Разработка интеллектуального метода для классификации уязвимостей и угроз в лентах новостей

Р. И. Горохова, П.В. Никитин

***Аннотация*— В данном исследовании рассматривается метод классификации уязвимостей и киберугроз с использованием данных из информационно-новостных ресурсов. Авторы предлагают применять методы текстового анализа и машинного обучения для автоматизации выявления и классификации угроз в информационных системах. Исследуются различные подходы, включая латентно-семантический анализ, тематическое моделирование, анализ n-грамм и векторное представление текста, что позволяет выявлять семантические связи и тематические структуры. Эти методы помогают адаптировать подходы к требованиям к интерпретируемости и точности данных о киберугрозах. В работе рассмотрен подход к анализу совместной встречаемости слов и словосочетаний в текстах для выявления тематических кластеров киберугроз. Основным инструментом анализа является расчет меры ассоциации между терминами, что позволяет количественно оценить их взаимосвязи и способствует построению более эффективных моделей классификации с использованием методов машинного обучения. Для визуализации результатов предлагается использовать стратегические диаграммы, основанные на двух ключевых показателях: центральности и плотности кластеров. Кластеры классифицируются на четыре категории в зависимости от значений указанных показателей, что позволяет более точно охарактеризовать их сущность и взаимосвязи с другими угрозами. Для анализа динамики изменений в кластерах используется метод направленных графов, который позволяет отслеживать трансформации компонентов киберугроз от одного временного периода к другому. На основе представленного подхода возможно создание более информативных систем мониторинга киберугроз, что в конечном итоге способствует повышению уровня защиты информационных систем от потенциальных атак. Таким образом, исследование открывает новые горизонты для автоматизации анализа и классификации угроз, обеспечивая более эффективные решения в области кибербезопасности.**

***Ключевые слова*— классификация уязвимостей, киберугрозы, информационные системы, текстовый анализ, семантические связи, кластеризация, анализ текстовых данных.**

I. Введение

Для разработки метода классификации уязвимостей и угроз источником данных может служить информационно-новостной ресурс [1-3], например РБК.ру, так как новости разнообразны, пополняются в большом объеме ежедневно и можно выполнять парсинг данных [4, 5, 6].

В исследовании предлагается использовать методы анализа текстовых данных и машинного обучения для выявления и классификации уязвимостей и угроз в информационных системах [7, 8].

Для создания интеллектуального метода классификации уязвимостей [9] и киберугроз на основе машинного обучения [10], можно рассмотреть несколько подходов, связанных с анализом совместной встречаемости слов и словосочетаний в текстах [11, 12, 13].

Латентно-семантический анализ (Latent Semantic Analysis, LSA), основанный на матричной факторизации для выявления скрытых семантических связей между словами. Способен выявлять семантические кластеры, устойчивость к шуму, возможность обработки больших объемов данных. Обладает сложностью интерпретации латентных факторов, чувствительностью к неточностям в текстах [14, 15].

Тематическое моделирование (Topic Modeling) использует латентное распределение Дирихле (LDA), для выявления скрытых тематических структур в коллекциях текстов [16]. Применимо для автоматического определения тем, и обладает гибкостью в моделировании сложных тематических структур. Однако, можно выделить сложность подбора оптимальных параметров субъективность интерпретации тем.

Анализ *n*-грамм основан на изучении частоты встречаемости последовательностей из n слов (n-грамм) в текстах. Достаточно прост в реализации, выявлении устойчивых словосочетаний, возможности учета контекста. У данного метода высокая размерность признакового пространства и сложность в необходимости подбора оптимального *n* [17, 18].

Векторное представление текста (Word Embeddings) использует нейронные сети для построения плотных векторных представлений слов, отражающих их семантические и синтаксические свойства. Дает эффективное представление текста, возможность учета контекста, интеграция с другими методами машинного

Статья получена 17 марта 2025.
Р.В. Горохова – Финансовый университет при Правительстве Российской Федерации (email: rigorokhova@fa.ru)

П.В. Никитин – Финансовый университет при Правительстве Российской Федерации (email: pvnikitin@fa.ru)

обучения. Требует больших объемов данных для обучения, является сложным в интерпретации векторных представлений [19, 20].

