Инженерный вестник Дона, №8 (2024)
ivdon.ru/ru/magazine/archive/n8y2024/9442

Модуль поиска деструктивной информации в изображениях

А.А. Джуров, Л.В. Черкесова, Е.А. Ревякина

Донской государственный технический университет, г. Ростов-на-Дону

Аннотация: Изображения на вебсайтах, социальных сетях, компьютерах могут содержать деструктивный контент и представлять собой угрозу для психики ребенка или подростка. Обычная классификация изображений не всегда правильно работает, и, соответственно, имеет ряд своих недостатков, из-за которых могут быть ложные срабатывания, что понижает точность классификации. В статье представлен метод в модуле Python, который позволяет обнаруживать вредоносный контент в изображениях. Метод, описанный в статье, основан на использовании библиотеки Yolov8, что обеспечивает хорошую классификацию изображений и дальнейшего анализа. С помощью разработанного метода удалось сократить количество ложных срабатываний, что привело к повышению эффективности его использования. В статье показана схема его работы, а также продемонстрирован поиск объектов в изображениях. Рассмотрены программы - аналоги и проведен их сравнительный анализ с разработанным методом.
Ключевые слова: Spacy, деструктивные изображения, информационная безопасность, pymorphy3, ultralytics, деструктивный текст, деструктивный контент, YoloV8, безопасность детей, сравнительный анализ, цифровой хэш.

Введение.

Интернет – среда зачастую более опасная, чем даже тот самый уличный двор со всеми хулиганами, вместе взятыми [1].

Деструктивный или опасный контент – любой контент (информация), который может нанести вред человеку или сподвигнуть его к причинению вреда другому [2]. К нему относится также шок-контент, содержащий насилие над людьми или животными, контент, пропагандирующий нетрадиционные ценности, употребление или распространение наркотиков, селфхарм (причинение вреда самому себе) или даже самоубийство.

Адресатами деструктивного контента можно стать случайно. Однако зачастую речь может идти о намеренном поиске пользователем такого контента или поиске злоумышленником пользователя, которому такой контент может оказаться «интересен» [3]. С развитием цифровых технологий люди получили обширный инструментарий для работы с информацией, который позволяет не только получить необходимую информацию о другом

человеке, но и создать вокруг него желаемый информационный фон (эхокамеру).

При этом более всего подвержены деструктивному воздействию именно молодые люди. Психика и ум детей и подростков находятся на очень «хрупком» этапе формирования, их система ценностей еще окончательно не сформирована. Поэтому молодые люди более зависимы от мнения окружающих (особенно тех, кому они подражают, с кем хотят ассоциироваться), они сложнее переживают негативное отношение к себе, зачастую менее склонны сочувствовать чужой позиции (беде, несчастью), представлять себя на месте другого, более жестоки и беспощадны. Деструктивный контент может содержаться в любых формах, включая и изображения.

Деструктивный контент в изображениях может включать в себя различные материалы, которые представляют угрозу для людей и общества [4]. На рис. 1 приведены некоторые примеры деструктивного контента в изображениях.

Рис. 1. – Виды деструктивного контента в изображениях

1) Изображения убийств и насилия могут быть неприемлемыми для просмотра по разным причинам, они могут вызывать сильные эмоции от возбуждения до крайнего отвращения и даже ужаса [5].

Определения таких изображений:

– графическое насилие. Это изображение особенно ярких, жестоких и реалистичных актов насилия в визуальных средствах массовой информации, таких как кино, телевидение и видеоигры [6];

– недопустимый контент. Например, видео-, аудиозаписи и изображения дорожно-транспортных происшествий, стихийных бедствий, уличных драк, физических нападений, жертвоприношений, пыток, протестов, беспорядков, грабежей и другие.

2) Психическое воздействие изображений с оружием. В социальной психологии существует теория, согласно которой простое присутствие оружия или его изображений может приводить к более агрессивному поведению у людей, особенно если они уже возбуждены [7].

3) Порнографическое изображение — это вульгарно-натуралистическое, непристойное изображение или словесное описание гениталий, полового акта, сексуальных действий [8].

