DOI 10.37882/2223–2966.2022.01.35

АКТУАЛЬНЫЕ МЕТОДЫ И ИНСТРУМЕНТЫ ВЫЯВЛЕНИЯ ПОТЕНЦИАЛЬНО ОПАСНОГО КОНТЕНТА В СЕТИ ИНТЕРНЕТ

CURRENT METHODS AND TOOLS FOR DETECTING POTENTIALLY DANGEROUS CONTENT ON THE INTERNET

P. Urvachev
V. Dovgal
K. Budnikov
S. Irkhin

Summary. The article describes the actualization of the problem of potentially dangerous content on the Internet and analyzes existing algorithms and tools for identifying content and current software products to identify potentially dangerous content.

Keywords: neural network, artificial intelligence, parsing, speech recognition, syntax analysis, analysis and processing of content.

Урвачёв Павел Михайлович
Старший преподаватель, Кубанский Государственный Технологический Университет, Краснодар
p.m.urvachev@gmail.com
Довгаль Владислав Витальевич
Кубанский Государственный Технологический Университет, Краснодар
d.vlad.recom@gmail.com
Будников Константин Андреевич
Кубанский Государственный Технологический Университет, Краснодар
kbudnikov999@gmail.com
Ирхин Сергей Эдуардович
Кубанский Государственный Технологический Университет, Краснодар
irhinsergei@gmail.com

Аннотация. В статье описывается актуализация проблемы потенциально опасного контента в сети Интернет, а также проводится анализ существующих алгоритмов и инструментов выявления контента и актуальных программных продуктов по выявлению потенциально опасного контента.

Ключевые слова: нейронная сеть, искусственный интеллект, парсинг, распознавание речи, синтаксический анализ, анализ и обработка контента.

Актуализация проблемы

Потенциально опасный контент в широком понимании — это не только факты распространения порнографических материалов и насилия, но и массовые «вбросы» информации, ложных новостей (фейкньюс), подменяющие реальную обстановку и понимание ситуации, а также факты кибербуллинга и распространение материалов деструктивными субкультурами и организациями в том числе, запрещёнными на территории РФ. Также потенциально опасный контент существует и в корпоративной, коммерческой среде. Начиная от вредоносных материалов и несертифицированных программных продуктов, наносящих прямой материальный ущерб, так и провокационный контент, нацеленный на буллинг конкретных коммерческих структур. Данный контент наносит как репутационный, имиджевый ущерб для компании, так и приводит к прямым материальным потерям вплоть до разорения (банкротства). Одним из наиболее ярких примеров последнего времени распространения опасного контента является трагедия в Казани 11 мая 2021 года. По данным правоохранительных органов, злоумышленник заранее анонсировал свои действия в социальных сетях и мессенджерах, но данному факту никто не придал значения.

Методиками распространения потенциально опасного контента так же пользуются различные политические движения. Один из последних примеров распространения такой информации — это вовлечение детей и молодёжи до 18 лет в несанкционированные митинги путём распространения в социальных сетях и мессенджерах деструктивного контента по методикам фейкньюс. Данный пример несёт прямую угрозу государственной безопасности.

Так же остро стоит проблема распространения среди детей и молодёжи деструктивного контента запрещённой на территории РФ экстремистской организации АУЕ. Распространяемый контент носит насильственный характер, и целью распространения данного

контента является дестабилизация обстановки в молодёжной среде.

Примером массового распространения опасного контента террористическими и экстремистскими организациями является хакерская атака в 2014 году на мобильные устройства пользователей через незащищенные сети Wi-Fi московского метрополитена. На мобильные устройства пользователей вместо привычного авторизационного окна по номеру телефона появлялось сообщение с деструктивным содержанием, а именно демонстрация запугивающей информации о грядущих террористических актах.

Примером распространения опасного контента в корпоративной среде является случай распространения в сети негативных материалов о продукции компании, в том числе через отзывы покупателей.