Выбор конкретного подхода будет зависеть от особенностей данных об уязвимостях и угрозах, таких как объем и структурированность текстовых описаний, необходимость в интерпретируемости результатов, требования к точности и полноте обнаружения связей между терминами [21, 22].

Интеграция методов текстового анализа и интеллектуальных алгоритмов машинного обучения позволяет автоматизировать процесс выявления и классификации уязвимостей и угроз в описаниях кибератак. Это способствует более эффективной защите информационных систем от киберугроз [23].

II. Цель исследования

Цель данного исследования состоит в рассмотрении интеллектуального метода классификации уязвимостей и киберугроз на основе текстового анализа и машинного обучения. Исследование направлено на автоматизацию процесса выявления и классификации угроз на основании данных из информационно-новостных ресурсов, что будет способствовать более эффективной защите информационных систем от киберугроз.

III. Основная часть

Ключевым аспектом является анализ совместной встречаемости слов и словосочетаний в текстах [24]. Рассмотрим подход, который позволяет выявлять тематические кластеры и взаимосвязи между ключевыми элементами описания кибератаки.

Расчет меры ассоциации между терминами будет выполняться по формуле (1), что дает возможность количественно оценить силу связей между ключевыми элементами. Это, в свою очередь, может быть использовано для построения более эффективных моделей классификации на основе машинного обучения.

Изменения в кластерах на протяжении всего времени их существования происходят на уровне всех анализируемых смежных сегментов. Создание нейронной сети происходит через i и j. Мера ассоциации вычисляется по формуле 1.

$$s_{ij} = \frac{2ma_{ij}}{k_i k_j} \qquad (1),$$

где a_{ij} – численная оценка силы связи между терминами атак i и j, определяется на основе количества исследуемых атак, которые имеют сходство между собой по этим двум терминам, k_i и k_j– суммарный вес связей для i и j терминов, m – характеристика суммарного веса общих связей в нейронной сети [25].

Для анализа текстового описания кибератаки в целях выявления и классификации уязвимостей и угроз предлагается использовать стратегические диаграммы, основанные на текстовых данных [26].

Стратегические диаграммы строятся на основе двух показателей:

1. Степень важности и связанности кластера с другими темами - центральность (*sc*).

2. Внутренняя связность элементов внутри кластера - плотность (*p*).

В зависимости от значений этих показателей кластеры можно разделить на 4 категории:

- 1: кластер отражает основную киберугрозу, которая косвенно связана с второстепенными угрозами, значения показателей ($sc > 0, p > 0$). Например, кластер, связанный с фишинговыми атаками, может косвенно включать другие угрозы, такие как кража учетных данных.
- 2: кластер включает узкоспециализированную киберугрозу, значения показателей ($sc \leq 0, p > 0$). Например, кластер, связанный с конкретной уязвимостью в программном обеспечении.
- 3: кластер включает киберугрозы со слабо развитой направленностью, значения показателей ($sc \geq 0, p \leq 0$). Например, кластер, включающий общие рекомендации по кибербезопасности, но без привязки к конкретным угрозам.
- 4: кластер отражает киберугрозы, которые имеют потенциал для дальнейшего развития, показатели этого кластера представлены значениями ($sc > 0, p \leq 0$). Например, он может включать новые или недостаточно изученные виды атак, которые могут стать актуальными в будущем.

Начальная точка (нулевая координата) определяется на основании пересечения медианных значений $\{sc_i\}$ и $\{p_i\}$, где i - обозначает количество кластеров киберугроз.

Для анализа изменений в кластерах киберугроз применяются направленные графы. Эти графы иллюстрируют, как элементы работы кластеров изменяются от одного временного периода к другому.

