ИЗДАТЕЛЬСТВО грамота

Филологические науки. Вопросы теории и практики
Philology. Theory & Practice

ISSN 2782-4543 (online)
ISSN 1997-2911 (print)

2023. Том 16. Выпуск 11 | 2023. Volume 16. Issue 11
Материалы журнала доступны на сайте (articles and issues available at): philology-journal.ru

RU

Программные инструменты создания и анализа массивов текстов коротких электронных сообщений пользователей социальных сетей

Логинова А. О., Горожанов А. И., Алейникова Д. В.

Аннотация. В рамках исследования преследуется цель разработки алгоритма создания и анализа массива текстов коротких электронных сообщений (постов) в социальных сетях с помощью общедоступных программных инструментов. Научная новизна состоит в том, что для решения подобной проблемы применяется междисциплинарный подход, учитывающий последние достижения прикладной и математической лингвистики и информационной безопасности, с привлечением актуальной нормативной базы. В ходе работы, согласно предложенной графической модели, посредством плагина Web Scraper был собран текстовый материал исследования объемом около 1,5 МБ; сформирован массив текстов коротких электронных сообщений, конвертированный в пригодный для дальнейшей обработки формат CSV; проведен базовый анализ этого массива текстов посредством общедоступного программного комплекса PolyAnalyst, который включил такие процедуры, как извлечение терминов, сущностей и ключевых слов, анализ тональности и определение тематики текстов. В результате была доказана функциональность созданного алгоритма, определены перспективы дальнейших исследований – работа с текстовыми данными большого объема и анализ этих данных для нахождения в них деструктивного контента.

EN

Software tools for creating and analyzing a text data bank of short electronic messages from social network users

Loginova A. O., Gorozhanov A. I., Aleynikova D. V.

Abstract. The research aims at developing an algorithm for creating and analyzing a text data bank of short electronic messages (posts) from social networks using free software tools. The scientific novelty lies in the fact that to solve such a problem, an interdisciplinary approach is used, taking into account the latest achievements of applied and mathematical linguistics and information security, with the involvement of the current regulatory framework. In the course of the work, according to the proposed graphical model, textual research material of ca. 1.5 MB was collected using the Web Scraper plug-in; a text data bank of short electronic messages was generated, converted into a CSV format suitable for further processing; a basic analysis of this data bank was carried out using PolyAnalyst free software package, which included such procedures as the extraction of terms, entities and keywords, sentiment analysis and determination of the subject matter of texts. As a result, the functionality of the created algorithm was proven, prospects for further research were identified – working with big text data and analyzing this data to find destructive content in them.

Введение

Сегодня значительное количество исследований в сфере искусственного интеллекта привлекает в качестве инструментов программные решения для обработки естественного языка (англ. natural language processing, NLP) (Токтарова, Попова, Сагдуллина и др., 2023), которые применяются, например, для разработки проблемы обучения нейронной сети на больших объемах данных естественного языка, что, в свою очередь, требует высокого уровня формализации текста.

Актуальность данного исследования обусловлена тем, что современный этап развития технологий требует обработки не только объемных текстов, но также и веб-контента, в частности текстов коротких электронных сообщений пользователей социальных сетей. Социальные сети являются в наши дни, хотели бы мы этого или нет, неотъемлемой частью повседневной жизни и представляют собой (в широком понимании) платформы для обеспечения общения в виртуальном пространстве, в том числе для обмена различными типами цифровых данных, а также создания контента, включая текстовые сообщения, фотографии или видеофрагменты (Islam, Latif, Ahmed, 2019). При этом предлагаемый контент далеко не всегда может рассматриваться

Научная статья (original research article) | https://doi.org/10.30853/phil20230560