По данным исследованиям проводимого РАЭК (Российская ассоциация электронных коммуникаций) совместно с Фондом развития интернета и МГТС интернет-риски классифицируются по 5 категориям:

1. Контентные риски: возникают в процессе использования материалов, содержащих противозаконную, неэтичную и вредоносную информацию — насилие, агрессию, эротику и порнографию, нецензурную лексику, пропаганду суицида, наркотических веществ и т.д.
2. Коммуникационные риски: связаны с межличностными отношениями Интернет-пользователей и включают в себя незаконные контакты (например, с целью встречи), киберпреследования, кибернижения, груминг и др.
3. Потребительские риски: злоупотребление правами потребителя: риск приобретения товара низкого качества, подделок, контрафактной и фальсифицированной продукции, хищение денежных средств злоумышленником через онлайн-банкинг и т.д.
4. Технические риски: возможность повреждения ПО, информации, нарушение ее конфиденциальности или взлома аккаунта, хищения паролей и персональной информации злоумышленниками посредством вредоносного ПО и др. угроз.
5. Интернет-зависимость: непреодолимая тяга к чрезмерному использованию Интернета. В подростковой среде проявляется в форме увлечения видеоиграми, навязчивой потребности к общению в чатах, круглосуточном просмотре фильмов и сериалов в Сети.

Для большинства этих категорий процесс потребления пользователями контента в сети является неотъемлемым. Таким образом основной обобщающей угрозой в сети интернет служит именно опасный контент в различных его проявлениях, будь то факты кибербуллинга, либо агитация и вербовка в деструктивные субкультуры и запрещенные организации. Также стоит отметить, что по данным этого исследования молодежь и дети предпочитают вести себя агрессивно именно в онлайн, так как они убеждены, что в онлайн пространстве их действия останутся безнаказанными. Так ответили 46% опрошенных. Также стоит отметить, что каждый пятый ребенок, являющийся жертвой кибербуллинга в возрасте от 12 до 13 лет, не обращался за помощью. Среди подростков, ставшие жертвами кибербуллинга 27% обращаются к друзьям за помощью и всего 8% за помощью к семье. Помимо этого, РАЭК отмечает, что половина опрошенных подростков становились жертвами сексуального насилия в сети. Также по данным системы мониторинга и анализа социальных медиа "Крибрум" кибербуллингу подвергается порядка 3.5 миллиона подростков Российской Федерации. Такие данные предоставляет программа для школ "ТравлиNet".

Исходя из данных АНО "Центр изучения сетевого мониторинга молодежной среды" более 93000 пользователей интернета сети в РФ так или иначе интересовались нападениями на школы. Более того система центра обнаружила в социальных сетях более 7500 сообществ, посвященных созданию и распространения опасного контента.

Эти цифры говорят нам о массовости таких явлений как кибербуллинг, флейминг, троллинг. И "культура" такого поведения и распространяемого контента носит угрозу национальной безопасности, так как все эти явления приводят к дестабилизации обстановки в молодежной среде и обществе в целом.

Методы и алгоритмы выявления текстового контента

Недопустимый контент в сети по большей части содержится именно в текстовом представлении, включая в себя: сообщения, публикации, статьи, посты в социальных сетях. Они могут хранить в себе фейкньюс, оскорбления, призыв и т.д. Для их обнаружения требуется использовать методы синтаксического анализа, от которых требуется не только находить определенные слова, но и проверять общее содержание текста.

Наиболее удобным и практичным представляется способ задания некоторого словаря запрещенных слов, поиск которых будет производиться по выбранному тексту, а затем, при их обнаружении, отправляться на модерирование.

Синтаксический анализ текста или парсинг в информатике — процедура сравнения линейной последова-