Определяется количество кластеров киберугроз в новостной ленте РБК.ру как $sct = \left(c_i^t\right)$, где t - это временной период, а $i = 1, \ k_t$ — общее количество кластеров.

При переходе от периода t к t+1, каждая значение sc_i^t (sc - показатели кластера в момент времени t) преобразуется в новое множество $\left(sc_j^{t+1}\right)$, $j = 1, \ldots, K_{t+1}$, где K_{t+1} - количество кластеров в период $t+1$. Аналогично, переходя от t − 1 к t (в обратном порядке) каждый кластер sc_i^t изменяется из упорядоченного множества $\left(sc_j^{t-1}\right), j = 1, \ldots, K_{t-1}$, где K_{t-1} - количество кластеров в период $t-1$. В процессе перехода к t+1 или t-1 учитываются изменения лишь тех кластеров, которые соответствуют определенным критериям: scp – типы изменения кластеров, scf – типы формирования кластеров, Δ – разница в объемах общих составляющих между кластерами в соседних периодах:

$$\frac{scf:\langle sc_i^t \cap sc_1^{t+1}\rangle}{\langle sc_i^t\rangle} - \frac{\langle sc_i^t \cap sc_n^{t+1}\rangle}{\langle sc_i^t\rangle} = \Delta_{1,n}, \#(1)$$

$$\frac{scp:\langle sc_i^t \cap sc_1^{t-1}\rangle}{\langle sc_i^t\rangle} - \frac{\langle sc_i^t \cap sc_n^{t-1}\rangle}{\langle sc_i^t\rangle} = \Delta_{1,n}, \#(2)$$

где ∩ – пересечение множеств.

Выполнение условий отбора гарантирует, что кластеризация киберугроз будет проводиться правильно, и будет отражать актуальные изменения в структуре и составе кластеров при переходе между временными периодами [26].

Это позволяет отслеживать динамику развития, трансформации, появления и исчезновения различных тематических кластеров, связанных с киберугрозами, что является важным для понимания актуальных тенденций и прогнозирования будущих угроз.

Для отслеживания динамики изменения кластеров во времени строятся графы, где ребра показывают переходы между кластерами от одного периода к другому.

Каждый кластер помечается метками, отражающими его тематическое содержание по классифицируемой угрозе. Метки назначаются на основе наиболее часто встречающихся терминов, имеющих наибольшее количество связей внутри кластера.

Схема на рисунке 1 для *t=1* показывает формирование кластеров.

