – локальный вес признака i в фрагменте j, G_i – глобальный вес признака i во всех фрагментах, N_j – нормализация вектора фрагмента A_j. Далее для построения тематической модели документа к его матричному представлению в пространстве n-грамм применяется метод латентно-семантического анализа, который основан на применении ортонормированной неотрицательной матричной факторизации (ОНМФ). Исходное матричное представление текста (матрица A) аппроксимируется произведением двух матриц с неотрицательными элементами: $A \approx W_k \cdot H_k$. Точность аппроксимации обеспечивается нахождением таких матриц $W_k \in \mathbb{R}^{m \times k}$ и $H_k \in \mathbb{R}^{k \times n}$, которые минимизируют целевую функцию [11]: $f(W_k, H_k) = \frac{1}{2}\|A - W_k H_k\|_F^2 + \frac{\alpha}{2}\|W_k^T W_k - I\|_F^2$, где $k \ll \min(m, n)$, $\alpha \ge 0$ – параметр ортонормированности W_k (при $\alpha > 0$ накладывается дополнительное условие: $W_k^T \cdot W_k = I$). Левая матрица (матрица тематик, W_k) служит для отображения между пространством тематик и пространством текстовых признаков, а правая (матрица фрагментов, H_k) – для представления фрагментов в пространстве тематик, фрагментам соответствуют столбцы матрицы H_k [11, 25, 26]. Требование ортонормированности накладывается на левую матрицу W_k для того, чтобы ее транспонирование позволило отображать пространство признаков в сформированное пространство тематик. Таким образом, получив тематическое представление одного документа, можно любой документ отобразить в пространство его тематик. Отметим, что применение ОНМФ к исходному матричному представлению A возможно за счет свойства неотрицательности получаемых весов признаков. Далее под тематической моделью будем понимать совокупность (L_m, W_k, H_k). Сформированная тематическая модель имеет следующие свойства. Во-первых, столбцы матрицы фрагментов W_k соответствуют выделенным тематикам. Элементы матрицы W_k неотрицательны, поэтому их можно рассматривать как вклад (вес) признаков (n-грамм) в соответствующую тематику, т.е. элементы матрицы W_k задают весовые коэффициенты n-грамм, объединяемых в каждой тематике. Чем больше значение j-го элемента в i-м столбце по сравнению с другими элементами столбца (i-й тематике), тем более характерна j-я n-грамма для данной тематики. Следовательно, выделенные тематики можно описывать n-граммами, имеющими наибольший вес. Аналогично, элементы матрицы фрагментов (H_k) неотрицательны, поэтому их можно рассматривать как вклад (вес) тематик в соответствующий фрагмент. Чем больше значение j-го элемента в i-м столбце по сравнению с другими элементами столбца (представление i-го фрагмента в пространстве тематик), тем сильнее фрагмент относится к j-й тематике.

Следует отметить, что описанная матричная модель представления документа является достаточно общим подходом, конкретный вид представления определяется схемой нормализации и кодирования глобальных и локальных весов. В рамках настоящего исследования были проверены разные схемы и для задачи аннотирования и поиска ключевых слов. В результате наилучшие результаты показала схема с бинарным локальным весом (есть или нет такая n-грамма в тексте) и глобальным весом на основе энтропии [22]:

$$G_i = 1 - \sum_{j=1}^{N}\left(\frac{p_{i,j}\log(p_{i,j})}{\log(N)}\right), \quad p_{i,j} = \frac{t_{i,j}}{\sum_{k=1}^{N} t_{i,j}},$$

где $t_{i,j}$ – число вхождений i-й n-граммы в j-й фрагмент. Нормализация для поиска ключевых слов и аннотирования не используется $N_j = 1$ [22–24]. В задаче расчета меры сходства при ранжировании выдачи используется схема с логарифмическим локальным весом:

$$L_{i,j} = 1 + \log(t_{i,j})$$

и глобальным весом на основе IDF:

$$G_i = 1 + \log\left(\frac{N}{n_i}\right),$$

где N число фрагментов в тексте, а n_i – число фрагментов, содержащих i-ю n-грамму. При этом используется косинусная нормализация:

$$N_i = 1/\sqrt{\sum_{i=1}^{m}(L_{i,j}G_i)^2}.$$

3.3. Выделение ключевых слов

В предложенном подходе каждую найденную тематику можно описать набором ключевых n-грамм,

поэтому ключевым признакам будут соответствовать не слова текста на исходном языке, а n-граммы. Очевидно, что набор ключевых n-грамм текста является неприменимым и с точки зрения понимания человеком, и с точки зрения формирования из них поискового запроса. Поэтому необходимо от n-грамм перейти к ключевым словам. Для выделения ключевых слов текста по его сформированной тематической модели (т.е. по совокупности матриц W_k и H_k) было предложено каждое слово z анализируемого документа сначала представить в виде вектора в пространстве n-грамм данного документа: $z = [z_{1,\ i}, z_{2,\ i}, \ldots, z_{m,\ i}]^T$. Таким образом, множество слов текста представляется также в виде матрицы Z, строки которой соответствуют n-граммам, а столбцы – словам. Затем матрица Z отображается в сформированное пространство тематик путем ее умножения слева на транспонированную матрицу W_k. Результатом данного перемножения будет матрица $C = W_k^T \cdot Z$, которая соответствует представлению множества слов документа в пространстве тематик данного документа, задаваемом матрицей W_k. Тогда для формирования набора ключевых слов последовательно для каждой строки (тематики) матрицы C выбираются p слов с индексами, соответствующими максимальным значениям нормы этих элементов в рассматриваемой строке (в работе в качестве нормы вектора использовался поэлементный максимум). При этом возникает ряд проблем. (1) Длинные слова получают больший вес, так как имеют больше непустых n-грамм. (2) Разные словоформы одного и того же слова отбираются как кандидаты на ключевое слово, поскольку подход языково-независимый, и все они имеют близкий вес. (3) Зачастую, несмотря на ортогональность тематик (за счет ОНМФ), одни и те же слова оказываются кандидатами для разных тематик. Это происходит вследствие того, что, хотя ключевые n-граммы не пересекаются по темам, но могут входить в одни и те же слова. Для решения первой проблемы в работе были исследованы разные схемы нормировки с учетом длины слова и весов тематик. И был выбран простой вариант нормировки на длину слова в символах, таким образом среди различных словоформ одного и того же слова приоритет получали наиболее короткие. Для решения второй проблемы применяется языково-независимый метод кластеризации слов на основе алгоритма DSCAN [29] с использованием строчного расстояния Левенштейна [30] между словами, который позволяет находить одну, наиболее общую словоформу (наиболее близкую всему кластеру схожих слов). Для решения третьей проблемы применен подход, основанный на расчете для каждого кандидата на роль ключевого слова относительного веса, т.е. веса слова K в тематике j (максимальный среди всех тематик), деленного на вес слова K в тематике i (второй после максимального, наибольший вес по тематикам). В результате отбираются слова-кандидаты с наибольшим таким отношением, т.е. наиболее характерные только для одной тематики.