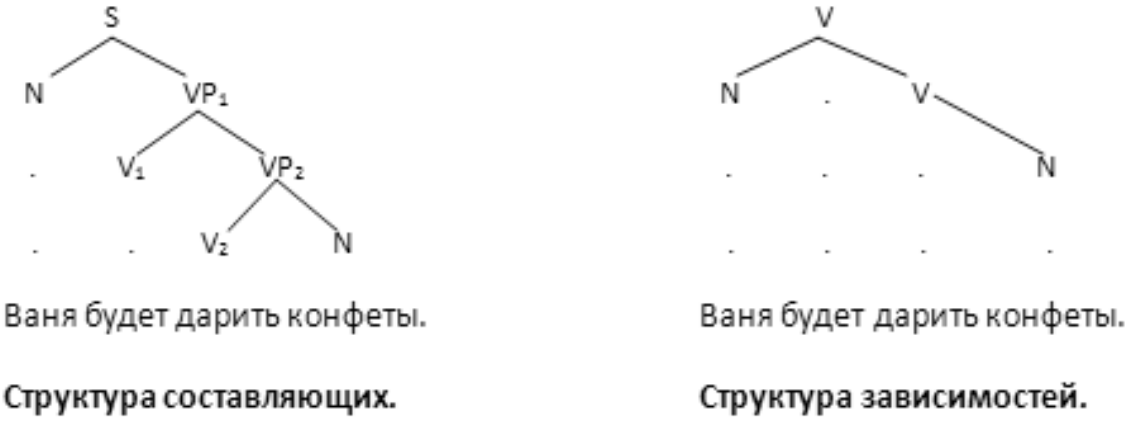

Рис. 1. Виды синтаксического строения текста

тельности символов формального или естественного языка с некоторой формальной грамматикой. В процессе анализа заданный текст преобразуется в удобную для использования структуру данных, наиболее часто — в дерево, которое отображает структуру последовательности, доступную для обработки.

Зачастую, представление синтаксического строения текста задается в виде дерева зависимостей, либо дерева составляющих, либо некоторого совмещения этих двух способов (см. Рисунок 1).

Грамматика составляющих — структурный отрезок предложения, представляющий собой более тесно связанные составляющие меньшего размера. Такая грамматика построена на утверждении, что любая сложная грамматическая единица состоит из двух более простых непересекающихся единиц. Составляющая, имеющая более одного слова, считается группой, а слово, представляющее корневой узел, описывающий группу — вершина группы.

Грамматика зависимостей — формальная модель структурного синтаксиса. Представляется иерархическим построением предложения, где между компонентами существует отношение зависимости. Все связи в предложении считаются подчинительными: вершина предложения — сказуемое или его знаменательная часть; предлоги — управляющие относящимися к ним формами существительных.

Основные типы алгоритмов преобразования исходного потока информации:

- Восходящий парсер — результаты получаются из правых частей, начиная с токенов и заканчивая начальным символом;
- GLR-парсер — в информатике расширенный алгоритм LR-парсера, предназначенный для разбора по недетерминированным и неоднозначным грамматикам;
- LR-анализатор — синтаксический анализатор для исходных кодов программ, написанных на некотором языке программирования, который читает входной поток слева направо и производит наиболее правую продукцию контекстно-свободной грамматики;
- Нисходящий парсер — результаты получаются, начиная со стартового символа и до получения некоторой последовательности токенов;
- Метод рекурсивного спуска — алгоритм, реализуемый путём взаимного вызова процедур, где каждая процедура соответствует одному из правил контекстно-свободной грамматики или БНФ;
- LL-анализатор — нисходящий синтаксический анализатор для некоторого подмножества контекстно-свободных грамматик, известных как LL-грамматики.

В данный момент практически для каждого языка программирования можно найти и использовать существующие библиотеки веб-скрейпинга, результаты работы которых затем проверять на наличие запрещенного контента.

Некоторые известные:

1. Python — BeautifulSoup, Selenium;
2. JavaScript — Cheerio, Osmosis.

Области применения:

- Структурированные данные (CSS, HTML, XML и т.д.);
- SQL-запросы;
- Лингвистика и естественные языки (машинный перевод, генераторы текстов);
- Регулярные выражения (нахождение подстроки в строке).

