Модуль поиска деструктивной информации в тексте

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

Донской Государственный Технический Университет, г. Ростов-на-Дону

Аннотация: Деструктивная информация в тексте довольна распространена и опасна для как для детей и подростков, так и для взрослых. Текущие методы поиска деструктивной информации в тексте: «поиск по ключевым словам», «метод обратной частоты документов» имеют ряд недостатков, из за которых могут быть ложные срабатывания, что понижает точность их работы. В статье рассмотрен новый разработанный метод поиска деструктивной информации в тексте, который используется в модуле Python. Данный метод использует библиотеки Spacy, pymorphy3, что позволяет детально рассматривать предложение и вникать в его смысл. Разработанный метод позволяет уменьшить ложные срабатывания и увеличить тем самым эффективность его использования. В статье показаны схемы разбора предложений, алгоритм работы нового метода, а также рисунки, демонстрирующие его работу. Показан сравнительный анализ нового метода с аналогами.
Ключевые слова: Spacy, деструктивный контент, информационная безопасность, TF-IDF, поиск по ключевым словам, pymorphy3, Net Nanny, CyberPatrol, Окулус, защита детей.

Введение.

Словосочетание «деструктивный контент» у всех на слуху, но далеко не каждый сможет с уверенностью сказать, что это, собственно, такое. И неудивительно, ведь единодушия в данном вопросе нет. В широком смысле к деструктивному контенту относятся все материалы, распространяющиеся в сети, которые могут негативно повлиять на пользователей.

Президент Российской Федерации Владимир Путин, в целях борьбы с деструктивным контентом, поручил создать реестр такого рода информации.

Эксперты, в рамках круглого стола Общественной палаты РФ, проанализировав более 1,5 миллиона сообщений в социальных сетях, чья деятельность запрещена на территории России, выявили девять категорий деструктивного контента в интернете, которые станут основой реестра:

- сексуализация несовершеннолетних [1];
- насилие в отношении людей;
- самоповреждение;
- насилие в отношении животных [2];

– киберунижение [3];

– оккультные услуги, направленные на причинение вреда жизни, здоровью, репутации и имуществу;

– подрыв нормы поведения в семье и школе;

– нетрадиционные модели межполовых отношений и гендерной идентификации;

– ложная информация о пользе или отсутствии вреда от употребления снюсов, вейпов, электронных сигарет и других заменителей никотина.

Директор Регионального общественного центра интернет-технологий Рустам Сагдатулин отметил, что основная проблема такого контента заключается в том, что он не регулируется уголовным или административным правом, а находится в «серой» зоне. «Для нас важно не просто создать рубрикатор, но и добиться существенного сокращения аудитории, которая взаимодействует с таким контентом, уменьшить возможности делиться такой информацией, передавать её друг другу», — заявил он [4].

Таким образом, возникает потребность в разработке специализированных систем поиска и категоризации информационных ресурсов. Кроме того, в связи с постоянно растущим объемом информационных ресурсов автоматизация процесса идентификации характера входной текстовой информации с целью дальнейшей блокировки опасного контента позволит не только сократить трудовременные затраты, но и минимизировать субъективизм и вероятность ошибок, обусловленных влиянием человеческого фактора.

Аналоги.

Существует множество программ и сервисов, которые помогают в поиске и блокировке деструктивного контента в интернете. Некоторые из них:

1. Net Nanny — программа для блокировки сайтов с порнографией, насилием, алкоголем и табаком [5]. Net Nanny — это программное обеспечение для фильтрации интернета, которое помогает родителям контролировать и ограничивать доступ своих детей к определенному контенту и сайтам в Интернете.

Net Nanny работает путем мониторинга интернет-активности на компьютере или мобильном устройстве ребенка и применения заранее определенные правила фильтрации. Эти правила могут включать блокировку сайтов с порнографией, насилием, алкоголем и табаком, а также социальных сетей, онлайн-чатов и других потенциально неприемлемых веб-сайтов.

Программа также может отслеживать и докладывать об активности детей в Интернете, включая попытки доступа к заблокированным сайтам и отправку личной информации. Данную программу родители могут настроить так, чтобы получать всю указанную информацию и в случае чего принять соответствующие меры.