В правовом пространстве такое определение содержится в Федеральном законе от 29.12.2010 №436-ФЗ «О защите детей от информации, причиняющей вред их здоровью и развитию»: информация, представляемая в виде натуралистических изображения или описания половых органов человека и (или) полового сношения либо сопоставимого с половым сношением действия сексуального характера.

4) Изображение наркотиков может рассматриваться как пропаганда, если оно используется в целях распространения наркотиков, или обучения способам их приготовления или употребления, или же склонения к их употреблению.

Просмотр изображений с употреблением наркотиков может грозить ребёнку следующими негативными последствиями для психики:

– концентрация на негативном. Ребёнок не будет заинтересован ни в чём, кроме как в подпитке эмоций от наркотиков;

– изменение поведения. Появятся агрессия, злость, начнутся проблемы во взаимоотношениях в семье, школе, с друзьями;

– деградация психики. Запоминать информацию станет всё сложнее, ни о какой трезвой оценке ситуации не может идти и речи.

5) Просмотр изображений с употреблением сигарет может грозить ребёнку негативными последствиями для психики. Например, исследования показывают, что сцены курения в фильмах могут побудить подростков начать курить.

Следует отметить, что такая информация может содержать материалы, которые не следует рассматривать как поощряющие или пропагандирующие курение, и их просмотр может нанести вред психике.

Аналоги программных средств.

1) Kaspersky Internet Security - программа, которая защищает устройство от вирусов, шпионского программного обеспечения и других угроз [9]. Также приложение позволяет производить блокировку web-сайтов с порнографией, насилием и алкоголем. Комплексное программное обеспечение Kaspersky Internet Security (KIS) разработана компанией Kaspersky Lab и предназначено для защиты устройства от различных интернет-угроз, среди которых: вирусы, шпионское и рекламное программное обеспечение, фишинг, атаки хакеров и других опасностей, встречающихся в сети интернет. Программа Kaspersky Internet Security предоставляет не только базовую защиту от вредоносных программ, но также имеет функции фаервола, защиты от рекламы и фишинга, контроля доступа к интернету, родительского контроля и множество других полезных инструментов, которые направлены на безопасность и конфиденциальность при работе в сети интернет.

Достоинства Kaspersky Internet Security:

– Высокая степень защиты от вирусов и опасного программного обеспечения, высокая эффективность в обнаружении и блокировки различного вредоносного программного обеспечения.

– Защита онлайн-банкинга. Приложение имеет специальные функции, например, безопасное браузерное окно, с помощью которого можно безопасно производить финансовые транзакции.

– Защита конфиденциальной информации. Программа Kaspersky Internet Security позволяет блокировать доступ к личным данным и выполнять защиту конфиденциальных файлов или папок.

– Защита при работе в сети. Программное обеспечение имеет эффективные инструменты для контроля web-сайтов, блокировки деструктивного контента и фильтрация нежелательной почты.

– Приятный и интуитивно понятный интерфейс. Программа Kaspersky Internet Security обладает простым в использовании пользовательским интерфейсом, с помощью которого легко управлять настройками безопасности.

Недостатки Kaspersky Internet Security:

– Ресурсоемкость. Программное обеспечение Kaspersky может замедлить работу устройства, особенно это заметно при выполнении сканирования в реальном времени.

– Высокая стоимость. Программа Kaspersky Internet Security, в сравнении с другими антивирусными программами, может быть дороже.

– Ограничения в работе функционала. В некоторых функциях программы существует ограничения, например, родительский контроль и защита Wi-Fi, в сравнении с конкурентами.

– Влияние на производительность. В редких случаях антивирусное программное обеспечение Kaspersky может увеличить нагрузку на ресурсы устройства или замедлить интернет-соединение.

– Возможные проблемы совместимости. В некоторых случаях пользователи сталкиваются с проблемами совместимости с другими программами или операционной системой при использовании программы Kaspersky.

2) McAfee - программное обеспечение, предназначенное для защиты от вирусов, шпионского программного обеспечения, а также для блокировки web-сайтов с деструктивным контентом, таким, как: порнография, насилие, алкоголь и другое [10]. Компания McAfee специализируется на разработке такого программного обеспечения, которое позволяет защитить компьютер или сеть от вирусов, троянов и других угроз, встречающихся в сети Интернет. Компания была основана Джоном Макфеем в 1987 году и считается одной из самых известных и уважаемых компаний в области информационной безопасности. Программное обеспечение компании McAfee предназначено для защиты различных устройств, таких, как: компьютер, мобильные телефоны и сервера. Продукты McAfee включают в себя: антивирусное программное обеспечение, систему контроля и управления угрозами, а также различные инструменты для обеспечения безопасности информации.