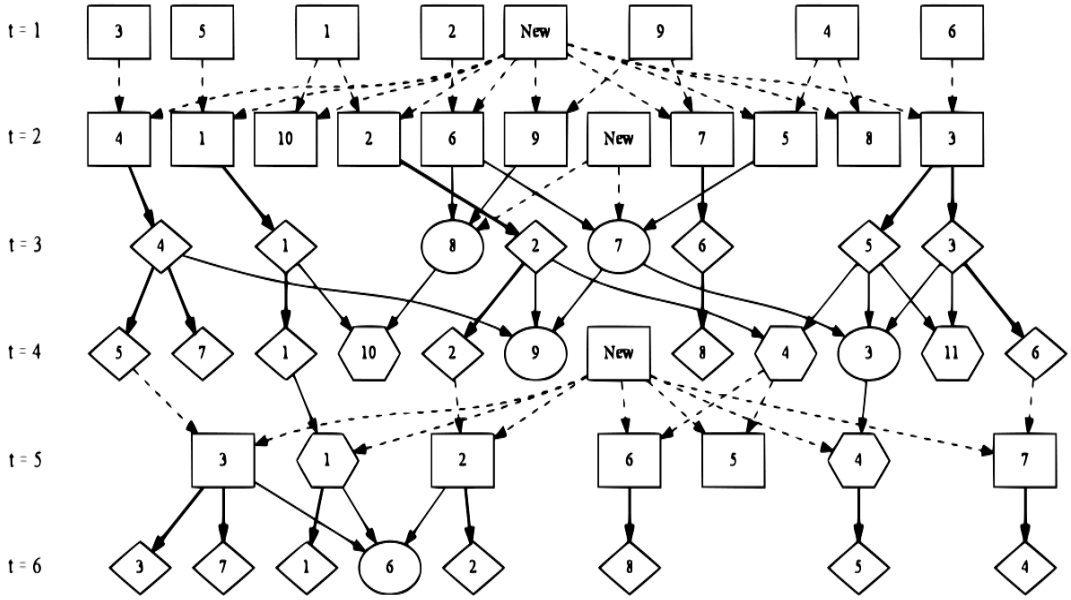

Рис. 1. Формирование кластеров [26]

Таким образом, использование стратегических диаграмм позволяет провести комплексный анализ текстового описания кибератаки, выявить ключевые темы, связанные с уязвимостями и угрозами, а также отследить их эволюцию во времени. Это способствует более эффективной классификации и пониманию природы выявленных киберугроз.

Анализ изменений в тематических кластерах, выявленных в текстовом описании кибератаки, позволяет классифицировать связанные с ними уязвимости и угрозы [27].

Предлагаемый подход основан на анализе изменений в тематических кластерах, выявленных в текстовом описании кибератаки.

IV. Заключение

В результате проведенного анализа совместной встречаемости слов и словосочетаний в текстах информационных ресурсов был разработан подход для выявления тематических кластеров киберугроз и их взаимосвязей. Используя разработанную формулу ассоциации, удалось количественно оценить связи между ключевыми терминами, что создает основы для более эффективного машинного обучения в классификации угроз.

Внедрение стратегических диаграмм, основанных на показателях центральности и плотности, позволяет структурировать кластеры киберугроз, обеспечивая их классификацию по типам и значимости. Это делает возможным более глубокое понимание динамики киберугроз и их развития во времени, что, в свою очередь, способствует повышению уровня защиты информационных систем.

Предложенный метод анализа и классификации предоставляет мощный инструмент для мониторинга и оценки киберугроз, а также помогает предсказывать потенциальные уязвимости. Реализация направленных графов для отслеживания изменений в кластерах обеспечивает актуальные данные для принятия оперативных решений в сфере кибербезопасности.

Библиография

[1] Мещеряков, И. В. Роль СМИ в государственной информационной политике по обеспечению национальной безопасности / И. В. Мещеряков // Наука Красноярья. – 2016. – Т. 5, № 6. – С. 70-82. – DOI 10.12731/2070-7568-2016-6-70-82. – EDN XHGMBZ.

[2] Головин, Ю. А. Возрастание роли СМИ в обеспечении информационной безопасности / Ю. А. Головин, А. Н. Орлов // Знание. Понимание. Умение. – 2013. – № 2. – С. 147-152. – EDN QYUIDN.