3.4. Фильтрация шума и автоматическое аннотирование

Другая важная задача связана с удалением информационного шума из документа. Она заключается в оценке значимости (релевантности) отдельных фрагментов текста и последующего удаления из результирующего документа наименее значимых фрагментов. При этом оставляемые фрагменты должны описывать все главные темы исходного текста. Рассматриваемая задача тесно связана с задачей автоматического аннотирования, для решения которой необходимо сформировать аннотацию, состоящую из наиболее значимых фрагментов текста. На сегодняшний день наиболее популярные методы автоматического аннотирования, которые вычисляют релевантность фрагментов текста, основаны на тематическом моделировании текстов с использованием латентно-семантического анализа, примененного к коллекции отдельных фрагментов (например, предложений) анализируемого документа [27, 28]. В текущей работе используется ранее предложенный авторами метод вычисления релевантности фрагментов текста, основанный на оценке весов тематик в нормализированном пространстве тематик, получаемом с помощью неотрицательной матричной факторизации [27]. Первый этап предложенного метода состоит в нормировке пространства тематик, т.е. в приведении длин вектор-столбцов матрицы W_k к единице. Этот шаг обусловлен тем, что неотрицательная матричная факторизация дает не единственное решение [11]. Второй этап заключается в оценке глобальных весов полученных тематик во всем документе. Каждая из строк матрицы H_k соответствует вектору, показывающему, насколько сильно представлена соответствующая тематика в каждом из фрагментов. Следовательно, чем больше длина вектор-строки матрицы H_k, тем соответствующая тематика больше представлена во всем документе. Поэтому предложено оценивать глобальный вес выделенных тематик как норму соответствующих вектор-строк матрицы H_k. На основе полученных оценок глобальных весов тематик документа релевантность j-го фрагмента текста R_j вычисляется как норма вектора, являющегося результатом поэлементного умножения вектора глобальных весов тематик и вектора весов тематик в рассматриваемом фрагменте (т.е. соответствующего вектор-столбца матрицы H_k), где k – число тематик:

$$R_j = \sum_{i=1}^{k} (\|w_i\|^2 \|h_i\| h_{i,j}).$$

Таким образом, идея предложенного метода удаления информационного шума заключается в выделении основных тематик в тексте документа и нахождении предложений (фрагментов) текста, которые наилучшим образом описывают выделенные тематики, путем расчета их релевантности. Для составления результирующего документа выбираются предложения с наибольшими значениями полученной релевантности, чтобы их суммарная длина не превышала заданного количества слов.

Релевантность фрагмента текста показывает его информационную значимость в рассматриваемом документе, поэтому релевантность можно рассматривать в качестве оценки количества информации, содержащейся в данном фрагменте. Исходя из этого, можно определить минимальное число фрагментов текста, требующееся для покрытия заданного процента содержащейся в тексте информации. Для построения результирующего документа, не содержащего информационного шума, выбираются его фрагменты с максимальными релевантностями, сумма которых не превышает заданный процент информации (в данной работе использовался порог 30%).

3.5. Вычисление релевантности документов выдачи тематикам документа-образца

Финальным этапом работы предложенного метода поиска информации по образцу является сортировка полученной поисковой выдачи по степени близости тематической направленности документов выдачи тематикам документа-образца. Сама выдача формируется в результате выполнения запроса, состоящего из выявленных ключевых слов, к внешней поисковой машине. Вычисление релевантности для каждого документа в поисковой выдаче основано на его представлении в пространстве тематик документа-образца и реализовано путем выполнения следующих шагов:

1. Представление найденного документа в виде матрицы B размерности число фрагментов на число n-грамм.

2. Проецирование матрицы найденного документа B в пространство тематик документа-образца. Для этого матрицу B умножают слева на транспонированную матрицу W_k. Результатом данного перемножения будет матрица $C = W_k^T \cdot B$, которая соответствует представлению найденного документа в пространстве тематик документа-образца, задаваемом матрицей W_k. Следует отметить, что ортонормированность разложения как раз необходима, чтобы получить возможность выполнить эту операцию – проекцию произвольного документа в пространство тематик документа-образца.

3. Расчет релевантности нового документа как нормы (в данной работе использовалась норма Фробениуса) матрицы C: чем больше $\|C\|_F$, тем сильнее документ выдачи соответствует тематикам документа-образца.

Для проверки предложенного подхода было проведено экспериментальное исследование на подготовленном тестовом наборе данных, состоящем из сообщений из набора данных джихадистского форума Ansar1 в качестве “плохих” примеров и данных из тематик набора 20 Newsgroups [41] на политические и религиозные темы (talk.politics.misc, talk.politics.guns, talk.politics.mideast, talk.religion.misc, alt.atheism, soc.religion.christian) в качестве нейтральных примеров. С применением описанных выше методов поиска по образцу были получены высокие оценки точности распознавания, а именно, средняя точность (average precision) на уровне 0.95 и выше [22].

4. МОНИТОРИНГ И ТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ ПОТОКОВ ТЕКСТОВЫХ СООБЩЕНИЙ

Следующей задачей после обнаружения источников потенциально террористической или экстремистской информации является постановка этих источников на контроль – мониторинг. В рамках этого контроля необходимо периодически скачивать новую информацию из этих источников, строить описательные модели, по которым можно понять, что именно и как интенсивно обсуждается. Также полезно строить прогнозные модели, которые позволяли бы оценить, какая тема находится на восходящем тренде обсуждения, а какая теряет свою актуальность. Важно уметь находить и анализировать исключения – сообщения, тематика которых существенно отличается от предыдущих, а также находить случаи изменения тренда, т.е. ситуации, когда прогноз активности какой-либо тематики существенно отличается от наблюдаемой.

4.1. Представление тематик потока текстовых сообщений в виде многомерного временного ряда

Для решения этих задач реализованы алгоритмы, которые позволяют для выбранного множества сетевых пользователей или групп за заданный период времени определить обсуждаемые темы, а также визуализировать динамику активности этих тем и получить описание этих тем в виде набора ключевых слов. Все это делается также с использованием языково-независимого подхода – методов латентно-семантического анализа, основанных на ОНМФ, но в другой постановке по сравнению с разделом 3. В данном случае интерес представляют

не тематики отдельных документов и их фрагментов, а поток сообщений в целом, причем опять же в условиях наличия проблем, связанных с лексикой и качеством текстов, включая проблему замаскированных ключевых слов.

В работе предлагается модель представления потока текстовых сообщений в виде многомерного временного ряда, в котором каждая компонента показывает изменение веса тематики во времени. Сама идея использовать многомерные временные ряды тематик для анализа потока текстовых документов не нова [19]. Но в настоящей работе характерные тематики потока выявляются с использованием методов ортонормированной неотрицательной матричной факторизации матрицы всех документов (не фрагментов) в пространстве n-грамм. Такой подход раньше не применялся, и он имеет ряд особенностей, связанных с возможностью строить интерпретируемые модели.

Формально поток текстовых сообщений представляется множеством пар $x = \{(d_1, t_1), ..., (d_n, t_n)\}$, где каждый элемент (d_j, t_j) – анализируемый объект $(1 \le j \le n)$, где d_j – документ, содержащий текстовые данные, t_j – временная метка. Для формирования тематической модели по потоку его текстовые данные описываются набором признаков, изменения значений которых определяет изменение набора обсуждаемых тем с течением времени. Коллекция документов $(d_1, ..., d_n)$ представляется в виде числовой матрицы $D \in \mathbb{R}^{m \times n}$, строки которой соответствуют признакам (n-граммам), а столбцы – документам. Тогда матрица D задает модель поведения потока сообщений в виде m-мерного временного ряда, показывающего изменение весов соответствующих признаков в потоке сообщений. Пространство признаков при использовании модели n-грамм имеет высокую размерность, которая равна числу различных n-грамм во всей коллекции. Кроме того, в таких признаках игнорируются семантические взаимосвязи между n-граммами, а матрица D содержит большое число строк и является разреженной. Последнее приводит к бессмысленности применения методов прогнозирования временных рядов для строк матрицы D. Использование тематических моделей представления документов позволяет уменьшить пространство признаков за счет объединения разных, но семантически связанных n-грамм в один признак – тематику. Полученная таким образом тематическая модель, сформированная по потоку текстовых документов $\{(d_1, t_1), ..., (d_n, t_n)\}$, представляет собой совокупность (L_m, W_k, H_k, T_n), где (L_m, W_k, H_k) – тематическая модель коллекции документов $(d_1, ..., d_n)$, $T_n = (t_1, ..., t_n)$ – временные метки документов. Столбец H_j $(1 \le j \le n)$ матрицы H_k соответствует тематической направленности в момент времени t_j. Требование ортонормированности W_k является необходимым, так как тематическая модель потока сообщений может применяться для анализа дальнейшей активности в потоке $y = \{(d_{new}, t_{new})\}$, где d_{new} – новое сообщение, а $t_{new} > t_n$. Поэтому необходимо иметь возможность представлять новые документы d_{new} в уже сформированном тематическом пространстве (L_m, W_k): $H_{new} = W_k^T \cdot A_{new}$, где A_{new} – вектор представления d_{new} со словарем n-грамм L_m.