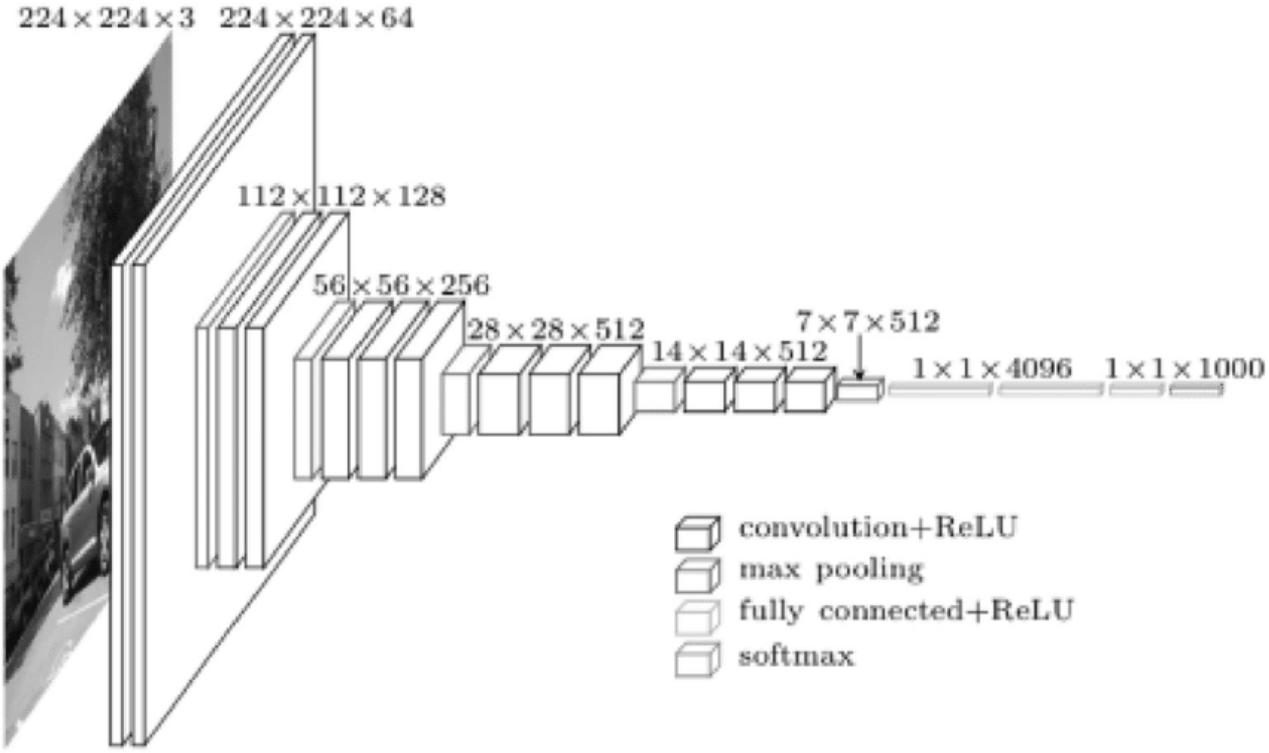

Рис. 2. Пример обработки изображения

Методы выявления графического и видео контента

На сегодняшний день нейросетевые методы, алгоритмы и технологии искусственного интеллекта (ИИ) являются наиболее перспективными в распознавании и выявлению графического контента. Существуют готовые решения для организации и интеграции в свою систему нейронной сети: Google Cloud Vision API, Платформа Amazon Rekognition, Apple VoiceOver

Для более гибкой настройки предполагаемой системы, безусловно, потребуется разобраться, что из себя представляет процесс обнаружения. Сначала необходимо выбрать как обрабатывать изображение, с чего начать. Основные подходы:

- метод скользящего окна — пошаговая обработка каждого определенного региона изображения;
- подход с предложением регионов — оптимизация предыдущего метода путём добавления алгоритма предложения регионов, где скорее всего находится искомый объект (R-CNN модели);
- обнаружение в один проход — анализ всего изображения (архитектуры YOLO и SSD) (см. Рисунок 2).