Достоинства:

– Надежность. Программа McAfee является надежным поставщиком антивирусных программ. Также компания имеет долгую и большую историю на рынке кибербезопасности.

– Широкий набор инструментов. Программное обеспечение McAfee имеет множество разных функций, среди которых: антивирусная защита, защита конфиденциальности, анти-спам, защита от вредоносных программ.

– Простота использования. Продукты McAfee имеют интуитивно понятный интерфейс. Это делает программные продукты McAfee лёгкими в использовании как для обычных пользователей, так и для бизнес-клиентов.

– Поддержка на многих устройствах. Приложение McAfee может работать на различных устройствах, таких как: Windows, Mac, Android и iOS.

Недостатки:

– Влияние на производительность. Пользователи приложения McAfee отмечают, что использование продукта снижает производительность устройств.

– Стоимость подписки. Подписка на продукцию McAfee является более дорогой, чем у конкурентов.

– Навязчивые уведомления. Пользователи жалуются на частые и навязчивый уведомления от продукции McAfee.

Продукция McAfee предлагает большое количество функций по защите устройств. Несмотря на то, что компания известна своей надежностью, все же некоторые пользователи могут столкнуться с недостатками, среди которых влияние на производительность и высокая стоимость подписки.

3) Google SafeSearch - сервис, который производит фильтрацию результатов поиска Google, а также блокирует деструктивный контент: сцены курения, насилие, порнографией [11].

Google SafeSearch - это функция, которая при поиске производит фильтрацию деструктивного контента, например, порнографии или откровенных изображений, и блокирует его при поиске в Google. Функционал SafeSearch может помочь создать для детей и подростков безопасное пространство в сети интернет. Также функционал предназначен не только для детей, но и взрослых, которые предпочитают избегать нежелательный контент в поисковой системе Google.

Когда включена функция SafeSearch, результаты поиска в Google фильтруются с целью исключения материалов, предназначенных для взрослых. Родители могут управлять этой функцией, включая ее или выключая с помощью настроек поиска Google. Также, настроить SafeSearch

могут и администраторы сети, для определенных учетных записей в рабочей среде, образовательных учреждениях и других организациях, для обеспечения безопасности в сети интернет.

Google SafeSearch помогает всем, кто использует поисковую систему Google, обеспечить безопасный и не деструктивный контент.

Достоинства Google SafeSearch:

– Фильтрация нежелательного контента. Сервис SafeSearch может помочь защитить отображение деструктивного контента: откровенные изображения, порнографию и другой нежелательный контент. Это делает сервис подходящим для обучения ребенка в интернете.

– Удобство использования. Функция SafeSearch доступна бесплатно и легка в использовании. Родители могут без труда включать и отключать сервис в настройках Google.

– Защита детей. Функция SafeSearch помогает создать для детей и подростков безопасное окружение в интернете, фильтруя деструктивный контент, не подходящий по возрасту.

– Поддержка на уровне сети. Сервис SafeSearch устанавливается не только родителями, но и администраторами организаций и образовательных учреждений, для обеспечения безопасности учетных записей.

Недостатки Google SafeSearch:

– Неидеальная фильтрация информации. Функционал SafeSearch в редких случаях может пропускать деструктивный контент или блокировать те сайты, которые не являются деструктивными. Это может привести к неудобству при использовании устройства с функцией SafeSearch.

– Ограничение доступа к полезному контенту. Иногда фильтрация SafeSearch может блокировать полезную информацию, такую, как академические или научные материалы.

– Возможные обходы фильтрации информации. Функцию SafeSearch можно обойти, если использовать нестандартные методы поиска материалов в сети интернет.

Функционал Google SafeSearch имеет значительные преимущества в обеспечении безопасного пространства для детей и подростков. Однако некоторые ограничения функции SafeSearch могут вызывать недовольства пользователей. Это требует постоянное совершенствования и обновления сервиса.

Разработанный модуль.