как безопасный (Шуликов, 2023; Джаффарова, 2021). Распространение деструктивного контента в социальных сетях становится серьезной социальной проблемой. Подобный контент содержит призывы к разжиганию ненависти, оскорбительные выражения, дезинформацию, спам, насилие, деструктивный графический контент и др. Социальные сети стремятся контролировать содержание с целью предотвращения серьезных последствий. Однако масштабы реализации вредоносной информации возрастают. Согласно данным, опубликованным Институтом Алана Тьюринга и организацией Ada, почти 90% людей в возрасте от 18 до 34 лет хотя бы раз сталкивались с деструктивным контентом в Интернете (Почти 90% молодых людей подвергаются воздействию вредоносного контента в соцсетях (RSpectr. 21.03.2023. https://rspectr.com/novosti/pochti-90-molodyh-lyudej-podvergayutsya-vozdejstviyu-vredonosnogo-kontenta-v-soczsetyah?utm_source=telegram%29которого). Важность регулирования процессов в интернет-средствах массовой коммуникации (далее – интернет-СМК) обусловлена возросшим интересом к социальным сетям как к площадке для реализации политических и маркетинговых сценариев. Отсюда государства столкнулись с необходимостью блокировки и разграничения доступа пользователей к контенту, который оценивается как деструктивный на законодательном уровне (Шуликов, 2023).

Задачи исследования формулируются следующим образом:

1) собрать текстовый материал исследования посредством плагина Web Scraper;

2) сформировать массив текстов коротких электронных сообщений и конвертировать его в пригодный для дальнейшей обработки формат;

3) провести базовый анализ сформированного массива текстов посредством программного комплекса PolyAnalyst.

В работе применяются следующие методы: моделирование процесса сбора и обработки массива текстов коротких электронных сообщений (на этапе формирования последовательности действий); простая случайная выборка (для отбора аккаунтов в социальной сети, посты которых войдут в массив текстов коротких сообщений); программная обработка текста (для сбора, накопления и структурирования данных массива текстов).

Материалом исследования послужил массив текстов коротких электронных сообщений пользователей социальных сетей на английском языке, собранный из веб-архива аккаунтов (Internet Archive. https://web.archive.org), которые были заблокированы модераторами как аккаунты социальных ботов, задействованные в предвыборной кампании Дональда Трампа.

Теоретическую базу исследования составляют работы в области прикладной и математической лингвистики (Горожанов, 2023; Потапова, Потапов, 2022; Баранов, 2017), в сфере информационной безопасности (Логинова, Алейникова, 2023; Мамченко, Мещеряков, Галин, 2022; Горожанов, Гусейнова, Писарик, 2022; Минаев, Реброва, Симонов, 2021; Islam, Latif, Ahmed, 2019), а также в области юриспруденции (Шуликов, 2023; Джаффарова, 2021).

Для создания массива текстов коротких электронных сообщений пользователей социальной сети и выделения собственно текста поста и метаданных поста использовался парсер Web Scraper. Базовый анализ сформированного корпуса текстов проведен посредством программного комплекса PolyAnalyst (поясним, что «базовый анализ» является техническим термином PolyAnalyst).

Практическая значимость исследования состоит в возможности использования его материалов для создания лингвистических корпусов данных веб-ресурсов с целью проведения широкого спектра междисциплинарных исследований. Кроме того, представленный в статье алгоритм может быть использован для подготовки иллюстративного материала в рамках чтения учебных дисциплин, посвященных проблемам прикладной и математической лингвистики, интерпретации текста, информационной безопасности и пр.

Обсуждение и результаты

Последовательность создания и анализа массива текстов коротких электронных сообщений представим в виде графической модели (см. Рис. 1).

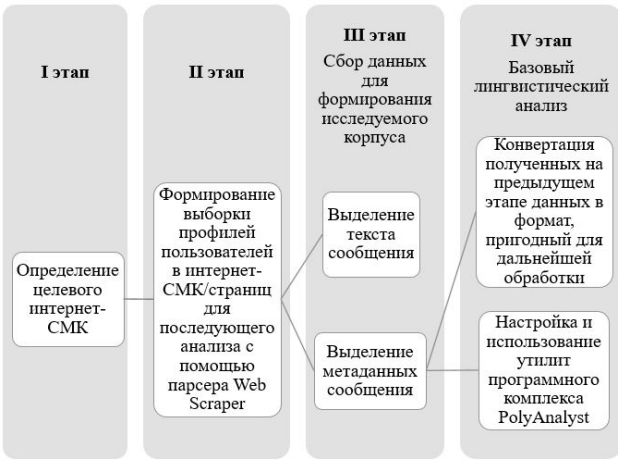

Рисунок 1. *Графическая модель последовательности создания и анализа массива текстов коротких электронных сообщений*

Рассмотрим каждый этап графической модели, реализованный на практике.