Внутри выбранного региона необходимо решить следующие задачи:

- поиск границ объекта (резких переходов/изменений яркости) — чаще всего используется детектор границ Кенни, основные этапы которого: сглаживание, поиск градиентов, подавление немаксимумов, двойная пороговая фильтрация, трассировка области неоднозначности. Итогом является двоичное изображение, содержащее границы;
- обнаружение структурных элементов — наилучшим для этого методом будет преобразование Хафа, которое служит для поиска на изображении аналитически заданных фигур (прямых, окружностей и любых других, описуемых уравнением);
- определение положения объекта путём распознавания объекта по построенному эталону или поиска изменений/смещений по кадрам. (см. Рисунок 3)

Существует три метода распознавания:

- Шаблонное сравнение.
- Структурные системы, где объект описывается как граф, узлами которого являются элементы входного объекта, а дугами — пространственные отношения между ними.
- Нейросети — системы, состоящей из нейронов входного, промежуточных и выходного (распознающего) слоя, которая на основе весов и поро-

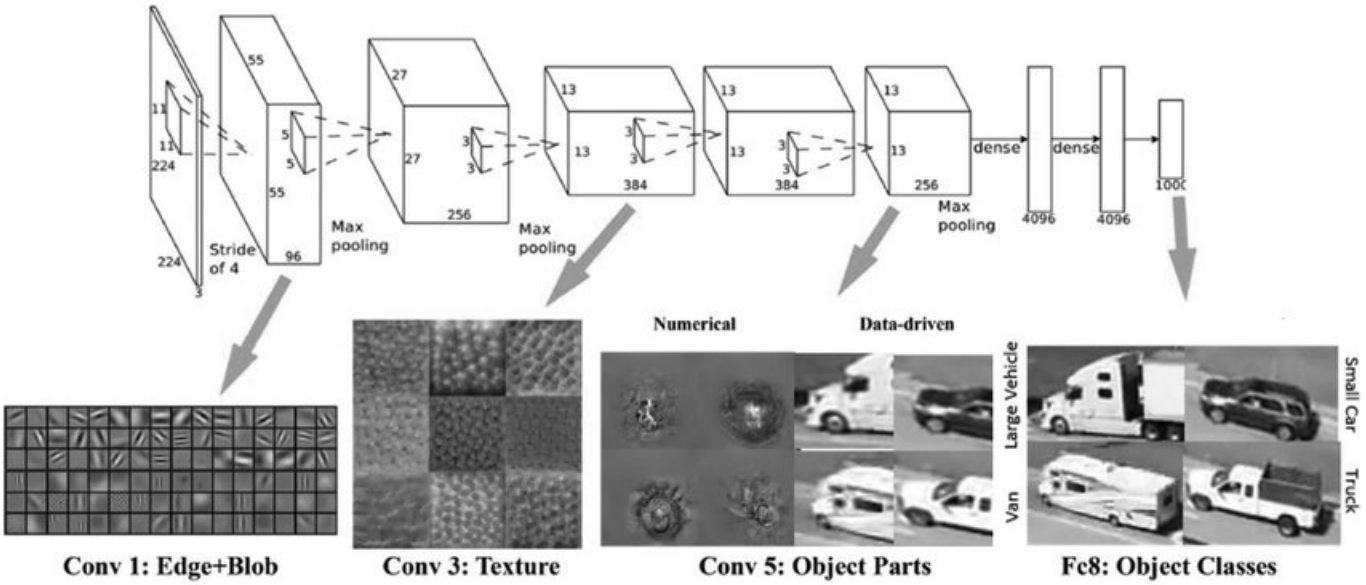

Рис. 3. Анализ характеристик выбранного региона

гов возбуждения этих самых нейронов производит анализ изображения.

Чаще всего, конечно же, к использованию популярны модели нейронных сетей, в которых используются алгоритмы глубокого обучения. Они в общем смысле заключаются в выявлении, анализе и запоминании некоторых абстрактных «деталей» заданных классов объектов.

Существует два подхода:

- Обучение модели с нуля. Для этого потребуется большой набор данных и готовая архитектура сети, которая будет изучать объекты и строить модель.
- Использование предварительно обученной модели. Здесь, как правило, используют подход трансферного обучения — точная настройка предварительно обученной модели: в существующую сеть вводятся новые данные с писанными ранее неизвестными классами объектов.

После обучения важно провести большую работу по тестированию и корректировке модели, чтобы добиться оптимальных параметров точности и скорости распознавания, а также избежать недо- и переобучения модели. Это случаи, когда сеть плохо срабатывает или слишком зависит от обучающих данных.