Для интерпретации получаемых моделей, помимо возможности визуализировать поведение тематики с помощью графиков временного ряда, необходимо обеспечить эксперту возможность понимать, о чем именно идет речь в найденной тематике. А это возможно сделать только с помощью выделения ключевых слов или фраз для тематики. Поэтому возникает задача генерации языково-независимых ключевых слов, характерных для каждой из тематик. Для ее решения также используется подход, описанный в пункте 3.3, который основан на проецировании наиболее часто встречающихся в анализируемом наборе сообщений слов в пространство тематик и выборе в качестве кандидатов слов с наибольшим весом в тематике.

4.2. Применение методов анализа временных рядов для прогнозирования поведения тематик потока сообщений и поиска исключений

Как показали экспериментальные исследования на эталонных наборах данных DarkWeb и на реальных собранных данных, использование для поиска корреляций, автокорреляций и прогнозирования тематических временных рядов, получаемых напрямую с помощью методов латентно-семантического анализа, затруднительно, поскольку такие ряды оказываются зашумленными, нестационарными, с “пропущенными” периодами (когда обсуждение темы в сообществе временно полностью прекращается, а потом возобновляется). Поэтому было принято решение анализировать сглаженные тематические ряды. Были рассмотрены методы экспоненциального сглаживания, регуляризированные сплайны и непараметрическая взвешенная локальная регрессия (loess). Последняя показала наилучшие результаты при выборе параметра регуляризации для loess по скорректированному информационному критерию Акаика (AICC) [31]. Для поиска коррелирующих тематик были реализованы алгоритмы расчета близости сглаженных тематических временных рядов на основе евклидового расстояния и расстояния на основе динамического выравнивания временной шкалы DTW (dynamic time wrapping) [32]. Последнее позволяет находить похожие по поведению во времени тематики с учетом “сжатия и растяжения” соответствующих временных

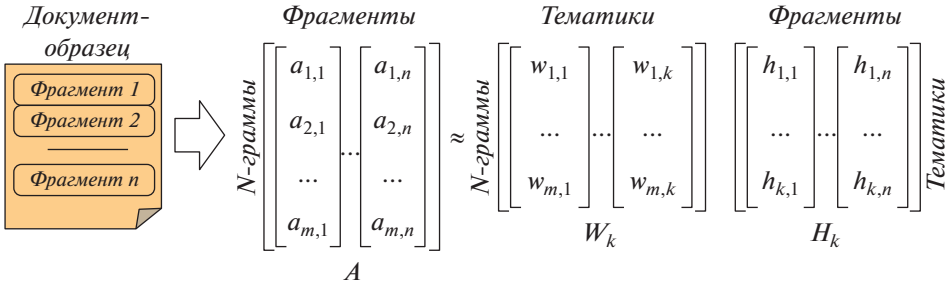

Рис 3. Матричное представление документа.

рядов, оно показало наиболее интересные результаты на практике.

Были протестированы алгоритмы прогнозирования изменения тематик в сообществе на основе прогнозирования сглаженного многомерного временного ряда соответствующих тематик с помощью линейной ARIMA [33]. Экспериментальное исследование на наборе данных DarkWeb показало достаточно высокую точность прогнозирования на маленький горизонт (несколько дней), но плохую точность на длительные периоды. Это, возможно, связано с тем, что в эксперименте не было данных для длительного периода, а также с тем, что, судя по всему, тематики в потоке текстовых сообщений существуют не очень долго и достаточно быстро полностью меняются.

Были разработаны два новых алгоритма обнаружения аномального содержания в потоке сообщений, использующие предложенное тематическое представление потока текстовых документов. В первом алгоритме рассматривается пользовательский поток документов $x = \{(d, t)\}$, где документ d представляет собой объединенные текстовые данные пользователя, к которым он обращался за время $[t, t + t_0)$, при этом t_0 выбирается достаточно "длительным". По потоку x строится тематическая модель поведения пользователя (L_m, W_k, H_k, T_n). Тогда поток x можно представить в виде множества упорядоченных пар $F = ((H_1, t_1), ..., (H_n, t_n))$, где $t_1 < t_2 < ... < t_n$. Выборка F рассматривается как k-мерный временной ряд, по которому строится прогноз (в работе использовалась линейная ARIMA) на следующие p шагов: $((H^f_{n+1}, t_{n+1}), ..., (H^f_{n+p}, t_{n+p}))$. После чего определяется решающая функция:

$$f((d, t_{n+j}), (L_m, W_k)) = \|H^f_{n+j} - h\|_1 = a_j,$$

где h – представление документа d в (L_m, W_k), a_j – уровень аномальности контента d за время $[t_{n+j}, t_{n+j+t0})$. Здесь уровень аномальности соответствует отклонению тематической направленности от ожидаемых значений, поэтому чем меньше значение a_j, тем менее аномально содержание сообщений пользователя. Во втором подходе к обнаружению аномального содержания сообщений пользователя используется оценка принадлежности отдельных документов к характерным тематикам анализируемого потока. Рассматривается поток $x = \{(d, t)\}$, где документ d соответствует тексту, который сгенерировал или прочитал пользователь в момент времени t. По потоку x строится тематическая модель (L_m, W_k, H_k, T_n). Вычислять оценку степени принадлежности произвольного документа d' к тематикам пользователя предложено как норму вектора h', являющегося представлением документа d' в (L_m, W_k), т.е. с помощью решающей функции

$$f(d', (L_m, W_k)) = \|h'\| = b,$$

при этом чем сильнее документ d' соответствует характерным тематикам анализируемого потока, тем меньше значение нормы и, значит, менее аномально сообщение d' данного пользователя.