На сколько эффективна программа Net Nanny зависит от возраста и на сколько ребенок опытен, а также какие настройки используются у него на устройстве. Фильтрация интернета не всегда является 100% защитой, например, дети и подростки всегда должны знать о правилах поведения и безопасности в интернете, а также знать, как общаться с подозрительными людьми или с нежелательным контентом.

Достоинства:

– Управление родительским контролем: программа Net Nanny позволяет родителям контролировать и ограничивать доступ своих детей и подростков к специальным web-сайтам, который выберет взрослый. Это помогает обеспечивать безопасность детей в интернете.

– Фильтрация контента: у программы существует возможность фильтровать контент. Данная возможность позволяет блокировать доступ к

нежелательному или вредоносному контенту, среди них сайты с порнографией, насилием и другими деструктивными материалами в сети.

– Мониторинг активности: программа Net Nanny может предоставлять отчеты родителям о посещаемых web-сайтах, и какая была активность в интернете у ребенка. Данная функция может помочь родителям за действиями свои детей в интернете и своевременно реагировать на угрозы.

Недостатки:

– Ложные срабатывания: в редких случаях Net Nanny может блокировать web-сайты с полезной или не деструктивной информацией, так как фильтры программы не могут быть всегда точными. Это может привести к неудобствам при использовании интернета на компьютере.

– Возможность обхода: программа Net Nanny представляет из себя родительский контроль, но некоторые пользователи, могут находить способы, чтобы обойти блокировку. Это может уменьшить эффективность программы для ограничения деструктивного контента.

– Ограниченные возможности: программа Net Nanny представляет только базовые функции контроля и фильтрации контента в интернете. Если необходимы какие-то дополнительные функции, например, мониторинг сообщений или блокировка определенных сообщений, то лучше обратиться к другим программам.

2. CyberPatrol - программа блокировки деструктивного контента, такого как: порнография, насилие, алкоголь, табак, а также социальных сетей и других онлайн-сервисов [6]. CyberPatrol - это программное средство, которое позволяет блокировать деструктивный контент в интернете. В данной программе в качестве инструмента, используется родительский контроль. Особенности CyberPatrol:

– Блокировка вредоносного контента: программа CyberPatrol дает возможность блокировать доступ к web-сайтам с нежелательным или

вредоносным контентом, например, порнография, насилие или наркотики. Эта возможность может помочь создать безопасную сетевую среду для детей и подростков, и предотвратить деструктивное воздействие на них.

– Фильтрация контента: в программе присутствует возможность фильтрации контента. Она позволяет настраивать блокировку определенных категорий сайтов или ключевых слов. Можно выбрать какой тип контента может быть заблокирован, чтобы соответствовать предпочтениям ребенка.

– Мониторинг активности: программа CyberPatrol помогает отслеживать активность детей и подростков в интернете. Также есть возможность проверить какие web-сайты посещал ребенок или какие приложения использовал. Это позволяет родителям всегда быть в курсе того, заходил ли их ребенок на деструктивный сайт или нет.

– Управление временем: программа предлагает функции управления временем использования интернета для детей и подростков. Данная функция позволяет родителям устанавливать время, которое ребенок может провести в интернете, а после истечения времени интернет пропадет на устройстве.

Можно отметить, что программа CyberPatrol - это инструмент родительского контроля. Эффективность программы зависит от регулярных обновлений и настройки фильтров, которые устанавливают родители. Как и в любом другом программном обеспечении, есть возможность ложного срабатывания или обхода системы, поэтому важно, вместе с использованием CyberPatrol обучать ребенка безопасно использовать интернет.

Достоинства:

– Блокировка деструктивного контента: приложение CyberPatrol позволяет блокировать доступ к web-сайтам, которые содержать деструктивный контент, такой как: порнография, насилие, наркотики и другой деструктивный контент. Данная возможность приложения помогает создать безопасное окружение для детей и подростков.

– Фильтрация контента: в приложении множество различных фильтров, которые позволяют родителям выбирать, какой вид контента должен отображаться ребенку, а какой должен быть заблокирован. Также программа дает возможность выбрать категории для блокировки или написать ключевые слова, по которым будет производиться блокировка контента.

– Мониторинг и отчеты: приложение CyberPatrol дает возможность отследить активность ребенка в интернете: web-сайты и приложения. Также приложение имеет возможность формировать отчеты, для дальнейшего анализа родителями.