Здесь также существует большое количество готовых решений-библиотек преимущественно для языка Python: Tensorflow, ImageAI, OpenCV. Они содержать в себе все необходимые функции для формирования, настройки, обучения и использования моделей нейронной сети.

Кроме того, стоит упомянуть методы машинного обучения распознавания объектов, они предлагают отличные от глубокого обучения подходы:

- извлечение функций HOG с помощью модели SVM;
- модели «мешков слов» с функциями SURF и MSER;
- алгоритм Виолы-Джонса.

Процесс машинного обучения состоит из нескольких этапов:

- формирование набора обучающих и тестировочных данных;
- выбор нужных характеристик для каждого объекта (не автоматически, как при глубоком обучении).

Методы и алгоритмы выявления аудио контента

Для определения принципов выявления аудио контента обратимся к методам распознавания речи. Здесь можно либо сравнивать исходные данные с неким шаблоном, используя динамическое программирование, либо преобразовать речь в текст, выделяя из потока речи отдельные лексические элементы — фонемы, которые далее объединяются в морфемы, в этом помогут метод дискриминантного анализа, скрытые Марковские модели или нейронные сети.

Стандартная архитектура систем распознавания включает в себя:

- Модуль шумоочистки для выделения полезного сигнала;

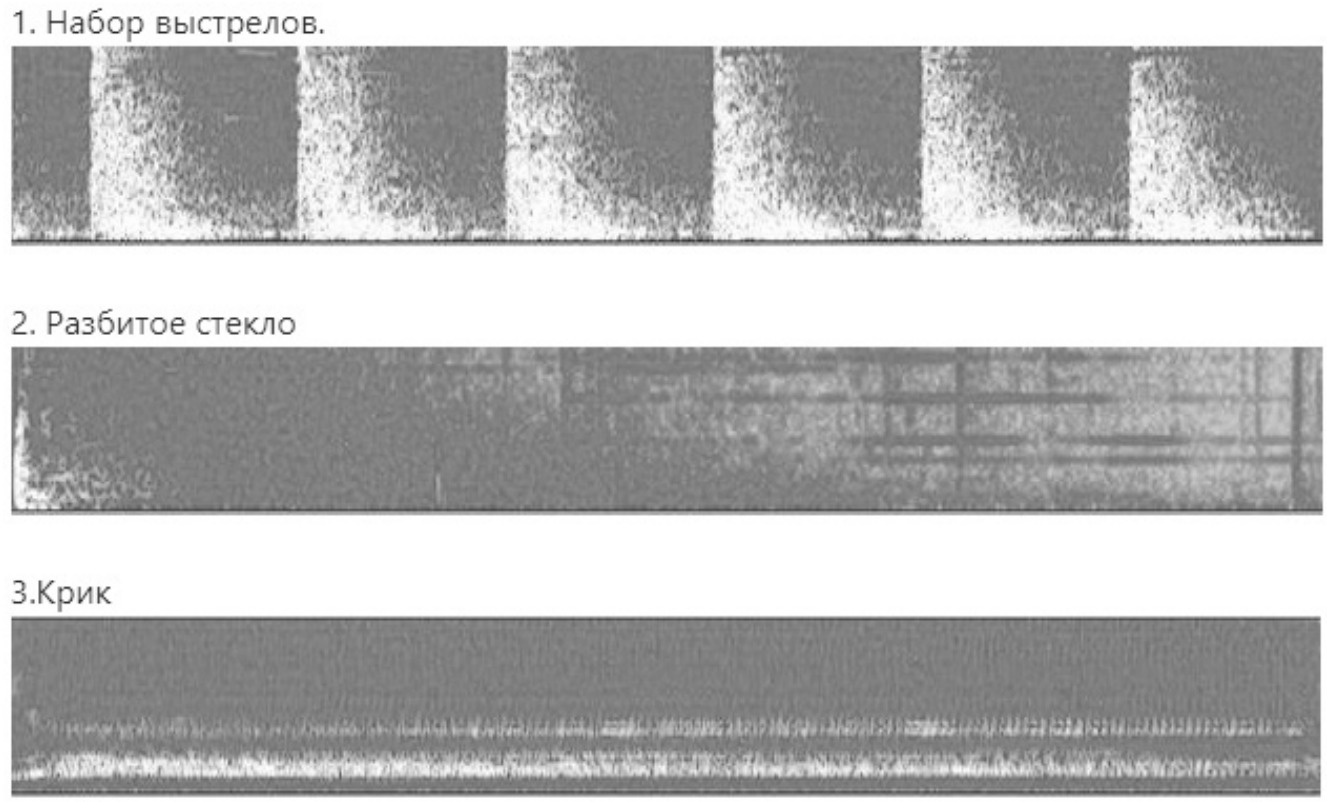

Рис. 4. Образы некоторых звуков

- Акустическая модель для оценки схожести речевого сегмента с заранее определенной моделью звука;
- Языковая модель для определения наиболее вероятных последовательностей фонем/морфем;
- Декодер для определения наиболее вероятной последовательности слов, которая и является конечным результатом.