4.3. Экспериментальное исследование предложенного подхода

Для демонстрации предложенного подхода рассмотрим следующий сценарий. В качестве документа-образца была взята статья от 22 февраля 2017 года с запрещенного экстремистского сайта kavkazcenter.com под название "Дорогой друг Эрдогана помог Асадитам войти в Африн", где критикуются действия РФ в Сирии и выражается поддержка террористам, противостоящим этим действиям. По этой статье с помощью методов из раздела 3 были выявлены следующие ключевые слова: асадитам, курдские, асаду, турецких, эрдогана, русские, вошли, африн, войск, коммунистов, курдам, спину, заявил. Очевидно, что они не являются специфической экстремистской лексикой и по ним можно найти много легитимной информации. На основе этих слов был сформирован поисковый запрос в сеть "Вконтакте" (vk.com), а полученная выдача была обогащена и ранжирована с помощью методов раздела 3. В результате в числе наиболее релевантных оказались сообщения ряда подозрительных пользователей с име-

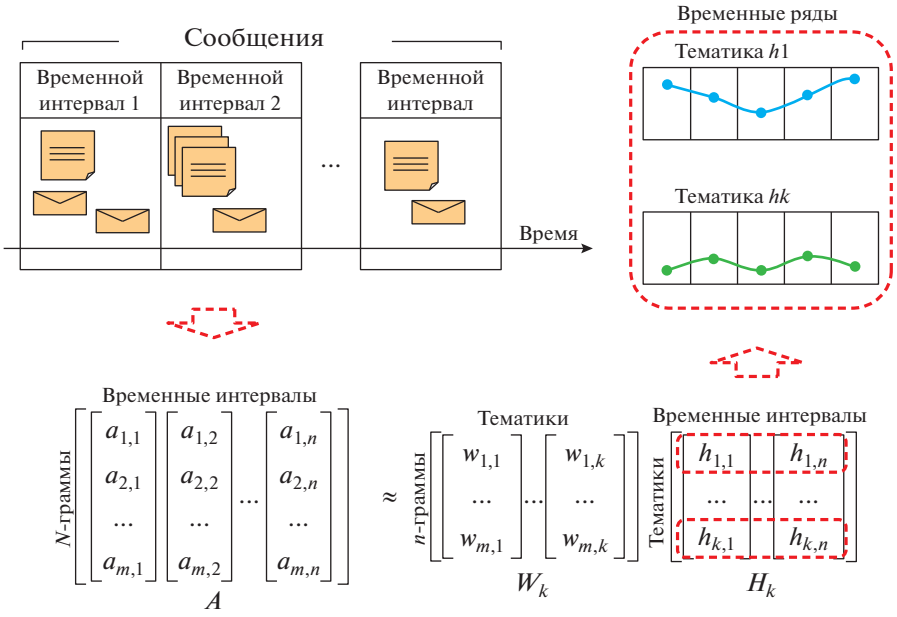

Рис 4. Модель представления содержания потока сообщений как многомерного временного ряда весов тематик.

нами: natisc_na_ivanov, islamword1071, club68716929. Визуальный анализ их страниц в vk подтвердил предположения, что они распространяют экстремистскую информацию, в том числе разжигают рознь по национальному и религиозному признаку. В рамках мониторинга из этих аккаунтов vk были выкачены сообщения за период с 1 ноября 2017 по 22 февраля 2018, к ним были применены методы тематического анализа и построения соответствующих временных рядов с тремя тематиками и прогнозом, описанными в предыдущем разделе. Результат приведен на рис. 5.

На рис. 5 цветом показано поведение и прогноз трех выявленных тематик, обсуждаемых этими пользователями. Прерывистые линии – веса тематик во времени, сплошные линии – сглаженные (по loess) значения тематик. Список ключевых слов первой (синей в цветном варианте рисунка) тематики содержит слово “Сирия”, второй (коричневой в цветном варианте рисунка) содержит слово “Украина” и третьей (зеленой в цветном варианте рисунка) слово “США”. При этом, согласно прогнозу, интенсивность обсуждения Сирии далее будет также возрастать, для обсуждения Украины ожидается плавный рост, обсуждение США останется на том же уровне.

Если применить алгоритм поиска исключений, описанный в разделе 4.2, то самым нетипичным окажется документ из зеленой тематики (про США), где просто перепечатана статья с выступлением президента США Трампа про обращение к Северной Корее. И действительно, с точки зрения остального потока сообщений, где преобладает антироссийская пропаганда, указанная обычная новостная статья действительно выглядит исключением.

5. ПРОГНОЗИРОВАНИЕ УГРОЗ, ИСХОДЯЩИХ ОТ ПОЛЬЗОВАТЕЛЕЙ ИНТЕРНЕТ СООБЩЕСТВА

На сегодняшний день большинство подходов для оценки угрозы и прогнозирования рисков, исходящих от отдельных пользователей сетевого сообщества, используют в качестве источника информации генерируемый членами социальных структур контент. Но в реальности такие данные могут быть недоступны, потому что пользователи социальных сетей, форумов, чатов могут использовать приватный или закрытый режим общения, недоступный постороннему наблюдателю. Такие режимы поддерживаются в большинстве серверов современных социальных сетей и Интернет форумов. Наиболее известным примером популярного среди террористов и экстремистов мессенджера является Телеграм [34], в котором присутствует возможность общения без последующего раскрытия текста переписки.

Рассмотрим множество пользователей, для которых известны их связи (через подписки, репосты, общие группы, ответы в ветках или просто через адресную книгу). Это множество можно представить в виде графа, где вершины – пользователи, а ребра – связи, направленные, и, возможно, с весами. Множество пользователей можно разбить на три группы:

• “неопасные”, те, про кого известно, что они не связаны с радикальными структурами, не ге-

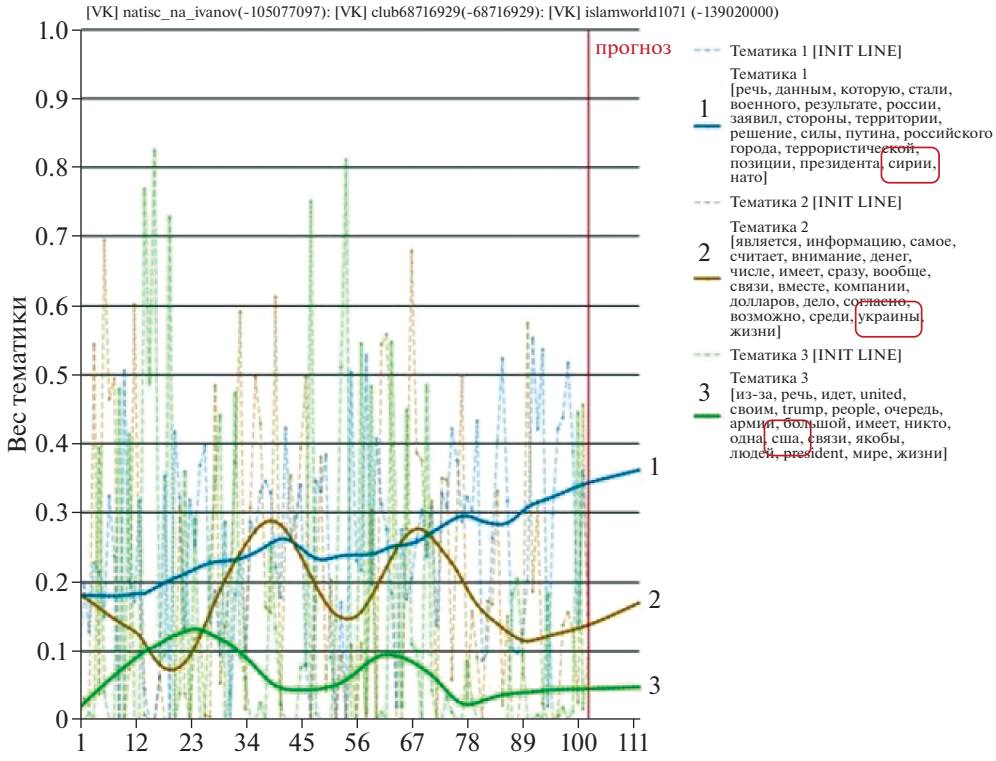

Рис 5. Скриншот визуализации выявленных тематик, их ключевых слов и прогноза для выбранных трех экстремистских аккаунтов в сети vk.com.

нерируют и не распространяют экстремистский контент;

• “опасные”, которые распространяют террористическую или экстремистскую информацию, или из других, не обязательно электронных, источников известно об их членстве в экстремистской организации;

• “неизвестные” – нет данных об их активности и переписке, они представляют наибольший интерес.