На первом этапе необходимо выбрать интернет-СМК, ресурсы которого будут подвергнуты анализу. На втором этапе необходимо определить пул аккаунтов/страниц, информация с которых будет подвергнута анализу. В нашем случае была применена простая случайная выборка ряда аккаунтов социальных интернет-ботов.

Третий этап исследования заключается в непосредственном сборе информации из веб-ресурса с помощью онлайн-парсера Web Scraper. Данный программный продукт позволяет путем корректной настройки правил сбора информации отбирать только необходимую для реализации задач исследования информацию, игнорируя рекламные блоки, тексты всплывающих подсказок и прочие возможные «шумовые» атрибуты социальных сетей. Web Scraper доступен для свободного использования в качестве плагина практически к любому веб-браузеру. В нашем случае использовался веб-браузер Google Chrome. Для запуска плагина необходимо зайти в меню веб-браузера, выбрать вкладку «Дополнительные инструменты» / «Инструменты разработчика». В открывшемся меню нужно выбрать Web Scraper (см. Рис. 2).

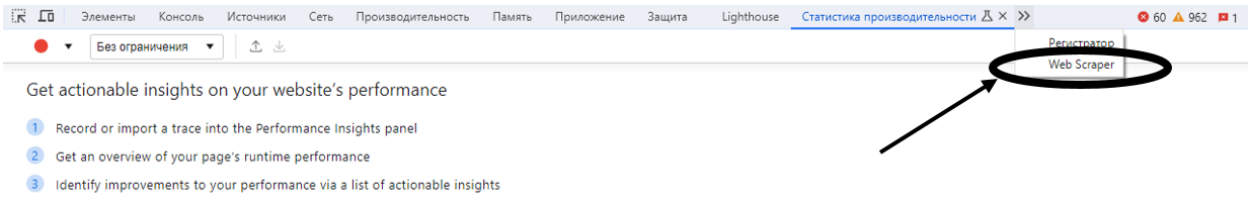

Рисунок 2. *Инструмент Web Scraper в меню разработчика веб-браузера*

В меню выбранного инструмента “Create new sitemap” создается новый набор правил для сбора информации с интересующей исследователя веб-страницы. Для этого вводятся URL или серия URL, если необходимо собрать данные с профиля в социальной сети, который размещен на нескольких веб-страницах, и наименование набора правил (см. Рис. 3).

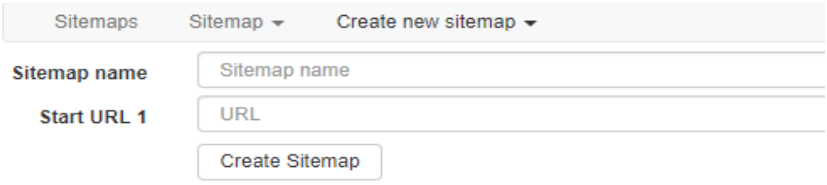

Рисунок 3. *Создание нового набора правил для сбора информации с веб-страницы*

В меню выбранного инструмента каждого нового профиля аккаунта в социальной сети необходимо создавать новый набор правил. На следующем шаге нужно задать группу элементов страницы профиля в социальной сети, информация из которого будет собираться в специальный «контейнер». Для этого в открывшемся окне выбирается “Add new selector”, вносятся запрашиваемые данные (см. Рис. 4).