Для этого существует множество, как бесплатных онлайн сервисов (Google Docs, Speechpad, Dictation, wreally), платных приложений (RealSpeaker, Voco, Express Scribe), так и библиотек для языков программирования (Kaldi, PocketSphinx, Vosk-api для Python). В дальнейшем предстоит уже анализ текстовой информации, как описано выше.

Стоит ещё упомянуть, что как нежелательный аудио контент можно охарактеризовать: агрессию (разговор на повышенных тонах, крик), сигнализацию автомобилей, разбивающиеся стекла, ключевые слова (“полиция”, “помогите”), выстрелы, крик/плач ребенка. Данные случаи анализируются, исходя из характеристик аудиодорожки или её части, таких как громкость, высота, частота, скорость и тембр. (см. Рисунок 4)

Актуальные инструменты выявления и блокировки потенциально опасного контента

После проведения анализа систем мониторинга и блокировки потенциально опасного контента было выявлено 11 актуальных автоматизированных либо частично автоматизированных программных продуктов. Сводные данные анализа представлены в таблице 1. Среди основных критериев были выделены стоимость использования системы, а также особенности её организации и сферы применения.

Выводы

Анализируя предметную область, можно прийти к заключению, что актуальность проблемы носит острософиальный характер и в некоторых проявлениях может нести угрозу не только безопасности отдельных лиц, но даже и государственной безопасности. К примеру, распространение деструктивного контента в детской и молодёжной среде.

В результате анализа актуальных программных продуктов было установлено отсутствие комплексного подхода к анализу и мониторингу контента. В частности,

Таблица 1. Актуальные программные продукты по анализу потенциально опасного контента