Задача состоит в том, чтобы с помощью методов машинного обучения построить модель, которая будет прогнозировать, с какой вероятностью и к какому из двух классов (опасный или нет) будет принадлежать “неизвестный” пользователь, используя в качестве свойств только характеристики узлов в графе. Для прогнозирования предлагается использовать стандартные популярные алгоритмы машинного обучения, поэтому данный раздел посвящен, в основном, вопросу формирования подходящего для рассматриваемой задачи признакового пространства. Генерация признакового пространства и построение классификатора радикальных пользователей позволит прогнозировать угрозу, исходящую от пользователя. Под уровнем угрозы будем понимать вероятность того, что пользователь опасен, т.е. генерирует или просматривает информацию экстремистского или террористического содержания. Для этого рассматривается множество пользователей в качестве вершин графов социальной структуры, а признаковое пространство (предикторы) для них формируется на основе только характеристик графа, исключая из анализа всю информацию о текстах сообщений.

5.1. Подготовка тестового набора данных

Стоит отдельно отметить, что качественных эталонных наборов данных (чатов/форумов/социальных сетей с размеченными с точки зрения опасности пользователями) для проверки предлагаемых для решения этой задачи алгоритмов найдено не было. Из используемых наборов данных в открытом доступе есть данные с соревнований Kaggle [42, 43], в которых отсутствует информация о связях пользователей, а разметка пользователей осуществляется очень неточно, по “экстремистским” хэштегам и ключевым словам. В силу указанных особенностей такие наборы не применимы для экспериментальной оценки моделей прогнозирования опасности пользователя в текущей постановке задачи. Поэтому в настоящей работе для выявления радикальных пользователей и построения связей между пользователями использовался набор данных, самостоятельно сформированный на основе набора KavkazChat. Подготовка набора данных

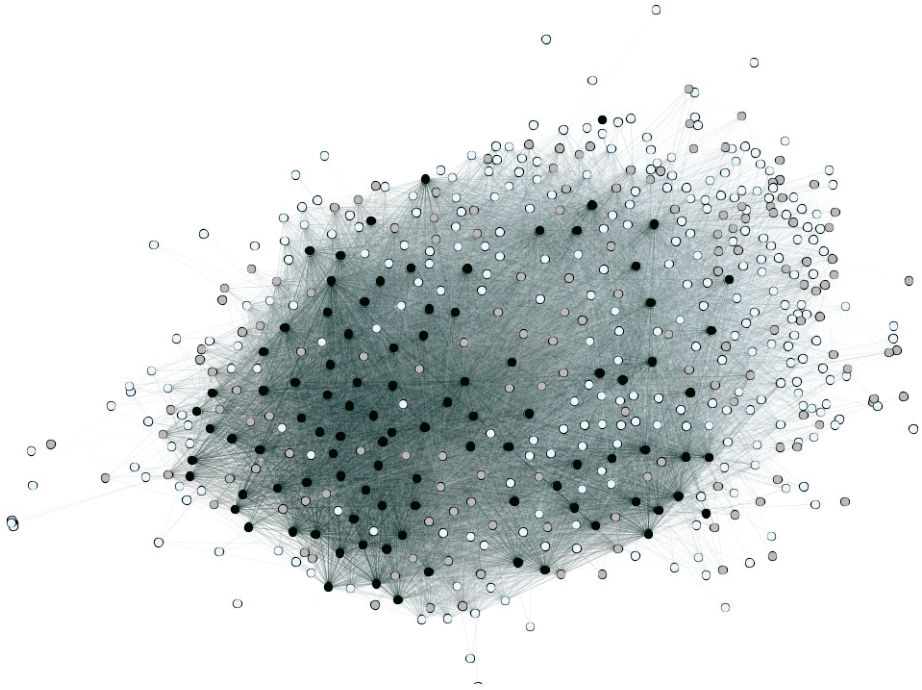

Рис 6. Пример визуализации графа сообщества по набору KavakazChat (черные – опасные, белые – нет, серые – неизвестные).

делится на следующие этапы: разметка пользователей и создание графа. В первой части применяется простой интерактивный способ обнаружения пользователей, чьи сообщения содержат опасный контент, что позволяет поставить метку “опасный”/”неопасный” пользователям социальной сети, основываясь на генерируемом ими контенте. Но при этом далее в экспериментах по классификации пользователей с неизвестной меткой контент не используется. Таким образом, для части пользователей будет ставиться метка “неизвестно”, и алгоритм классификации должен будет ее спрогнозировать уже без использования контента. Для оценки точности будем сравнивать метку, спрогнозированную по графовым признакам, с исходной меткой, полученной на основе анализа контента. Разметка пользователей на “опасный”/“неопасный” производилась с использованием языково-независимой модели представления текстовых сообщений на основе *n*-грамм, так как анализ именно символьных последовательностей хорошо зарекомендовал себя в вышеописанных задачах тематического моделирования. Для получения менее разряженных векторов признаков меньшей размерности при разметке использовался подход из области латентно-семантического анализа, но на основе нейронных сетей, а не ОНМФ. Был создан нейросетевой автоэнкодер (autoencoder) [40], у которого все веса – неотрицательные, нулевые смещения и ReLU-активации на каждом слое. Число нейронов на каждом слое было следующим N, 500, 100, 10, где N – размерность входного вектора представления текстов через *n*-граммы. Для настройки весов использовался алгоритм Adam [40] и среднеквадратичная ошибка (MSE) в качестве функции потерь.

Далее был сформирован набор текстов из всех сообщений форума KavkazChat с помощью простого поиска по ключевым словам – сообщения, содержащие слова “джихад”, “призываю”, “война”, “кафир”, “моджахед” и ряд других. Таких было найдено 248. Среди них вручную было отобрано первые 10 сообщений, гарантировано экстремистского содержания. После чего использовался Алгоритм 1. На каждой из его итераций анализировались 100 наиболее релевантных текстов, и в результате его работы менее чем за 6 итераций стало возможным определение 171 сообщения с радикальным содержанием, а также 135 пользователей (1.8% всех пользователей форума), генерирующих соответствующий контент. На каждой итерации случайно выбирались неразмеченные сообщения из всего корпуса таким образом, чтобы пропорция выделенных опасных к неразмеченным составляла 5 к 95. Затем полученный набор данных делился на 5 стратифицированных частей (5-fold cross-validation), и оптимизировались параметры метода опорных векторов с RBF ядром. Константа регуляризации, ширина ядра и другие параметры отбирались с помощью кроссвалидации выборки, максимизируя среднее значение меры F1-меры (гармоническое среднее

между точностью и полнотой). После этого обучался классификатор на всем наборе данных (с пропорцией отклика 5%), с использованием набора найденных кросс-валидацией лучших гиперпараметров, а применялся обученный классификатор для разметки всего множества сообщений набора KavkazChat. В качестве признаков использовались выходы среднего слоя автоэнкодера (аналог весов скрытых тематик) для каждого текста. Далее сообщения ранжировались по убыванию соответствия классу с “радикальными” текстами, и вручную выбирались среди ста наиболее потенциально радикальных только те, которые содержали опасный контент, за исключением сообщений пользователей, которые уже были помечены как “опасные” (сгенерировали хотя бы одно сообщение, которое уже было помечено как радикальное). Из отсортированного списка сообщений выбирались первые 100, которые размечались вручную, после чего добавлялись во множество опасных сообщений для дообучения, и описанная выше процедура переходила на следующую итерацию. В итоге было получено множество опасных пользователей – тех, кто сгенерировал хотя бы одно сообщение с радикальным содержанием.