[3] Бурова, Ю. В. Потенциал региональных СМИ в противодействии терроризму (на примере Республики Мордовия) / Ю. В. Бурова // Век информации. – 2017. – № 2-1. – С. 180-181. – EDN YGRTCL.
[4] Плоткина, П. В. Международные проблемы противодействия фейковым новостям как инструменту информационных войн / П. В. Плоткина, Э. К. Тохчукова // Теории и проблемы политических исследований. – 2023. – Т. 12, № 5А-6А. – С. 54-62. – DOI 10.34670/AR.2023.94.56.012. – EDN IOCCEZ.
[5] Обидов, О. С. Современное состояние и перспективы развития СМИ в обеспечении информационной безопасности в современных международных отношениях / О. С. Обидов // Известия Академии наук Республики Таджикистан. Отделение общественных наук. – 2019. – № 1(254). – С. 57-60. – EDN ELFOXT.
[6] Запорожцева, В. М. Парсинг сайтов как метод сбора данных для лингвистических исследований / В. М. Запорожцева // Молодой ученый. – 2024. – № 24(523). – С. 496-499. – EDN QCMGMO.
[7] Методы оценки защищенности компьютерных систем информационной поддержки цифровой экономики / А. А. Грушо, Н. А. Грушо, М. И. Забежайло, Е. Е. Тимонина // International Journal of Open Information Technologies. – 2019. – Т. 7, № 4. – С. 61-66. – EDN ZCGPAD.
[8] Eyzenakh, D. S. High performance distributed web-scraper / D. S. Eyzenakh, A. S. Rameykov, I. V. Nikiforov // Proceedings of the Institute for System Programming of the RAS. – 2021. – Vol. 33, No. 3. – P. 87-100. – DOI 10.15514/ISPRAS-2021-33(3)-7. – EDN SIPWXY.
[9] Bogatenko, T. R. Application of machine learning and statistics to anaesthesia detection from EEG data / T. R. Bogatenko, K. S. Sergeev, G. I. Strelkova // Izvestiya of Saratov University. New Series. Series: Physics. – 2024. – Vol. 24, No. 3. – P. 209-215. – DOI 10.18500/1817-3020-2024-24-3-209-215. – EDN HKYBMM.
[10] Analysis of Machine Learning Models by Solving the Text Data Classification Problem / A. V. Pchelin, N. A. Kononov, V. S. Serova [et al.] // Journal of Computational and Engineering Mathematics. – 2021. – Vol. 8, No. 2. – P. 33-45. – DOI 10.14529/jcem210203. – EDN VREXYT.
[11] Даринская Л. А. Библиометрический анализ как способ вхождения в проблему исследования / Л. А. Даринская, А. С. Гуслина // Вестник Санкт-Петербургского университета. – 2010. – № 3. – С. 71-79.
[12] Классификация текстов различными методами машинного обучения для выявления признаков киберагрессии / Ю. Е. Гапанюк, Д. А. Попова, К. Р. Рабцевич [и др.] // Естественные и технические науки. – 2022. – № 6(169). – С. 277-281. – EDN OHHHNM.
[13] Хотин, Д. Ю. Выявление фейковых фрагментов в текстовых новостных сообщениях с помощью машинного обучения / Д. Ю. Хотин // Вестник науки. – 2024. – Т. 1, № 6(75). – С. 1568-1576. – EDN MHTVUE.
[14] Гусев, П. Ю. Разработка системы классификации текстов по научным специальностям с применением методов машинного обучения / П. Ю. Гусев // Вестник Новосибирского государственного университета. Серия: Информационные технологии. – 2021. – Т. 19, № 1. – С. 39-47. – DOI 10.25205/1818-7900-2021-19-1-39-47. – EDN DWLTXW.
[15] Дронов, С. В. Оптимизация кластерных разбиений с привлечением техники латентного анализа классов / С. В. Дронов // Известия Алтайского государственного университета. – 2023. – № 1(129). – С. 89-94. – DOI 10.14258/izvasu(2023)1-14. – EDN ZIMQKI.
[16] Алмаев, Н. А. Тематический анализ дискуссий с применением метода Латентного размещения Дирихле / Н. А. Алмаев, О. В. Мурашева // Институт психологии Российской академии наук. Социальная и экономическая психология. – 2022. – Т. 7, № 1(25). – С. 47-69. – DOI 10.38098/ipran.sep_2022_25_1_03. – EDN YECKGR.
[17] Зырянов, М. С. Влияние длины n-грамм при пословной токенизации на эффективность выявления текстов экстремистской направленности методами машинного обучения / М. С. Зырянов // Научный аспект. – 2024. – Т. 22, № 5. – С. 2949-2961. – EDN KVHJHN.
[18] Kechik, D. A. Method of estimation of frequency variation relying on estimation of shift of spectral peaks / D. A. Kechik, Yu. P. Aslamov, I. G. Davydov // Системный анализ и прикладная информатика. – 2021. – No. 1. – P. 53-61. – DOI 10.21122/2309-4923-2021-1-53-61. – EDN CDEBVV.
[19] Grefenstette, G. Competing Views of Word Meaning: Word Embeddings and Word Senses / G. Grefenstette, P. Hanks // International Journal of Lexicography. – 2023. – Vol. 36, No. 2. – P. 211-219. – DOI 10.1093/ijl/ecad005. – EDN NQHRQL.
[20] Панамарева, О. Н. Реализация кластеризации новостных потоков на основе векторных представлений текста / О. Н. Панамарева, В. В. Лука, Д. А. Сухарев // Известия Тульского государственного университета. Технические науки. – 2024. – № 7. – С. 304-309. – DOI 10.24412/2071-6168-2024-7-304-305. – EDN VPTOZH.
[21] Sandhu, T. Exploration of Word Embeddings with Graph-Based Context Adaptation for Enhanced Word Vectors / T. Sandhu, Z. Kobti // The International FLAIRS Conference Proceedings. – 2024. – Vol. 37. – DOI 10.32473/flairs.37.1.135597. – EDN PWIKWF.
[22] Кандилас, В., Апхэм, С. П., Унгар, Л. Х. Анализ сообществ знаний с использованием кластеров переднего плана и фона. [Электронный ресурс]. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.146.3141&rep= rep1&type=pdf (дата обращения: 02.10.2024).
[23] Нгуен, Т.Т. и др. Многоцелевая система обучения с глубоким подкреплением // Инженерные приложения искусственного интеллекта, Том 96, ноябрь 2020, 103915. [Электронный ресурс]. - https://doi.org/10.1016/j.engappai.2020.103915(дата обращения: 02.10.2024).
[24] Пальмов, С.В., Артюшкина, Е.С. Глубокое обучение: определение и отличительные особенности. //Форум молодых ученых. 2020. № 3 (43). С. 311-316.
[25] Чистова, Е.В., Шелманов, А.О., Смирнов И.В. Применение глубокого обучения к моделированию диалога на естественном языке. // Труды Института системного анализа Российской академии наук. 2019. Т. 69. № 1. С. 105-115.
[26] Потемкин, А.В. Обработка разнородной информации с помощью глубокого обучения нейронных сетей. // Мягкие измерения и вычисления. 2019. № 9 (22). С. 44-48.
[27] Малый, Н. Отслеживание и прогнозирование областей роста в науке [Электронный ресурс]. http://www.scimaps.org/exhibit/docs/small.pdf (дата обращения: 01.10.2024).