Название	Особенности	Стоимость подписки	Сфера применения
Darvvin	1) Предоставляется по модели Software As A Service (SaaS) 2) Проведение работ по внедрению в закрытые Интернет-сообщества 3) Каждая опасная тематика представлена отдельным модулем в система	1) Для города (г. Краснодар) — 24,422,100 руб/год 2) Для региона РФ (Краснодарский край) — 97,758,350 руб/год	Системы корпоративного и коммерческого использования
Bitcop	1) Интеграция с профессиональным прикладным ПО 2) Учёт отработанного времени пользователя 3) Оценка продуктивности пользователя 4) Оперативная сводка состояния рабочий программ пользователя 5) Мониторинг поисковых запросов 6) Возможность работы в скрытом режиме, исключая обнаружение пользователем	Lite: при подписке на год 120 руб/мес Pro: при подписке на год 168 руб/мес Enterprice: 3300 руб за сотрудника (от 10 сотрудников)	Системы корпоративного и коммерческого использования
SkyDNS	1) Предоставляется по модели Software As A Service (SaaS) 2) Отдельные настройки для каждого устройства 3) Собирает статистику посещения сайтов	1) Семейный пакет: 495р в год 2) Бизнес пакет: 4500р в год (подключение до 50 устройств)	Системы комплексного мониторинга
UserGate Web Filter	1) Блокировка по категориям сайтов 2) Морфологический анализ 3) Белые и чёрные списки	1) Родительский контроль: 890р в год 2) Цена корпоративного пакета устанавливается индивидуально	Системы комплексного мониторинга
Panda Gate Defender eSeries	1) Блокировка по категориям сайтов 2) Настройка VIP-пользователей 3) Белые и чёрные списки	Цена пакета сообщается после беседы с менеджером	Системы корпоративного и коммерческого использования
Kaspersky Total Security	1) Контроль активности в Интернете 2) Контроль использования программ 3) Настройка использования устройства по расписанию 4) Блокировка нежелательного контента 5) Контроль активности в соцсетях 6) Блокировка загрузки файлов	Для дома: от 1599р за 2 года (1 устройство) Для малого бизнеса: 4290р за 1 год (5 устройств) Для среднего бизнеса: 30690р за 1 год (10 устройств)	Системы комплексного мониторинга
Comodo Internet Security	1) Блокировка нежелательного контента 2) Блокировка загрузки файлов	29.99$ (2195р) в год	Системы родительского контроля
Qustodio	1) Контроль активности в Интернете 2) Контроль активности в соцсетях 3) Настройка использования устройства по расписанию 4) Блокировка нежелательного контента	1) Семейный пакет: 49.46$ (3620р) в год (5 устройств) 87.26$ (6387р) в год (10 устройств) 124.16$ (9088р) в год (15 устройств) 2) Бизнес пакет: от 6.91$ (от 505р) в год (подключение от 5–100 устройств) 3) Школьный пакет: от 6.91$ (от 505р) в год (подключение от 5–100 устройств)	Системы родительского контроля
NetPolice	1) Блокировка по категориям сайтов 2) Блокировка обмена сообщениями в соцсетях, форумах, сайтах 3) Блокировка загрузки файлов 4) Скрытый режим работы 5) Журнал подключения к сайтам	1) NetPolice Child: 500р в год 2) NetPolice Pro: 720р в год	Системы родительского контроля

Таблица 1 (продолжение). Актуальные программные продукты по анализу потенциально опасного контента

Название	Особенности	Стоимость подписки	Сфера применения
Mobicip	1) Ограничение времени экрана для каждого устройства 2) Блокировка сайта по фильтру. 3) Блокировка приложений	В зависимости от кол-ва устройств: 3.99$ (292р) в месяц (подключение 5 устройств) 4.99$ (365р) в месяц (подключение 10 устройств) 9.99$ (731р) в месяц (подключение 20 устройств)	Системы родительского контроля
McAfee	1) Защита домашней сети 2) Блокировка сайтов	1 устройство: 34.99$ (2561р) в год 5 устройств: 39.99$ (2927р) в год До 10 устройств: 44.99$ (3293р) в год	Системы родительского контроля
Norton	1) Умный брандмауэр 2) Блокировка несанкционированного доступа к веб-камере 3) Родительский Контроль	Delux: 44.7$ (3271р) в год Premium: 57.2$ (4186р) в год Standart: 35$ (2561р) в год	Системы родительского контроля

в программах родительского контроля нет комплексов программных средств, которые могли бы устанавливаться на все устройства в семье, начиная от смартфонов и персональных компьютеров, заканчивая smart-TV и игровыми станциями. В бизнес среде также было обнаружено отсутствие комплексного подхода к анализу и мониторингу контента. Отсутствует инструментарий для ручного мониторинга контента и выявление потенциально опасной информации кибер-волонтёрами.

Были рассмотрены и изучены существующие и активно используемые алгоритмы и программные средства по обработке и анализа контента разного рода: графического и аудио, видео и текст.

ЛИТЕРАТУРА

1. Гибридная нейро-экспертная система для идентификации значимых 3 событий на графиках временных рядов Частиков А.П., Урвачев П.М., Тотухов К.Е. // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. 2016. No 124. С. 756–769.
2. Распознавание паттернов в диаграммах управления на основе нейронных 1 сетей с подкреплением Частиков А.П., Урвачев П.М., Тотухов К.Е. // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. 2016. No 124. С. 770–789.
3. Регрессионный анализ для прогнозирования объема работ при ремонте 0 дорог Частиков А.П., Урвачев П.М., Аксенов Г.В. // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. 2014. No 99. С. 585–607.