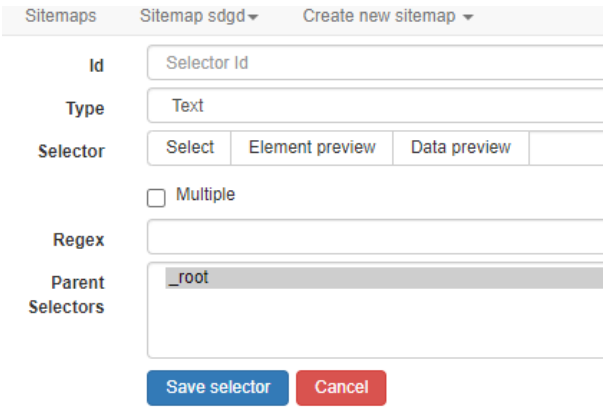

Рисунок 4. *Настройка правил сбора информации в «контейнер»*

В поле “ID” вносится наименование «контейнера». В поле “Type” необходимо выбрать значение “Element”, а с помощью кнопки “Select” – элемент страницы пользователя в социальной сети. Для создания корпуса текстов коротких электронных сообщений (постов) нужно выбрать блок, включающий текст сообщения и его метаданные (см. Рис. 5).

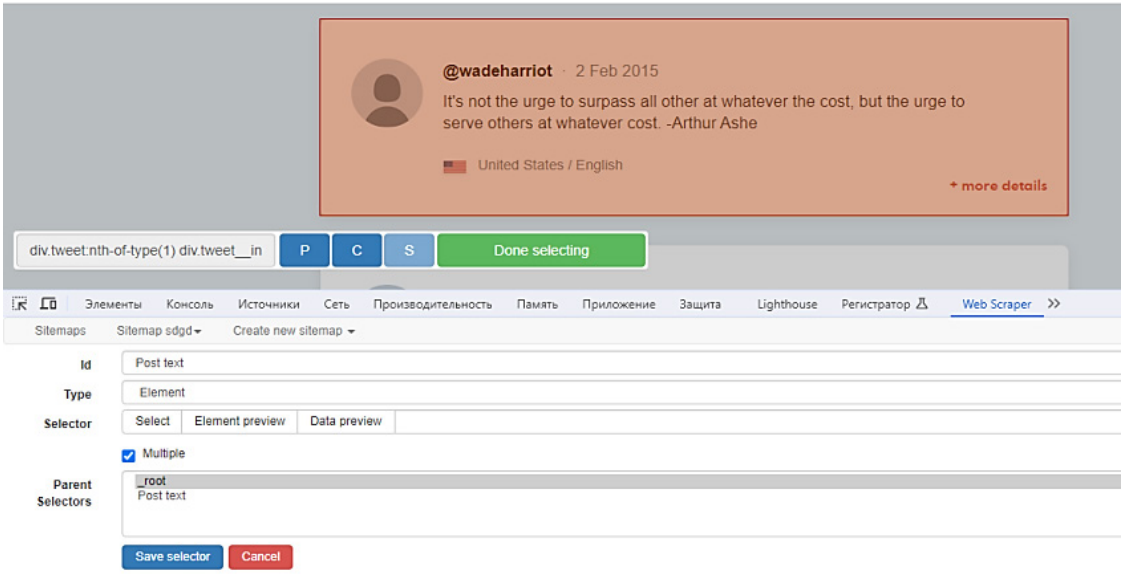

Рисунок 5. *Выбор элемента страницы пользователя социальной сети*

Далее следует закрепление выбора нажатием кнопки “Done selecting”. В случае, если элемент страницы повторяется, например, страница пользователя в социальной сети состоит из множества опубликованных постов, необходимо поставить отметку “Multiple”. Тогда при сборе информации со страницы пользователя в созданный контейнер войдут данные не одного, а всех опубликованных постов. Нажатие на кнопку “Save selector” завершает формирование контейнера.

Для распределения собираемой парсером информации в созданный контейнер “Post text” были созданы следующие категории информации, находящейся в блоке опубликованного сообщения (поста): “Name and date”, “Retweet” и “Message” (см. Рис. 6).