Development of an Intelligent Method Based on Machine Learning for Classification of Vulnerabilities and Threats in Information Systems

Rimma Gorokhova, Pyotr Nikitin

***Abstract* – This study considers a method for classifying vulnerabilities and cyber threats using data from information and news resources. The authors propose to apply text analysis and machine learning methods to automate the detection and classification of threats in information systems. Various approaches are explored, including latent semantic analysis, topic modeling, n-gram analysis, and vector representation of text, which allows identifying semantic relationships and thematic structures. These methods help to adapt approaches to the requirements for interpretability and accuracy of cyber threat data. The paper considers an approach to analyzing the co-occurrence of words and phrases in texts to identify thematic clusters of cyber threats. The main analysis tool is the calculation of the association measure between terms, which allows for a quantitative assessment of their relationships and contributes to the construction of more effective classification models using machine learning methods. To visualize the results, it is proposed to use strategic diagrams based on two key indicators: centrality and cluster density. Clusters are classified into four categories depending on the values of these indicators, which allows for a more accurate characterization of their essence and relationships with other threats. To analyze the dynamics of changes in clusters, the method of directed graphs is used, which allows tracking the transformation of cyber threat components from one time period to another. Based on the presented approach, it is possible to create more informative cyber threat monitoring systems, which ultimately contributes to increasing the level of protection of information systems from potential attacks. Thus, the study opens up new horizons for automating the analysis and classification of threats, providing more effective solutions in the field of cybersecurity.**

***Keywords* – vulnerability classification, cyber threats, information systems, text analysis, semantic links, clustering, text data analysis.**

REFERENCES

[1] Meshcheryakov, I. V. The role of the media in the state information policy to ensure national security / I. V. Meshcheryakov // Science of Krasnoyarsk. - 2016. - Vol. 5, No. 6. - P. 70-82. - DOI 10.12731 / 2070-7568-2016-6-70-82. - EDN XHGMBZ.