Разработанный модуль по поиску деструктивного контента в изображениях представляет из себя изучения контекста изображения, т.е. изучение текста, который содержится вместе с изображением. В большинстве случаев изображения на сайтах имеют текст, к которому они относятся. Экспериментальными методами была сформирована формула (1), по которой идет расчет деструктивного контента в изображениях.

$$Des = \frac{X+Y}{1+Y} - 0.1\,, \qquad (1)$$

где:

Des – процент деструктивного отношение изображения к тексту;

X – процент деструктивности текста (можно использовать известные методы);

Y- процент деструктивности текста (классификации изображения).

Формула (1) показывает деструктивное отношение изображения к тексту. Для примера, у нас будут следующие входные данные показанные в таблице 1.

На рис. 2 показан график отношения деструктивности изображения к тексту по указанным в таблице 1 входным данным.

Таблица №1

Входные данные для разработанного метода.

Деструктивность текста, %	Деструктивность изображения, %
30	30
40	40
50	50
60	60
70	70

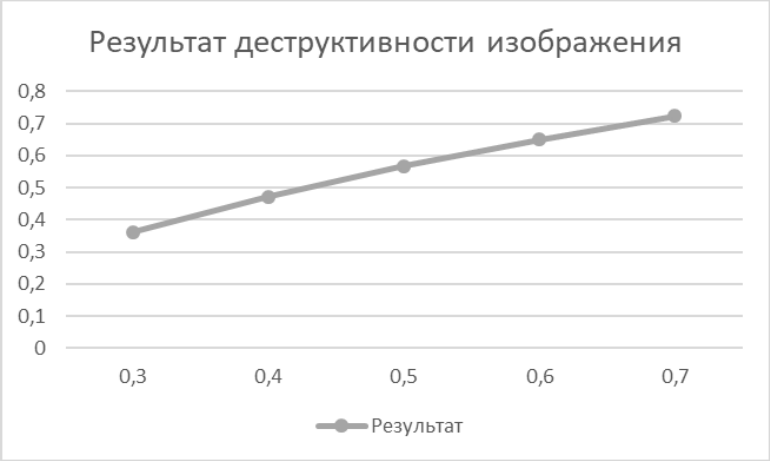

Рис. 2. – Схема результата деструктивности изображения для таблицы 1

По оси *X* отображается значение деструктивности изображения, по оси *Y* результат работы разработанного метода. Полученный результат, рассчитанный по формуле 1 показан в таблице 2.

Таблица №2

Результат расчета входных данных новым методом.

Результат, %
36,15
47,14
56,66
65
72,35

Вышеуказанная формула предусмотрена для того, чтобы избежать плавающий порог, при котором программа считает, что изображение

деструктивно. Например, если процент деструктивности изображения равен 60%, а порог, при котором изображение деструктивным считается от 80%, то такие пороги не показывают объективную оценку деструктивности. Предположим такое понятие, как сомнение нейронной сети, сомнением будет считаться результат от 30% до 70% по деструктивности изображения (*Y*), в пределах этих процентов будет работать реализуемый метод с помощью вышеуказанной формулы. Если у нейронной сети результат будет выше 70%, тогда нейронная сеть будет уверена в том, что изображение деструктивное. Если у нейронной сети результат будет ниже 30%, тогда нейронная сеть будет уверена, что изображение не является деструктивным.

В разработанном методе порогом деструктивности (*Des*) всегда будет считаться одно значение 50%, т.е, если происходит сомнение нейронной сети, то выполняется расчет по формуле (1) результатом которой будет значение в процентах. Если это значение будет равно или более 50%, тогда программа будет считать, что изображение деструктивное. Если результат формулы будет менее 50%, тогда программа будет считать, что изображение не деструктивное.

Новый метод использует для распознавания элементов на изображениях библиотеку YoloV8, которая имеет ряд преимуществ [12]:

– высокая точность обнаружения объектов. Модель демонстрирует рекордную точность на различных тестовых наборах данных;

– высокая скорость работы. YOLOv8 работает очень быстро, что делает его идеальным для обработки данных в реальном времени;

– простота использования. YOLOv8 имеет простой и интуитивно понятный API, который облегчает обучение и использование модели;

– гибкость. YOLOv8 поддерживает различные сценарии, включая детектирование объектов, сегментацию, классификацию и отслеживание;

– адаптивность к разным условиям. Модель сохраняет высокий уровень точности в различных условиях, что делает её надёжным инструментом в разнообразных приложениях;

– эффективное использование ресурсов. YOLOv8 способен эффективно использовать вычислительные ресурсы, что позволяет запускать его на различных устройствах, включая мобильные устройства и встроенные системы, без значительного снижения производительности.