Недостатки:

– Ложные срабатывания: в редких случаях приложение CyberPatrol может некорректно классифицировать web-сайты, из-за чего происходит блокировка полезного и безопасного контента. Это может создавать дискомфорт при работе в сети интернет.

– Обход системы: существуют опытные пользователи (включая подростков), которые могут обходить блокировки и ограничения приложения CyberPatrol, что уменьшает эффективность программы при блокировки деструктивного контента.

– Ограниченные возможности: приложение CyberPatrol предлагает только базовые функции блокировки и фильтрации контента. Если необходимы дополнительные функции, например, мониторинг сообщений или блокировка определённых приложений, то для этого понадобится другое программное обеспечение.

3. «Окулус» - система, разработанная Роскомнадзором, которая занимается автоматическим поиском деструктивного контента [7]. Система имеет функцию распознавания изображений и символов, которые изображены на картинке. Также одной из функций является анализ на

изображениях и видеоматериалах текста и противоправных действий. «Окулус» может автоматически обнаружить следующие правонарушения: экстремистская тематика, призывы к массовым незаконным материалам, суициду, пронаркотический контент, пропаганда ЛГБТ. Системе массовой информации представители Окулуса рассказали, что система предполагает классификацию изображений и видеороликов по заданным требованиям, которые включают основные типы деструктивного контента. По словам представителей, система «Окулус» работает как классификатор с заданным набором источников информации, с помощью которых производится анализ контента на соблюдения законодательства Российской Федерации. Система может проверять и анализировать конкретные страницы web-сайтов, пабликов, профилей в соцсетях. Программный комплекс не занимается сбором данных, а только лишь классифицирует их.

Все вышеуказанные программы и сервисы помогают родителям и другим пользователям защитить детей и подростков от деструктивного контента в сети Интернет. Важно помнить, что данные программы должны использоваться с умом и не полагаться на их 100% защиту.

Разработанное программное обеспечение.

В начале программа разбивает текст на предложения и для каждого предложения выполняется предварительная обработка текста.

Алгоритм предварительной обработки текста включает в себя следующие шаги, показанные на рис. 1:

1. Преобразование в нижний регистр слов.

2. Лемматизация: лемматизацией называют процесс приведения слов к их базовой форме [8]. Например, слова "бегать", "бегаю", "бегать" можно привести к одной базовой форме, слову "бегать". Процесс лемматизации может помочь уменьшить количество разных форм одного и того же слова.

3. Производится замена цифр на слова, например, «7» на «семь».

4. Удаление стоп-слов: под стоп-словами понимаются такие слова, которые не несут смысловой нагрузки и часто встречаются в текстах [9]. Стоп-словами обычно бывают: предлоги, союзы, местоимения и другие слова. Удаление стоп-слов позволяет убрать лишние слова для анализа текста и сконцентрироваться на основных деталях.

Программный модуль использует для поиска деструктивного контента в тексте 2 метода одновременно:

– Известный метод поиска обратной частоты документа (TF-IDF).

– Новый разработанный метод.

Метод, который на основе содержания текста в документе производит его классификацию, называется методом классификации текста обратной частоты документа (TF-IDF) [10]. Данный метод основан на том, что слова, которые часто встречаются в предложении, но редко встречаются в других предложениях, являются важными для содержания этого предложения.