Алгоритм 1. Выделение радикальных пользователей

Вход:

- M_D: множество опасных сообщений;
- M_{ALL}: все сообщения;
- USERS: множество пользователей;
- M_I: множество сообщений пользователя I;
- N: количество итераций;
- K: количество сообщений для ручной разметки на каждой итерации;

Выход:

- M_D: обновленное множество опасных сообщений.

1: Для всех i от 1 до N

2: $M_R = |M_R|$ произвольных сообщений из M_{ALL}, где $|M_R|/|M_D|=95/5$

3: Найти классификатор с оптимальным набором параметров с помощью, стратифицированной кросс-валидации на $M_R + M_D$

4: Обучить классификатор на наборе данных $M_R + M_D$

5: $M_E = \{\}$ – множество сообщений для исключения из процесса ручной разметки

Для всех USER из USERS:

Если пересечение (M_{USER}, M_D) != {}

То добавить M_{USER} к M_E

6: $M_A = M_{ALL} - M_E$ – множество сообщений для анализа

7: P = предсказать M_A с помощью классификатора и извлечь топ K релевантных текстов (ближайших к классу “радикальный”)

8: Вручную разметить P, после чего добавить релевантные сообщения в множество M_D

9: Вернуть M_D

Для того чтобы сформировать сетевую структуру социальной сети и представить ее в виде графа был использован Алгоритм 2. В этом алгоритме сначала инициализируется полносвязный ориентированный граф со всеми пользователями в качестве вершин и нулевыми весами ребер. Если пользователь A имел некоторую активность “в направлении” пользователя B (поставил лайк, сделал репост записи, написал сообщение и т.д.), то вес ребра, исходящего из вершины A в вершину B, увеличивается на единицу. Все ребра с нулевыми весами удаляются, а ненулевой вес ребра инвертируется. Это означает, что если пользователь A имел x “активностей” в сторону B, то вес соответствующего ребра из A в B считается равным $1/x$. Таким образом вычисленный вес ребра может быть использован в качестве меры “расстояния” между пользователями – чем “больше” активность, тем “ближе” пользователи. Такой подход оказался более эффективным при вычислении кратчайших путей между вершинами графа, подсчет пути без учета весов дуг. В экспериментальном исследовании также использовались графы с бинарными весами, в которых ненулевые веса заменялись на единицы. Для определения активности может быть использован большой набор взаимодействий: добавление в “друзья”, “репосты”, “лайки”, просмотры страницы, число

общих групп, количество входящих/исходящих сообщений, ответы в ветке форума и т.д. Кроме того, необходимо учитывать, что связи могут быть двунаправленными, а также для одной социальной сети можно построить несколько графов (свой для каждого типа активности).

Алгоритм 2. Создание графа социальной сети

Вход:

• USERS – множество пользователей социальной сети;

• W_{ij} – матрица, элементы которой – число “активностей” от пользователя i к j;

Выход:

• G: граф пользовательских связей;

```
1: G = полносвязный ориентированный граф с нулевыми весами
2: Для всех вершин i из USERS
3:    Для всех j из USERS/{i}
4:          если W_ij != 0
5:                то установить вес 1/W_ij ребру из i в j в графе G
6: Удалить ребра с нулевыми весами в G
7: Вернуть G
```

В ходе экспериментов было обнаружено, что многие признаки, которые рассчитываются с учетом направления ребер в графе, лучше отражают связь между пользователями, если это направление поменять на обратное. По построению графа связь из A в B имеет семантику “A что-то сделал с контентом B”. Обратная связь имеет семантику “B заинтересовал A”, и как раз на таких обратных связях экспериментальный результат оказался лучше.

Для набора данных KavkazChat были построены следующие графы пользовательских связей, по которым вычислялись признаки:

• ориентированный граф со связями только по репостам сообщений без весов у связей;

• ориентированный граф со связями только по репостам сообщений с весами у связей, обратно пропорциональными интенсивности (числу действий);

• ориентированный граф со связями только по ответам в ветке без весов у связей;

• ориентированный граф со связями только по ответам в ветке с весами у связей, обратно пропорциональными интенсивности (числу действий);

• ориентированный граф с обратными связями только по репостам сообщений без весов у связей;

• ориентированный граф с обратными связями только по репостам сообщений с весами у связей, обратно пропорциональными интенсивности (числу действий);

• ориентированный граф с обратными связями только по ответам в ветке без весов у связей;

• ориентированный граф с обратными связями только по ответам в ветке с весами у связей, обратно пропорциональными интенсивности (числу действий).

5.2. Построение пространства признаков

В работе рассматривались следующие признаки узлов, применяемые в анализе сетевых структур:

• “важность” узла по PageRank [35];

• “уровень авторитета” и “уровень посредника” (hub) по HITS (Hyperlink Induced Topic Search) [36];

• betweenness centrality [37] – доля кратчайших путей между любыми двумя вершинами, которые включают в себя данную вершину;

• proximity prestige [37] – нормализованное среднее расстояние от текущей вершины до всех достигаемых вершин;

• sociability [37] – число входящих в вершину ребер, деленное на число всех ребер, связанных с данной вершиной;

• три бинарных признака – наличие в ближайшем окружении хотя бы одного “опасного”/“неопасного”/“неизвестного” узла;

• три числовых признака – число “опасных”, число “неопасных” и число “неизвестных” ближайших соседей;

• шесть числовых признаков – минимальное расстояние с учетом весов ребер и без до ближайшего “опасного”/“неопасного”/“неизвестного” узла;

• три числовых признака – доля “опасных”/“неопасных”/“неизвестных” узлов среди всех ближайших соседей.

Для каждого пользователя считались все указанные выше признаки по всем типам сформированных графов. Например, для пользователя A отдельно его признаком становился PageRank,

рассчитанный по графу с прямыми связями, сформированными по репостам, отдельно признак PageRank по графу с прямыми связями, сформированными по ответам в общей ветке, отдельно признак PageRank по графу с обратными связями, сформированными по репостам и так далее.

5.3. Экспериментальное исследование

Для эталонного набора данных с разметкой пользователей, описанной в разделе 5.1, и с признаковым пространством, описанным в разделе 5.2, была проведена следующая серия экспериментов. В каждом эксперименте из исходного набора формировался тренировочный и тестовой в пропорции 70 к 30. На каждом таком наборе обучались следующие классификаторы: логистическая регрессия с L2 регуляризацией, дерево решений типа CART [44], случайный лес [45], ансамбли деревьев решений на основе градиентного бустинга XGBoost [38] и LightGBM [39]. При обучении подбирались оптимальные параметры каждого из алгоритмов с помощью стратифицированной кросс-валидации на тренировочном наборе. Для логистической регрессии и градиентных бустингов подбирался оптимальный параметр L2-регуляризации признаков на сетке возможных параметров. Для алгоритмов на основе деревьев решений выбирались максимальная глубина деревьев, минимальное число листьев в вершине каждого дерева, для ансамблей – оптимальный размер ансамбля. Для того чтобы избежать перебора всевозможных комбинаций параметров и эффективно выбирать очередную точку в пространстве гиперпараметров использовался подход последовательной оптимизации SMBO (Sequential Model-Based Optimization) [46].

Для сравнения моделей использовалась площадь под ROC кривой – AUC. По результатам серии экспериментов оценивалось среднее значение и разброс AUC.

Как видно из приведенных в таблице 1 результатов ни одна из моделей не получилась переобученной, и наилучшее качество показали современные бустинг ансамбли с регуляризацией. Отдельно была проведена проверка предположения о том, возможно ли классифицировать пользователей потенциально террористического или экстремистского сообщества, если вообще нет информации об “опасных” соседях, исходя только из признаков, рассчитанных без учета метки класса узлов-соседей. Это значит, что при расчете признаков по методу, изложенному в разделе 5.2, считалось, что у всех узлов класс – “неизвестный пользователь”. Результаты (представлены в таблице 2) получились неожиданно удачными и в общем близкими к результатам на всем признаковом пространстве.