Модель обучалась на более чем 5 тысяч изображений. Схема работы разработанного модуля изображена на рис. 3.

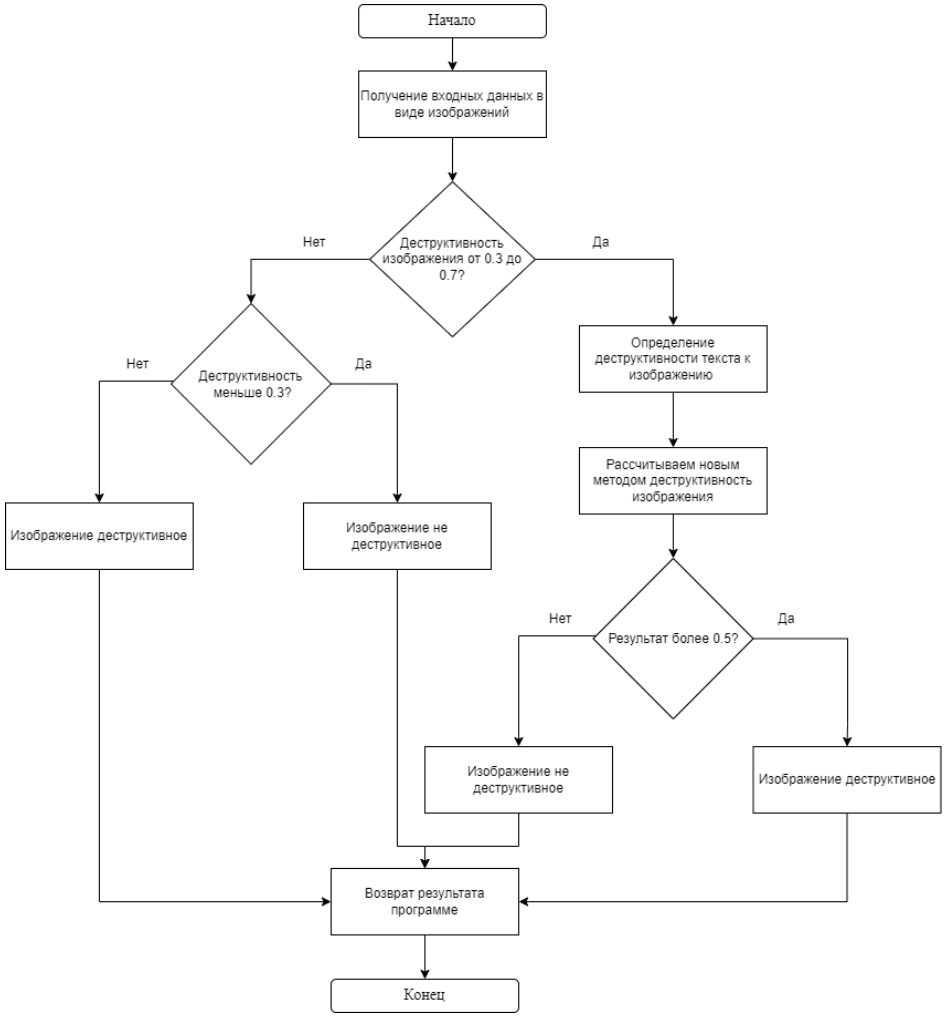

Рис. 3. – Схема работы нового модуля выявления деструктивности изображения при выполнении одной итерации

Рассмотрим следующий пример. На рис. 4 изображено, как врач делает укол пациенту.

Рис. 4. – Укол врача

Как сработает бинарный классификатор деструктивного контента в изображении на указанном выше рис. 4? Классификатор обнаружит на фото иглу и посчитает изображение деструктивным, так как игла может ассоциироваться с наркотиками, не вникая в подробности фото.