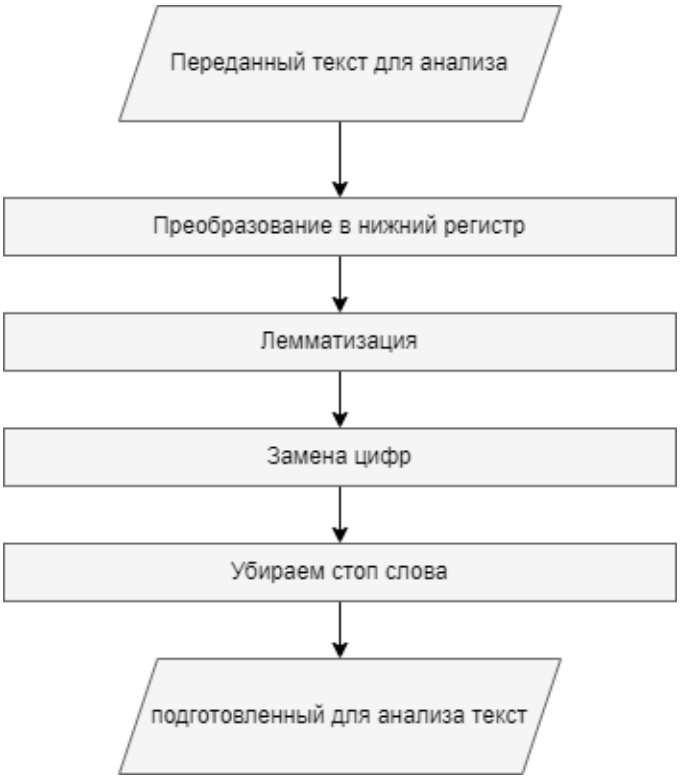

Рис. 1. - Алгоритм предварительной обработки текста в разработанной программе

В методе TF-IDF используются два показателя:

– Частота слова в предложении (TF) - это количество раз, которое слово встречается в данном предложении [11].

– Обратная частота документа (IDF) - это обратная величина частоты слова во всех предложениях [12].

Метод TF-IDF основан на вычислении коэффициента важности каждого слова в предложении, умножая его частоту на обратную частоту документа. Чем выше коэффициент важности, тем более важным является слово для описания содержания документа.

Формула обратной частоты слова в тексте показана формулой ниже.

$$IDF(t) = \log\left(\frac{N}{df(t)}\right)$$

Формула частоты слова t в предложении d показана формулой ниже.

$$\mathrm{TF}(t, d) = \frac{\text{количество вхождений слова } t \text{ в предложении } d}{\text{общее количество слов в предложении } d}$$

Формула расчета TF-IDF показана формулой ниже.

$$TF - IDF(t, d) = TF(t, d) * \log\left(\frac{N}{df(t)}\right),$$

где *IDF(t)* - обратная частота текста; *TF(t, d)* - частота слова *t* в предложении *d*; *df(t)* - количество предложений, в которых встречается слово *t*; *N* - общее количество предложений в коллекции.

TF-IDF придает больший вес слову, которое редко встречается в тексте.

Суть нового метода поиска деструктивного контента в тексте основан на разбитии предложения на части речи, и анализа, насколько к существительному можно применить деструктивный глагол. Расчет деструктивности предложения новым методом показан формулой 1.

$$D = \lg\left(exp^{(x+y)}\right), \quad (1)$$

где ***D*** – значение деструктивности текста; ***x*** – значение возможности применения деструктивного глагола к существительному; ***y*** – значение деструктивности глагола.

Рассмотрим частные случаи расчета формулы (1). Частные случаи представлены в таблице 1.

Таблица № 1

Частные случаи расчета формулы (1)

x, %	***y,%***	***D,%***
10	10	8,68589
20	20	17,37178
30	30	26,05767
40	40	34,74356
50	50	43,42945
60	60	52,11534
70	70	60,80123
80	80	69,48712
90	90	78,17301
100	100	86,8589

На рис. 2 изображен график функции по координатам, указанным в таблице 1.

Сокращения имеют следующее описание:

PROPN: имя собственное, например, Мэри, Джон.

VERB: глагол, например, run, бежит.

NOUN: существительное, например, девушка, кошка.

ADJ: прилагательное, например, большой.

ADP: дополнение, например, в, к, во время.

NUM: числительное, например, 1.

Рис. 2. - График функции lg(exp(x+y))

На рис. 3 и рис. 4 показан пример разбора предложений новым методом.