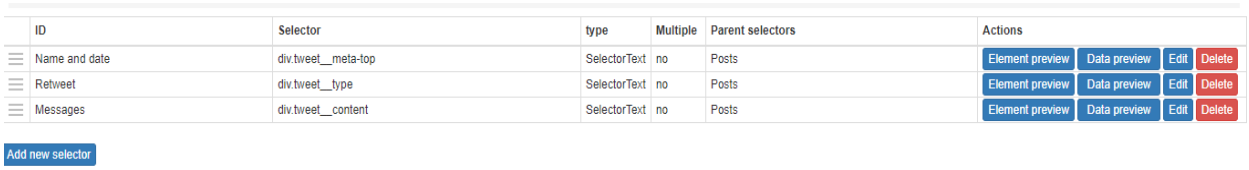

Рисунок 6. *Категории собираемой информации*

После того как контейнер для сбора информации и правила сбора были настроены, был инициирован парсер Web Scraper (выбирается созданный парсер и запускается процедура “Scrap” с интервалом запроса до 5000 мс). По окончании формирования контейнера – массива коротких электронных сообщений и их метаданных – парсер «предложит» сохранить файл в формате XLSX или CSV. Наш опыт работы с программным обеспечением PolyAnalyst показывает, что для дальнейшего анализа корпуса предпочтительно выбрать формат CSV.

На четвертом этапе для проведения базового лингвистического анализа мы воспользовались программным обеспечением PolyAnalyst. Анализ текстовых данных в предлагаемом программном обеспечении производится путем добавления узлов в скрипт и выполнения цепочек этих узлов при предварительном импорте анализируемых данных. Здесь «Узел» – это отдельное действие по обработке данных, например идентификация языков, на которых написаны анализируемые тексты массива; узел также содержит результаты выполнения представляемого им действия (см. Рис. 7).

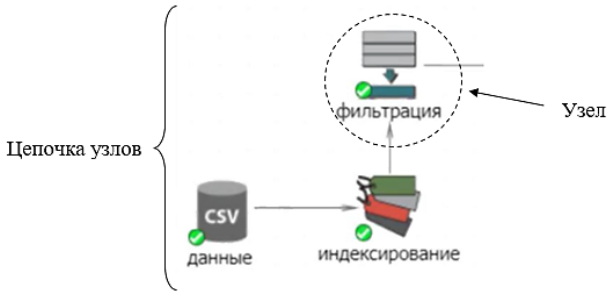

Рисунок 7. *Фрагмент цепочки узлов PolyAnalyst*

Для базовой обработки созданного массива текстовых данных на естественном (английском) языке построим следующую цепочку узлов (см. Рис. 8).

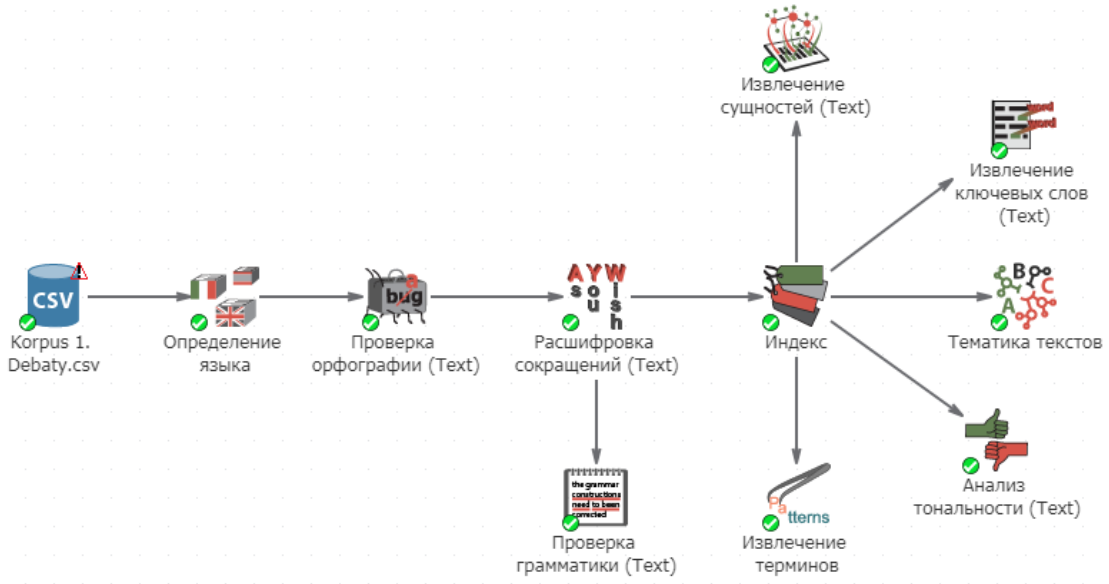

Рисунок 8. *Цепочка узлов PolyAnalyst*

Данная последовательность узлов в цепочке обусловлена порядком проведения базового анализа текста согласно руководству пользователя PolyAnalyst (https://www.megaputer.com/polyanalyst/).