[2] Golovin, Yu. A. Increasing role of the media in ensuring information security / Yu. A. Golovin, A. N. Orlov // Knowledge. Understanding. Skill. - 2013. - No. 2. - P. 147-152. - EDN QYUIDN.

[3] Burova, Yu. V. Potential of regional media in countering terrorism (on the example of the Republic of Mordovia) / Yu. V. Burova // Age of information. – 2017. – No. 2-1. – P. 180-181. – EDN YGRTCL.

[4] Plotkina, P. V. International problems of counteracting fake news as a tool of information wars / P. V. Plotkina, E. K. Tokhchukova // Theories and problems of political research. – 2023. – Vol. 12, No. 5A-6A. – P. 54-62. – DOI 10.34670/AR.2023.94.56.012. – EDN IOCCEZ.

[5] Obidov, O. S. Current state and prospects for the development of media in ensuring information security in modern international relations / O. S. Obidov // News of the Academy of Sciences of the Republic of Tajikistan. Department of Social Sciences. – 2019. – No. 1 (254). – P. 57-60. – EDN ELFOXT.

[6] Zaporozhtseva, V. M. Parsing websites as a method of collecting data for linguistic research / V. M. Zaporozhtseva // Young scientist. – 2024. – No. 24(523). – P. 496-499. – EDN QCMGMO.

[7] Methods for assessing the security of computer systems for information support of the digital economy / A. A. Grusho, N. A. Grusho, M. I. Zabezhailo, E. E. Timonina // International Journal of Open Information Technologies. – 2019. – Vol. 7, No. 4. – P. 61-66. – EDN ZCGPAD.

[8] Eyzenakh, D. S. High performance distributed web-scraper / D. S. Eyzenakh, A. S. Rameykov, I. V. Nikiforov // Proceedings of the Institute for System Programming of the RAS. – 2021. – Vol. 33, No. 3. – P. 87-100. – DOI 10.15514/ISPRAS-2021-33(3)-7. – EDN SIPWXY.

[9] Bogatenko, T. R. Application of machine learning and statistics to anesthesia detection from EEG data / T. R. Bogatenko, K. S. Sergeev, G. I. Strelkova // Izvestiya of Saratov University. New Series. Series: Physics. – 2024. – Vol. 24, No. 3. – P. 209-215. – DOI 10.18500/1817-3020-2024-24-3-209-215. – EDN HKYBMM.

[10] Analysis of Machine Learning Models by Solving the Text Data Classification Problem / A. V. Pchelin, N. A. Kononov, V. S. Serova [et al.] // Journal of Computational and Engineering Mathematics. – 2021. – Vol. 8, No. 2. – P. 33-45. – DOI 10.14529/jcem210203. – EDN VREXYT.

[11] Darinskaya L. A. Bibliometric analysis as a way of entering the research problem / L. A. Darinskaya, A. S. Guslina // Bulletin of St. Petersburg University. – 2010. – No. 3. – P. 71-79.

[12] Text classification using various machine learning methods to detect signs of cyber aggression / Yu. E. Gapanyuk, D. A. Popova, K. R. Rabtsevich [et al.] // Natural and technical sciences. - 2022. - No. 6 (169). - P. 277-281. - EDN OHHHNM.

[13] Khotin, D. Yu. Identification of fake fragments in text news messages using machine learning / D. Yu. Khotin // Bulletin of science. - 2024. - Vol. 1, No. 6 (75). - P. 1568-1576. - EDN MHTVUE.