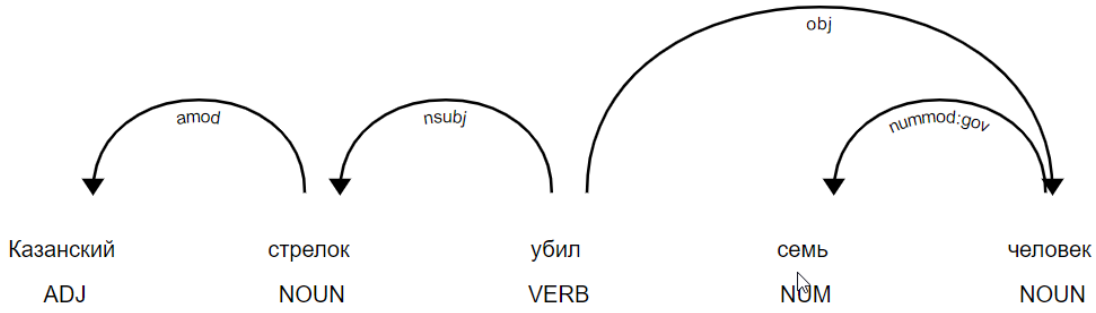

Рис. 3. – Пример разбора деструктивного предложения

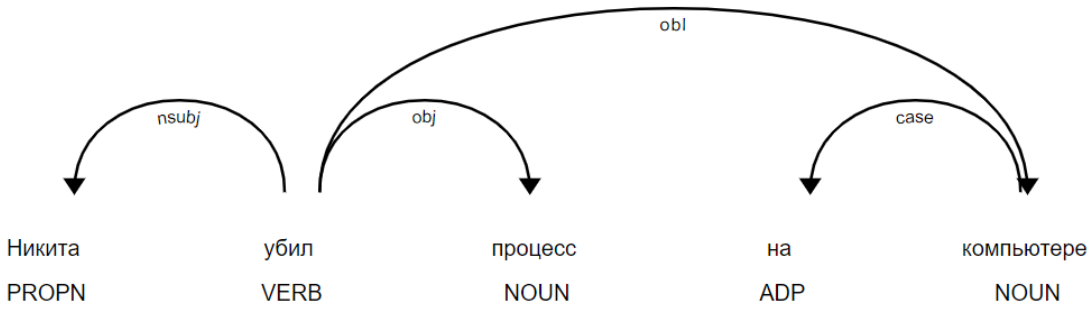

Рис. 4. – Пример разбора недеструктивного предложения

Подлежащее (nsubj) – именное подлежащее (субъект), первый аргумент предиката с самым высоким синтаксическим статусом.

Дополнение (obj) – второй именной аргумент предиката, обычно объектный. В типичном случае отношение связывает глагол в действительном залоге с именным зависимым в винительном падеже (прямым дополнением).

Атрибутивный модификатор (amod) – имя прилагательное, прилагательное-числительное или причастие, модифицирующее другое имя. Атрибутивный модификатор может стоять до и после главного элемента и обычно согласуется с ним по роду, числу, падежу и одушевленности.

Косвенное дополнение / обстоятельство (obl) – связывает синтаксически периферийные зависимые – косвенные аргументы или обстоятельства – с вершиной.

Количественный модификатор (nummod) – отношение, связывающее числительное в роли квантификатора (а также существительные тысяча, миллион, миллиард и т.п.) с синтаксическим хозяином, обозначающим квантифицируемое.

Предлог (case). Предлоги представляются как зависимые от имени, которым они управляют или которое вводят.

Для примера, можно разобрать предложение с рис. 3. Глаголом является слово «убил», далее программу интересует, к какому существительному применяется этот глагол. На рис. 3 - это «Дополнение (obj)» и существительным является слово «человек». К данному существительному можно с высокой вероятностью применить деструктивный глагол, что и показано в примере, где слово «убил» является деструктивным глаголом.

Разберем предложение с рис. 4. Глаголом является слово «убил», далее программу интересует, к какому существительному применяется этот глагол. На рис. 4 - это «Дополнение (obj)» и существительным является слово «процесс». К данному существительному невозможно применить деструктивный глагол, что и показано в примере, где слово «убил» является деструктивным глаголом, но само предложение не является деструктивным.

Общая схема работы разработанного модуля показана на рис. 5.