Таблица 1. Сравнение классификаторов на всех признаках в предложенном признаковом пространстве

Классификатор	*Train AUC*	*Test AUC*
Дерево решений	0.915 ± 0.005	0.912 ± 0.031
Логистическая регрессия	0.934 ± 0.010	0.920 ± 0.038
Random Forest	0.938 ± 0.007	0.930 ± 0.029
LightGBM	0.946 ± 0.005	0.941 ± 0.022
XGboost	0.957 ± 0.005	**0.943 ± 0.024**

Таблица 2. Результаты работы классификатора без использования информации о взаимном расположении с “известными” опасными пользователями

Классификатор	*Train AUC*	*Test AUC*
Дерево решений	0.897 ± 0.007	0.886 ± 0.030
Логистическая регрессия	0.889 ± 0.021	0.893 ± 0.048
Random Forest	0.931 ± 0.007	0.911 ± 0.032
LightGBM	0.936 ± 0.007	0.923 ± 0.032
XGboost	0.944 ± 0.006	**0.924 ± 0.033**

Это показывает, что при наличии обученной модели, можно оценить угрозу членов полностью неразмеченного аналогичного сообщества, т.е. графа, где достоверно неизвестно ни одного опасного пользователя. Хотя, безусловно, если информация об опасных пользователях в графе есть, результаты получаются лучше.

Для того чтобы выявить влияние отдельных признаков на результат (оценить важности каждого из предикторов), была построена описательная модель на основе регуляризированной логистической регрессии. Для этого выделены наиболее важные переменные с точки зрения их абсолютных весов в регрессионной модели, изменения которых значимо влияют на шанс принадлежности к классу “опасных” пользователей (Таблица 3).

В случае использования всех переменных признакового пространства наиболее важным оказывается количество “опасных” пользователей в графе обратных связей по ответам в общие ветки. Другими словами, это может быть интерпретировано, как количество “опасных” пользователей, которые писали сообщения в той же ветке, что и данный пользователь, после него. Ну и чуть менее важными оказались переменные, отражающие количество “опасных” пользователей, которыми интересуется данный участник форума.

Если же использовать только информацию о центральностях пользователей и считать всех пользователей в анализируемом сообществе “неизвестными”, то окажется, что наиболее важными будут уровень посредника (Hub) в графе пря-

Таблица 3. Коэффициенты топ-8 переменных в модели логистической регрессии

Переменная	Значение стд. коэф.
Количество “опасных” ближайших соседей в графе с обратными связями, построенными по ответам в ветках форума	+0.334
Количество “опасных” ближайших соседей в графе с прямыми связями, построенными по ответам в ветках форума	+0.278
Количество “опасных” ближайших соседей в графе прямых связей, построенных по “репостам” сообщений	+0.199
Proximity prestige узла в графе прямых связей, построенных по “репостам” сообщений	+0.106
Количество “неизвестных” ближайших соседей в графе с обратными связями, построенными в инвертированном графе связей по “репостам”	+0.096
Уровень посредника (hub) пользователя в графе прямых связей, построенных по “репостам”	–0.184
Минимальное расстояние до “неопасного” пользователя в графе прямых связей, построенного по общим веткам	–0.130
Минимальное расстояние до ближайшего “неизвестного” пользователя в графе обратных связей, построенном по ответам в ветках форума с учетом весов ребер	–0.088

Таблица 4. Коэффициенты топ-5 переменных в модели логистической регрессии (без использования информации о классах других участников сообщества)

Переменная	Значение стд. коэф.
Уровень посредника пользователя в графе прямых связей по ответам в общих ветках	+0.187
Авторитетность пользователя в графе прямых связей по “репостам”	–0.089
Proximity prestige пользователя в графе прямых связей по “репостам”	+0.066
Proximity prestige пользователя в графе прямых связей по ответам в общих ветках	+0.060
Page Rank пользователя в графе прямых связей по “репостам”	+0.058

мых связей по ответам в общих ветках, а также Proximity Prestige значения во всех графах прямых связей. Это можно упрощенно проинтерпретировать как то, что чем больше связей у члена в целом подозрительного сообщества, тем более вероятно, что он играет в нем ключевую роль и является “опасным”.

6. ЗАКЛЮЧЕНИЕ

В настоящей работе рассматривается важная проблема – обнаружение и мониторинг источников потенциально экстремистской или террористической информации в сети Интернет, а также прогнозирование угроз, исходящих от таких источников. Особое внимание уделялось проблемам, связанным с качеством текстов, включая наличие ссылок и хэштегов, многоязыковость, использование специальных, жаргонных или “замаскированных” ключевых слов, а также наличие опечаток и грамматических ошибок, в том числе преднамеренных. Также рассматривался вопрос оценки уровня угрозы пользователя, исходя из его окружения в социальной сети. Основными результатами, изложенными в статье, являются:

• Языково-независимый подход для информационного поиска по образцу потенциально экстремистской или террористической информации в сети Интернет с учетом обозначенных выше проблем, включая качество текстовых данных и наличие ссылочного окружения. Подход основан на n-граммном представлении текстов, латентно-семантическом анализе с использованием ортонормированной неотрицательной матричной факторизации и “обогащении” текстов документов за счет ключевых слов и аннотаций, полученных из данных по web ссылкам и хэштегам, которые встречаются в документе.

• Подход на основе анализа временных рядов для мониторинга источников потенциально экстремисткой или террористической информации в сети Интернет, позволяющий представлять тематическое содержание потока текстовых сообщений в виде многомерного временного ряда тематик этого потока с автоматически определяемыми ключевыми словами, характеризующими каждую из тематик. Обсуждается возможность использования стандартных методов анализа временных рядов для прогнозирования тематической направленности потока сообщений в контролируемом источнике и для поиска исключений – нетипичных сообщений и моментов изменения тренда тематик.

• Подход, позволяющий оценивать уровень угрозы, исходящей от пользователя, позволяю-

щий прогнозировать вероятность того, что пользователь участвует в создании или обсуждении экстремистских, или террористических тем, используя при этом только данные о его сетевом окружении и структуре социального графа, и не используя контентную информацию.