Как это изображение будет классифицировать разработанный модуль:

– определим, что изображено на фото;

– на фото изображена игла или укол иглой;

– разработанный метод определяет, что деструктивность фото 67%;

– так как разработанный метод работает в пределах 30% -70%, то данный метод проверяет деструктивность текста;

– для примера текст к фотографии такой: «27 марта в медицинском центре провели вакцинацию от гриппа», предположим, что деструктивность текста равна 10%;

– применив формулу получим результат деструктивности изображения 36,1%, что ниже 50%, что считается не деструктивным изображением.

А теперь проверим, как будет выглядеть деструктивность контента, если текст к рис. 4 будет такой: «Употребление наркотиков — это такой

кайф, когда колешься, то на седьмом небе от счастья». Деструктивность текста в данном случае, предположим, будет равна 92%. Применив формулу нового метода, получим значение деструктивности текста, равное 85,2%, что подлежит блокировке.

Также на рис. 5, рис. 6 показаны примеры поиска деструктивной информации в изображениях с помощью YoloV8.

Рис. 5. – Пример поиска сигарет и курения в изображениях с помощью YoloV8

Рис. 6. – Пример поиска оружия в изображении с помощью YoloV8

В таблице 3 показано сравнение классификации известными методами не деструктивных изображений с деструктивным текстом.

Таблица №3

Сравнение классификации известных методов не деструктивных изображений с деструктивным текстом.

Метод	Количество не деструктивных изображений с деструктивным текстом, шт.	Количество правильной классификации, шт.
Новый метод	100	100
Сравнительный анализ		91
Цифровой хэш		93

В таблице 4 показано сравнение классификации известными методами деструктивных изображений с не деструктивным текстом.

Таблица №4

Сравнение классификации известных методов деструктивных изображений с не деструктивным текстом.

Метод	Количество деструктивных изображений с не деструктивным текстом, шт.	Количество правильной классификации, шт.
Новый метод	100	100
Сравнительный анализ		89
Цифровой хэш		86

Заключение.

Доказана эффективность разработанного нового модуля для поиска деструктивной информации в изображениях за счет анализа контекста изображения.

Основные результаты заключаются в следующем:

1. Проанализированы современные методы и системы для обнаружения деструктивной информации в изображениях, и как основной общий недостаток – это бинарная классификация изображений, когда значение деструктивности либо равно 0, либо равно 1 без контекста к изображению;

2. Предложен метод и алгоритм для эффективного обнаружения деструктивного контента в изображениях. Разработанный модуль поиска деструктивного контента в изображениях анализирует контекст к изображениям и на анализе изображения и контекста к нему. За счет этого определяется необходимость блокировки деструктивного контента или нет.

3. Проведены экспериментальные исследования предложенного метода и модели в части обнаружения деструктивного контента в изображениях, в результате чего показано, что точность определения деструктивного контента в сравнении с другими методами составляет на 10% точнее, что является более высоким уровнем по сравнению с конкурирующими методами.

Литература

1. Титор С. Е., Каменева Т. Н. Деструктивное влияние интернета на поведение несовершеннолетних: результаты эмпирического исследования // Caucasian Science Bridge. 2022. №4. С. 126-135.

2. Деструктивный контент: что это и куда о нем сообщать // Лига безопасного Интернета URL: ligainternet.ru/destruktivnyj-kontent-chto-eto-i-kuda-o-nem-soobshhat/ (дата обращения 20.06.2024).

3. Zelensky A., Cherkesova L., Revyakina E., Kulikova O., Trubchik I. Searching for and blocking destructive content in images // International Scientific Conference «Fundamental and Applied Scientific Research in the Development of Agriculture in the Far East» (AFE-2022). 2023. №371. URL: e3s-

conferences.org/articles/e3sconf/abs/2023/08/e3sconf_afe2023_01056/e3sconf_afe2023_01056.html.

4. Шуликов К.А. Деструктивный контент: понятие, административно-правовая характеристика, виды // Вестник ННГУ. 2023. №2. С.176-182.

5. Альбо Т. М. Имад К. Пропагандистская деятельность ИГИЛ и методы борьбы с ней // Вестник ВУиТ. 2022. №2. С. 110-121.