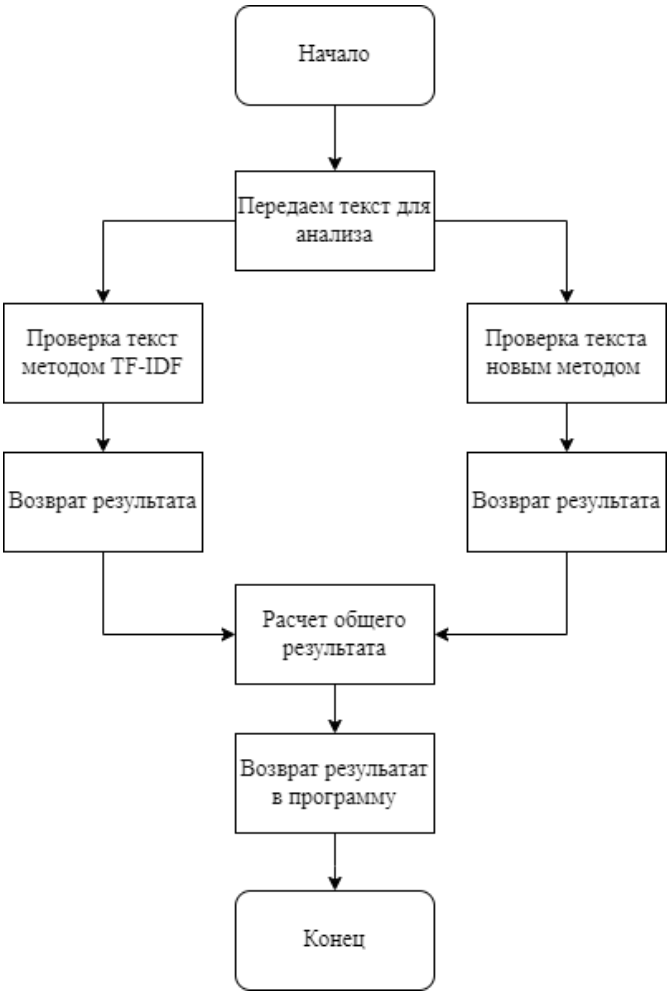

Рис. 5. – Общая схема работы разработанного модуля поиска деструктивной информации в тексте

Преимущества нового метода над другими известными:

– разбор предложения на части речи для анализа;

– не используется «поиск по ключевым словам»;

– не используется база данных (не требуется постоянная актуализация);

– деструктивный глагол в предложении не влияет на результат;

– более точное определение деструктивности предложения;

– гибкий метод, данные, полученные при помощи этого метода, возможно использовать для других анализов.

В таблице 2 показано сравнение классификации известных методов недеструктивных предложений на их деструктивность.

Таблица № 2

Сравнение классификации известных методов недеструктивных предложений на их деструктивность

Метод	Количество недеструктивных предложений, шт.	Количество правильной классификации, шт.
Новый метод	100	100
TF-IDF		95
По ключевым словам		83

В таблице 3 показано сравнение классификации известных методов деструктивных предложений на их деструктивность.

Таблица № 3

Сравнение классификации известных методов деструктивных предложений на их деструктивность

Метод	Количество деструктивных предложений, шт.	Количество правильной классификации, шт.
Новый метод	100	100
TF-IDF		98
По ключевым словам		100

Заключение.

Доказана эффективность разработанного нового модуля для поиска деструктивной информации в тексте за счет комбинированного использования нескольких методов классификации текста и использования нового метода поиска деструктивной информации.

Основные результаты заключаются в следующем:

1. Проанализированы современные методы и системы для обнаружения деструктивной информации в тексте, и, как основной общий недостаток, обозначено, что они фокусируются в основном на:

– бинарной классификации текста, когда значение деструктивности либо равно 0, либо равно 1 без детального разбора данных;

– использовании баз данных для хранение деструктивных сайтов, что является плохим решением, так как любая база данных требует постоянной актуализации.

2. Предложен метод и алгоритм для эффективного обнаружения деструктивного контента в тексте. Разработанный модуль поиска деструктивного контента в тексте анализирует части речи подробно в предложении и рассчитывает процент деструктивности предложения, потом на основе результата производится классификация, определяющая, деструктивен контент или нет.

3. Проведены экспериментальные исследования предложенного метода и модели в области обнаружения деструктивного контента в тексте, в результате чего показано, что точность определения деструктивного контента в сравнении с другими методами на 5% точнее, что является более высоким уровнем по сравнению с конкурирующими методами.

Литература

1. Юмашева И.А. Семейные ценности как инструмент внешней политики России // Культурологический журнал. 2021. №3 (45). URL: sciup.org/semejnye-cennosti-kak-instrument-vneshnej-politiki-rossii-170177122-en.

2. Осипова Н. Р., Васильев А. М. Актуальные вопросы криминализации жестокого обращения с животными: особенности законодательной конструкции и квалификации // Вестник науки и творчества. 2024. №5 (96). URL: elibrary.ru/item.asp?id=67855248.