СПИСОК ЛИТЕРАТУРЫ

1. Why big data analytics holds the key to tackling the changing terror threat // Journal of Advanced Analytics Intelligence quarterly, 1Q, 2015. http://www.sas.com/content/dam/SAS/en_us/doc/other1/iq-q115.pdf
2. *Hankin C.* IDEAS Factory – Detecting Terrorist Activities: Making Sense. http://gow.epsrc.ac.uk/NGBOViewGrant.aspx?GrantRef=EP/H023135/1.
3. *Nizamani S. et al.* Modeling suspicious email detection using enhanced feature selection. arXiv preprint arXiv:1312.1971. 2013.
4. *Sheehan I.S.* Assessing and comparing data sources for terrorism research. Evidence-based counterterrorism policy. Springer New York, 2012. V. 3. P. 13–40.
5. *Berger J.M., Morgan J.* The ISIS Twitter Census. The Brookings Project on US Relations with the Islamic World, Analysis Paper. 2015. №. 20. 65 p.
6. IDEAS Factory – Detecting Terrorist Activities: Making Sense. https://www.slideserve.com/fawzia/detecting-terrorist-activities-making-sense
7. Workshop on Link Analysis, Counterterrorism and Security at the SIAM International Conference on Data Mining. California, USA, 2005. http://research.cs.queensu.ca/home/skill
8. *Zhang Y., Zeng S., Fan L., Dang Y., Catherine A. Larson C.A., Chen H.* Dark web forums portal: searching and analyzing Jihadist forums. Proceedings of the 2009 IEEE international conference on Intelligence and security informatics (ISI'09). IEEE Press, Piscataway, NJ, USA. 2009. P. 71–76.
9. *Abbasi A., Chen H.* Applying authorship analysis to extremist-group web forum messages, IEEE Intelligent Systems. 2005. V. 20. P. 67–75.
10. *Sebastián A.* Ríos and Ricardo Muñoz. 2012. Dark Web portal overlapping community detection based on topic models. Proceedings of the ACM SIGKDD Workshop on Intelligence and Security Informatics (ISI-KDD '12). ACM, New York, NY, USA, Article 2, 7 pages.
11. *Kuang D., Choo J., Park H.* Nonnegative Matrix Factorization for Interactive Topic Modeling and Document Clustering. Partitional Clustering Algorithms. Springer International Publishing, 2015. P. 215–243.
12. *Tsarev D.V., Petrovskiy M.I., Mashechkin I.V.* Using NMF-based text summarization to improve supervised and unsupervised classification. Hybrid Intelligent Systems (HIS). 11-th IEEE International Conference on Application of Information and Communication Technologies. 2011. P. 185–189.
13. *Elovici Y., Shapira B., Last M., Zaafrany O., Friedman M., Schneider M. Kandel A.* Detection of access to terror-related Web sites using an Advanced Terror Detection System (ATDS). Journal of the American Society for Information Science and Technology. 2010. V. 61. P. 405–418.
14. *Agarwal S., Sureka A.* Applying Social Media Intelligence for Predicting and Identifying On-line Radicalization and Civil Unrest Oriented Threats. November 2015. arXiv:1511.06858 [cs.CY]
15. *Badia A., Kantardzic M.* Link Analysis Tools for Intelligence and Counterterrorism. Chapter in Intelligence and Security Informatics. Lecture Notes in Computer Science. V. 3495. P. 49–59.
16. *Ferrara E., Wang W.-Q., Varol O., Flammini A., Galstyan A.* Predicting online extremism, content adopters, and interaction reciprocity. International Conference on Social Informatics. Springer. 2016. P. 22–39. arXiv:1605.00659 [cs.SI].
17. *Sebastián A.* Ríos and Ricardo Muñoz. 2012. Dark Web portal overlapping community detection based on topic models. In Proceedings of the ACM SIGKDD Workshop on Intelligence and Security Informatics (ISI-KDD '12). ACM, New York, NY, USA, Article 2, 7 pages.
18. *Toure I.*, Gangopadhyay A. Analyzing terror attacks using latent semantic indexing, IEEE International Conference on Technologies for Homeland Security (HST). 2013 P. 334–337.
19. *Scanlon J.R., Gerber M.S.* Forecasting Violent Extremist Cyber Recruitment. IEEE Trans. Information Forensics and Security. 2015. V. 10. № 11. P. 2461–2470.
20. *Gaston L'Huillier, Hector Alvarez, Sebastián A. Ríos, and Felipe Aguilera*. 2011. Topic-based social network analysis for virtual communities of interests in the dark web. SIGKDD Explor. Newsl, 2011. V. 12. № 2. P. 66–73.
21. *Li Yang and Feiqiong Liu and Joseph Migga Kizza and Raimund K.* Ege Discovering Topics from Dark Websites IEEE Symposium on Computational Intelligence in Cyber Security, 2009. CICS '09. P. 175–179.
22. *Petrovskiy M., Tsarev D., Pospelova I.* Pattern Based Information Retrieval Approach to Discover Extremist Information on the Internet / Ghosh A., Pal R., Prasath R. (eds) Mining Intelligence and Knowledge Exploration. MIKE 2017. Lecture Notes in Computer Science, vol 10682. Springer, Cham.
23. *Manning C.D. et al.* Introduction to information retrieval. Cambridge: Cambridge university press, 2008. V. 1. P. 496.
24. *Chisholm E., Kolda T.G.* New term weighting formulas for the vector space method in information retrieval. Computer Science and Mathematics Division, Oak Ridge National Laboratory, 1999.
25. *Landauer T. K., Dumais S.T.* A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological review. 1997. V. 104. №. 2. P. 211.
26. *Lee D.D., Seung H.S.* Learning the parts of objects by non-negative matrix factorization. Nature. 1999. V. 401. №. 6755. P. 788–791.
27. *Tsarev D.V., Petrovskiy M.I., Mashechkin I.V., Popov D.S.* Automatic text summarization using latent semantic analysis. Programming and Computer Software. 2011. V. 37. №. 6. P. 299–305.
28. *Steinberger J., Ježek K.* Text summarization and singular value decomposition // Advances in Information Systems. Springer Berlin Heidelberg, 2005. P. 245–254.

29. *Ester M., Kriegel H.-P., Sander J., Xu X.* A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96) / Evangelos Simoudis, Jiawei Han, Usama M. Fayyad. AAAI Press, 1996. P. 226–231.
30. *Левенштейн В.И.* Двоичные коды с исправлением выпадений, вставок и замещений символов. Доклады Академии наук СССР. 1965. Т. 163. № 4. С. 845–848.
31. *Hurvich C.M., Simonoff J.S., Tsai C.L.* Smoothing Parameter Selection in Nonparametric Regression Using an Improved Akaike Information Criterion. Journal of the Royal Statistical Society B. 1998. V. 60. P. 271–293.
32. *Salvador S., Chan P.* FastDTW: Toward Accurate Dynamic Time Warping in Linear Time and Space. KDD Workshop on Mining Temporal and Sequential Data. 2004. P. 70–80.
33. Notation for ARIMA Models. Time Series Forecasting System. SAS Institute. Retrieved 19 May 2015.
34. *Shehabat A., Mitew T., Alzoubi Y.* Encrypted Jihad: Investigating the Role of Telegram App in Lone Wolf Attacks in the West. Journal of Strategic Security 10. 2017. № 3. P. 27–53.
35. *Page L., Brin S., Motwani R., Winograd T.* The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.
36. *Kleinberg J.M.* Authoritative sources in a hyperlinked environment. Journal of the ACM. 1999. V. 46. № 5–7. P. 604–632.
37. Wasserman, Stanley, and Faust, Katherine, Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences), (First Edition 1994) Cambridge University Press, Cambridge, UK, West 20th St., New York, USA, Melbourne.
38. *Chen T., Guestrin C.* Xgboost: A scalable tree boosting system. Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2016. P. 785–794.
39. *Ke G., Meng Q., Finley T., Wang T., Chen W., Ma W., Ye Q., Liu T.Y.* Lightgbm: a highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems. 2017. P. 3149–3157.
40. *Baldi P.* Autoencoders, unsupervised learning, and deep architectures. Proceedings of International Conference on Machine Learning Workshopon Unsupervised and Transfer Learning. 2012. P. 37–49.
41. The 20 Newsgroups data set, http://people.csail.mit.edu/jrennie/20Newsgroups/
42. Kaggle “How ISIS uses Twitter” dataset, https://www.kaggle.com/fifthtribe/how-isis-uses-twitter
43. Kaggle “ISIS religious texts dataset”. https://www.kaggle.com/fifthtribe/isis-religious-texts
44. *Breiman L., Friedman J.H., Olshen R.A., Stone C.J.* (1984). Classification and regression trees. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software. 1984. 366 p.
45. *Breiman L.* Bagging Predictors. Machine Learning. 1996. V. 24. № 2. P. 123–140.
46. *Hutter F., Hoos H., Leyton-Brown K.* Sequential model-based optimization for general algorithm configuration. Learning and Intelligent Optimization. 2011. P. 507–523.