Первый узел в цепочке позволяет импортировать данные из хранилища (память персонального компьютера и т. д.) в среду PolyAnalyst, он представлен в категории «Источники данных» палитры узлов. PolyAnalyst дает возможность импортировать широкий спектр типов и форматов данных, которые конвертируются на входе в табличную форму. В этой связи для удобства работы и простоты обнаружения технических ошибок, возникающих при загрузке данных, в среду PolyAnalyst были импортированы файлы с расширением CSV, содержащие анализируемый текстовый массив.

Результаты выполнения операции в предыдущем узле являются входными данными для последующего узла при линейном расположении узлов.

Второй и последующие узлы выбраны из категории «Текстовый анализ». Фрагмент цепочки, которую составляют узлы от *Импорт данных* до *Индекс*, представляет последовательность узлов для первичной обработки данных. Представленный фрагмент цепочки содержит последовательность узлов, позволяющих:

– определить перечень естественных языков, на которых составлены тексты обрабатываемого массива;
– проверить массив на наличие орфографических, грамматических и других ошибок;
– раскрыть значения аббревиатур для проведения корректного лексического анализа;
– проиндексировать единицы текстов.

Узел *Проверка грамотности* выведен за вектор первичной обработки, поскольку данный узел вычисляет значение параметра, от которого не зависит выполнение последующих узлов.

В узле *Индекс* поступающие на вход текстовые данные структурируются, разбиваются на токены – предложения или слова; собирается информация о каждом слове: статистика употребления, характеристики, условия употребления. Данный узел формирует основу для дальнейшей обработки текстовых данных в узлах *Извлечение сущностей*, *Извлечение ключевых слов*, *Тематика текстов*, *Анализ тональности* и *Извлечение терминов*.

Узел *Извлечение сущностей* позволяет экстрагировать наименования объектов реального мира, упоминающиеся в текстах массива, а также обозначения даты и времени, денежных сумм, номера телефонов, URL и другие объекты.

Итогами работы узла *Извлечение ключевых слов* являются перечень ключевых слов анализируемого массива текстов, их графическое представление в виде облака тегов, распределение по частоте или по значимости, а также результаты классификации ключевых слов по частям речи.

Узел *Тематика текстов* используется для автоматического определения тем, освещаемых в массиве текстов. Узел *Анализ тональности* применяется для поиска конструкций, выражающих эмоциональную оценку. Узел *Извлечение терминов* используется в настоящем исследовании для поиска и подсчета количества буквенных и числовых символов, а также знаков препинания и специальных символов.

Таким образом, был сформирован алгоритм создания и анализа массива текстов коротких электронных сообщений (постов) в социальной сети с помощью общедоступных программных инструментов, что и составляет цель нашей работы.

Заключение

Современный уровень технологического развития подчеркивает возрастающую необходимость работы с интернет-контентом. В данном случае серьезную озабоченность представляет прежде всего деструктивный контент. Сегодня распространенным направлением исследования подобного материала является машинное обучение, что требует серьезной профессиональной технической подготовки.

Однако, в отличие от широко применяемых в технических исследованиях методов поиска деструктивного контента на основе моделей машинного обучения, предложенный нами алгоритм с привлечением общедоступного программного обеспечения может быть использован для анализа контента веб-ресурсов более широким кругом исследователей в силу простоты реализации процедуры сбора и обработки данных, а также отсутствия необходимости иметь глубокие познания в области программирования.

В настоящей статье был описан алгоритм автоматизированного создания и обработки массива текстовых данных, полученных с выбранного веб-ресурса. Во-первых, с помощью онлайн-парсера Web Scraper был собран, накоплен и структурирован текстовый массив объемом около 1,5 МБ, имеющий потенциал для дальнейшей обработки. Во-вторых, посредством программного комплекса PolyAnalyst был проведен базовый анализ сформированного массива текстов.