3. Dilmac J. A. Humiliation in the Virtual World: Definitions and Conceptualization // Journal of Human Sciences. 2014. № 11(2). С. 1285-1296. URL: researchgate.net/publication/275535832_Humiliation_in_the_Virtual_World_Definitions_and_Conceptualization.

4. Эксперты назвали 9 типов деструктивно влияющего на психику детей сетевого контента // Газета.ru. URL: gazeta.ru/science/news/2022/11/15/19044343.shtml (дата обращения: 14.07.2024).

5. The Best Parental Control to Keep Your Kids and Family Safe Online URL: netnanny.com/ (дата обращения: 14.07.2024).

6. Welcome-SafeToNet // CyberPatrol URL: safetonet.com/ (дата обращения: 14.07.2024).

7. Зиборов О.В., Аветисян К.Р. Комплексный подход при противоборстве организации массовых беспорядков // Вестник Московского университета МВД России. 2023. №7. С. 100-104.

8. Khyani D., Siddhartha B.S. An Interpretation of Lemmatization and Stemming in Natural Language Processing // Journal of University of Shanghai for Science and Technology. 2021. № 22(10). С. 350-357.

9. Челышев Э. А., Оцоков Ш. А., Раскатова М. В., Щёголев П. Сравнение методов классификации русскоязычных новостных текстов с

использованием алгоритмов машинного обучения // Вестник кибернетики. 2022. №1 (45). URL: vestcyber.ru/jour/article/view/417.

10. Бушуев Е.М. Обзор подходов кластеризации поисковых ключевых фраз по семантической схожести методами машинного обучения // Вестник науки. 2023. №12 (69). С. 392-402.

11. Боровский А. В., Раковская Е. Е., Бисикало А. Л. Классификация коротких технических текстов с применением системы нечеткого вывода сугено // Вестник АГТУ. Серия: Управление, вычислительная техника и информатика. 2021. №1. С. 16-27.

12. Катермина Т.С., Тагиров К.М., Тагиров Т.М. Элементы искусственного интеллекта в решении задач анализа текстов // Computational nanotechnology. 2022. №2. С. 35-44.

References

1. Yumasheva I.A. Kul`turologicheskij zhurnal. 2021. №3 (45). URL: sciup.org/semejnye-cennosti-kak-instrument-vneshnej-politiki-rossii-170177122-en.

2. Osipova N. R., Vestnik nauki i tvorchestva. 2024. №5 (96). URL: elibrary.ru/item.asp?id=67855248.

3. Dilmac J. A. Journal of Human Sciences. 2014. № 11(2). pp. 1285-1296. URL: researchgate.net/publication/275535832_Humiliation_in_the_Virtual_World_Definitions_and_Conceptualization.

4. E`ksperty` nazvali 9 tipov destruktivno vliyayushhego na psixiku detej setevogo kontenta [Experts named 9 types of network content that destructively affects the psyche of children]. Gazeta.ru. URL: gazeta.ru/science/news/2022/11/15/19044343.shtml (accessed: 07/14/2024).

5. The Best Parental Control to Keep Your Kids and Family Safe Online URL: netnanny.com (date of access: 07/14/2024).

6. Welcome-SafeToNet URL: safetonet.com / (date of address: 07/14/2024).

7. Ziborov O.V., Avetisyan K.R. Vestnik Moskovskogo universiteta MVD Rossii. 2023. №7. pp. 100-104.

8. Khyani D., Siddhartha B.S. Journal of University of Shanghai for Science and Technology. 2021. № 22(10). pp. 350-357.

9. Chely`shev E`. A., Oczokov Sh. A., Raskatova M. V., Shhyogolev P. Vestnik kibernetiki. 2022. №1 (45). URL: vestcyber.ru/jour/article/view/417.

10. Bushuev E.M. Vestnik nauki. 2023. №12 (69). pp. 392-402.

11. Borovskij A. V., Rakovskaya E. E., Bisikalo A. L. Vestnik AGTU. Seriya: Upravlenie, vy`chislitel`naya texnika i informatika. 2021. №1. pp. 16-27.

12. Katermina T.S., Tagirov K.M., Tagirov T.M. Computational nanotechnology. 2022. №2. pp. 35-44.

Дата поступления: 17.06.2024

Дата публикации: 25.07.2024