[14] Gusev, P. Yu. Development of a text classification system by scientific specialties using machine learning methods / P. Yu. Gusev // Bulletin of the Novosibirsk State University. Series: Information Technology. – 2021. – V. 19, No. 1. – P. 39-47. – DOI 10.25205/1818-7900-2021-19-1-39-47. – EDN DWLTXW.

[15] Dronov, S. V. Optimization of cluster partitions using latent class analysis technique / S. V. Dronov // Bulletin of Altai State University. – 2023. – No. 1(129). – P. 89-94. – DOI 10.14258/izvasu(2023)1-14. – EDN ZIMQKI.

[16] lmaev, N. A. Thematic analysis of discussions using the Latent Dirichlet Allocation method / N. A. Almaev, O. V. Murasheva // Institute of Psychology of the Russian Academy of Sciences. Social and Economic Psychology. - 2022. - Vol. 7, No. 1 (25). - P. 47-69. - DOI 10.38098 / ipran.sep_2022_25_1_03. - EDN YECKGR.

[17] Zyryanov, M. S. The influence of the length of n-grams in word-by-word tokenization on the efficiency of identifying extremist texts using machine learning methods / M. S. Zyryanov // Scientific aspect. - 2024. - Vol. 22, No. 5. - P. 2949-2961. - EDN KVHJHN.

[18] Kechik, D. A. Method of estimation of frequency variation relying on estimation of shift of spectral peaks / D. A. Kechik, Yu. P. Aslamov, I. G. Davydov // System analysis and applied informatics. – 2021. – No. 1. – P. 53-61. – DOI 10.21122/2309-4923-2021-1-53-61. – EDN CDEBVV.

[19] Grefenstette, G. Competing Views of Word Meaning: Word Embeddings and Word Senses / G. Grefenstette, P. Hanks // International Journal of Lexicography. – 2023. – Vol. 36, No. 2. – P. 211-219. – DOI 10.1093/ijl/ecad005. – EDN NQHRQL.

[20] Panamareva, O. N. Implementation of clustering of news flows based on vector representations of text / O. N. Panamareva, V. V. Luka, D.

A. Sukharev // Bulletin of Tula State University. Technical sciences. – 2024. – No. 7. – P. 304-309. – DOI 10.24412/2071-6168-2024-7-304-305. – EDN VPTOZH.

[21] Sandhu, T. Exploration of Word Embeddings with Graph-Based Context Adaptation for Enhanced Word Vectors / T. Sandhu, Z. Kobti // The International FLAIRS Conference Proceedings. – 2024. – Vol. 37. – DOI 10.32473/flairs.37.1.135597. – EDN PWIKWF.

[22] Kandilas, V., Upham, S. P., Ungar, L. H. Knowledge Community Analysis Using Foreground and Background Clusters. [Electronic resource]. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.146.3141&rep= rep1&type=pdf (accessed: 02.10.2024).

[23] Nguyen, T.T. et al. Multi-target deep reinforcement learning system // Engineering Applications of Artificial Intelligence, Vol. 96, November 2020, 103915. [Electronic resource]. - https://doi.org/10.1016/j.engappai.2020.103915 (accessed: 02.10.2024).

[24] Palmov, S.V., Artyushkina, E.S. Deep learning: definition and distinctive features. // Forum of young scientists. 2020. No. 3 (43). P. 311-316.

[25] Chistova, E.V., Shelmanov, A.O., Smirnov, I.V. Application of deep learning to modeling dialogue in natural language. // Proceedings of the Institute for Systems Analysis of the Russian Academy of Sciences. 2019. Vol. 69. No. 1. P. 105-115.

[26] Potemkin, A.V. Processing heterogeneous information using deep learning of neural networks. // Soft measurements and calculations. 2019. No. 9 (22). P. 44-48.

[27] Small, H. Tracking and forecasting growth areas in science [Electronic resource]. http://www.scimaps.org/exhibit/docs/small.pdf (accessed: 01.10.2024).