Авторы проведенного исследования видят перспективу использования предложенного алгоритма для анализа больших объемов текстовых данных с целью идентификации в них деструктивного вредоносного контента и его купирования программными средствами.

Источники | References

1. Баранов А. Н. Лингвистика в лингвистической экспертизе (метод и истина) // Вестник Волгоградского государственного университета. Серия 2: Языкознание. 2017. Т. 16. № 2. https://doi.org/10.15688/jvolsu2.2017.2.2
2. Горожанов А. И. Создание лингвистического корпуса на основе инструментов обработки естественного языка: планирование программных решений // Филологические науки. Вопросы теории и практики. 2023. Т. 16. Вып. 5. https://doi.org/10.30853/phil20230252
3. Горожанов А. И., Гусейнова И. А., Писарик О. И. Уровневая модель информационной безопасности в условиях виртуального пространства // Вестник МГПУ. Серия: Филология. Теория языка. Языковое образование. 2022. № 2 (46). https://doi.org/10.25688/2076-913X.2022.46.2.11
4. Джаффарова Н. Т. Административная ответственность за правонарушения в области оборота информации: дисс. … к. юрид. н. М., 2021.
5. Логинова А. О., Алейникова Д. В. Выявление демаскирующих признаков социального бота на синтаксическом уровне генерируемого сообщения // Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии. 2023. № 1. https://doi.org/10.17308/sait/1995-5499/2023/1/139-147
6. Мамченко М. В., Мещеряков Р. В., Галин Р. Р. Социокиберфизическая система для выявления и блокирования деструктивного интернет-контента // Современные проблемы радиоэлектроники и телекоммуникаций. 2022. № 5.
7. Минаев В. А., Реброва А. Д., Симонов А. В. Выявление деструктивного контента в социальных медиа на основе моделей машинного обучения // Информация и безопасность. 2021. Т. 24. № 1.
8. Потапова Р. К., Потапов В. В. Интернет-меметика как эмоциогенная среда сетевой коммуникации // Известия Российской академии наук. Серия литературы и языка. 2022. Т. 81. № 2. https://doi.org/10.31857/S160578800019458-9
9. Токтарова В. И., Попова О. Г., Сагдуллина И. И., Белянин В. А. Технологии искусственного интеллекта в практике современного высшего образования // Вестник Марийского государственного университета. 2023. № 2 (50).
10. Шуликов К. А. Деструктивный контент: понятие, административно-правовая характеристика, виды // Вестник Нижегородского университета им. Н. И. Лобачевского. 2023. № 2.
11. Islam T., Latif S., Ahmed N. Using Social Networks to Detect Malicious Bangla Text Content // 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT). Dhaka, 2019.

Финансирование | Funding

RU Публикация подготовлена в рамках государственного задания на проведение научно-исследовательских работ № FSFU-2020-0020 «Перспективные технологии реализации информационной функции государства и обеспечения цифрового суверенитета».

EN The reported study was carried out as a part of state assignment to conduct scientific research No. FSFU-2020-0020 “Promising technologies for implementing the information function of the state and ensuring digital sovereignty”.

Информация об авторах | Author information

RU

Логинова Алина Олеговна[1]
Горожанов Алексей Иванович[2], д. филол. н., доц.
Алейникова Дарья Викторовна[3], к. пед. н.
[1, 2] Московский государственный лингвистический университет
[3] Московский государственный лингвистический университет;
Российский университет дружбы народов, г. Москва

EN

Loginova Alina Olegovna[1]
Gorozhanov Alexey Ivanovich[2], Dr
Aleynikova Darya Viktorovna[3], PhD
[1, 2] Moscow State Linguistic University
[3] Moscow State Linguistic University;
Peoples' Friendship University of Russia, Moscow

[1] *a.loginova@linguanet.ru,* [2] *a.gorozhanov@linguanet.ru,* [3] *festabene@mail.ru*

Информация о статье | About this article

Дата поступления рукописи (received): 12.09.2023; опубликовано online (published online): 25.10.2023.

Ключевые слова (keywords): корпусная лингвистика; массив текстовых данных; информационная безопасность; тексты коротких электронных сообщений; деструктивный контент; corpus linguistics; text data bank; information security; texts of short electronic messages; destructive content.