6. Шестерина А. М. Влияние технологий искусственного интеллекта на видеопроизводство в сфере продвижения сетевого контента // Вестник ЮУрГУ. Серия: Социально-гуманитарные науки. 2022. №1. С. 108-113. URL: vestnik.susu.ru/humanities/issue/view/734.

7. Макаренко С.И. Информационное противоборство и радиоэлектронная борьба в сетецентрических войнах начала XXI века. Санкт-Петербург: Наукоемкие технологии, 2017. 546 с.

8. Никишин В.Д. Вредоносная информация в интернет-медиа: "окно овертона" и взаимосвязь деструктивных сетевых течений // Lex Russica. 2022. №11 (192). С. 131-148.

9. Зеленский А.А., Григорян А.И., Черкесова Л.В., Ревякина Е.А. Разработка монитора безопасности от деструктивных влияний веб-сайтов и социальных сетей интернета // Вестник НГУ. Серия: Информационные технологии. 2021. №1. С. 48–60. URL: intechngu.elpub.ru/jour/issue/view/15/showToc.

10. Калининский Д.С. Сравнительный анализ антивирусных решений для обеспечения безопасности // Международный журнал гуманитарных и естественных наук. 2023. №7-1 (82). С. 203-207. URL: intjournal.ru/wp-content/uploads/2023/08/Mezhdunarodnyj-ZHurnal-7-1.pdf.

11. Google SafeSearch // Google Центр безопасности URL: safety.google/search/ (дата обращения 22.06.2024).

12. Terven J., Cordova-Esparza D., Romero-González J. A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS // Machine Learning and Knowledge Extraction 5. 2023. №4. C. 1680-1716. URL: mdpi.com/2504-4990/5/4/83.

References

1. Titor S. E., Kameneva T. N. Caucasian Science Bridge. 2022. №4. p. 126-135.

2. Destruktivny`j kontent: chto e`to i kuda o nem soobshhat` [Destructive content: what is it and where to report it]. Liga bezopasnogo Interneta URL: ligainternet.ru/destruktivnyj-kontent-chto-eto-i-kuda-o-nem-soobshhat/ (accessed 06/20/2024).

3. Zelensky A., Cherkesova L., Revyakina E., Kulikova O., Trubchik I. International Scientific Conference «Fundamental and Applied Scientific Research in the Development of Agriculture in the Far East» (AFE-2022). 2023. №371. URL: e3s–conferences.org/articles/e3sconf/abs/2023/08/e3sconf_afe2023_01056/e3sconf_afe2023_01056.html.

4. Shulikov K.A. Vestnik NNGU. 2023. №2. p.176-182.

5. Al`bo T. M. Imad K. Vestnik VUiT. 2022. №2. p. 110-121.

6. Shesterina A. M. Vestnik YuUrGU. Seriya: Social`no-gumanitarny`e nauki. 2022. №1. p. 108-113. URL: vestnik.susu.ru/humanities/issue/view/734.

7. Makarenko S.I. Informacionnoe protivoborstvo i radioe`lektronnaya bor`ba v setecentricheskix vojnax nachala XXI veka [Information warfare and electronic warfare in the network-centric wars of the early 21st century]. Sankt-Peterburg: Naukoemkie texnologii, 2017. 546 p.

8. Nikishin V.D. Lex Russica. 2022. №11 (192). p. 131-148.

9. Zelenskij A.A., Grigoryan A.I., Cherkesova L.V., Revyakina E.A. Vestnik NGU. Seriya: Informacionny`e texnologii. 2021. №1. p. 48–60. URL: intechngu.elpub.ru/jour/issue/view/15/showToc.

10. Kalininskij D.S. Mezhdunarodny`j zhurnal gumanitarny`x i estestvenny`x nauk. 2023. №7-1 (82). p. 203-207. URL: intjournal.ru/wp-content/uploads/2023/08/Mezhdunarodnyj-ZHurnal-7-1.pdf.

11. Google SafeSearch. Google Centr bezopasnosti URL: safety.google/search/ (accessed 06/22/2024).

12. Terven J., Cordova-Esparza D., Romero-González J. Machine Learning and Knowledge Extraction 5. 2023. №4. p. 1680-1716. URL: mdpi.com/2504-4990/5/4/83.

Дата поступления: 27.06.2024

Дата публикации: 5.08.2024