УДК 004.032.26:82-4

# ИССЛЕДОВАНИЕ СОДЕРЖАНИЯ МЕДИЦИНСКИХ ЭССЕ: НЕЙРОСЕТЕВЫЕ ПОДХОДЫ И ТРАДИЦИОННЫЕ МЕТОДЫ

**[1]Меликбекян А.А., [2]Борбат А.М., [3]Новикова Т.О., [4]Павлов К.А.**

*[1]ФГБОУ «МИРЭА – Российский технологический университет», Москва, e-mail: melikbekyan.ashot@yandex.ru*

*[2]ООО «Деметрамед», Москва, e-mail: aborbat@yandex.ru*

*[3]ФГБУ «НМИЦ онкологии им. Н. Н. Блохина» Минздрава России, Москва, e-mail: tn.path1910@yandex.ru*

*[4]ФГБУ ГНЦ ФМБЦ им. А.И. Бурназяна ФМБА России, Москва, e-mail: pavlovkos@gmail.com*

Цель исследования – привлечь внимание к современным методам машинного анализа текстов среди российских исследователей в медицинской сфере, а также продемонстрировать их применение и эффективность на примере анализа эссе клинических ординаторов. Анализ производился на основе 297 эссе клинических ординаторов ФГБУ ГНЦ ФМБЦ им. А.И. Бурназяна на тему «Почему я не патологоанатом?». Использовались методы нейросетевого анализа для численного представления текстов и классические методы, включая анализ частотности n-грамм и создание облаков слов. С помощью нейросетевых методов были успешно выявлены тексты с различной степенью схожести, в том числе и потенциальные случаи плагиата. Однако применение этих методов для кластеризации оказалось неэффективным из-за узконаправленной тематики текстов и однородности авторов. Традиционные методы анализа позволили глубже проникнуть в смысловое содержание текстов, выявляя основные темы и акценты в рамках исследования. Несмотря на активное развитие и высокую эффективность нейросетевых методов, традиционные подходы к анализу текстовых данных продолжают оставаться актуальными и весьма информативными. Комбинирование этих подходов может дать наиболее полное и многогранное представление о содержании анализируемых текстов, что подтверждается результатами настоящего исследования.

**Ключевые слова: анализ текстовых данных, патологическая анатомия, нейронные сети, машинное обучение, педагогика**

# RESEARCH ON THE CONTENT OF MEDICAL ESSAYS: NEURAL NETWORKS APPROACHES AND TRADITIONAL METHODS

**[1]Melikbekyan A.A., [2]Borbat A.M., [3]Novikova T.O., [4]Pavlov K.A.**

*[1]MIREA – Russian Technological University, Moscow, e-mail: melikbekyan.ashot@yandex.ru*

*[2]Limited liability company «Demetramed», Moscow, e-mail: aborbat@yandex.ru*

*[3]National Medical Research Center of Oncology named after N.N. Blokhin of the Ministry of Health of the Russian Federation, Moscow, e-mail: tn.path1910@yandex.ru*

*[4]Russian State Research Center − Burnasyan Federal Medical Biophysical Center of Federal Medical Biological Agency, Moscow, e-mail: pavlovkos@gmail.com*

The purpose of this study is to highlight the importance of modern methods of machine text analysis for Russian researchers in the medical field. We will also demonstrate the application and effectiveness of these methods by analyzing essays of clinical residents. Materials and methods. The analysis was performed on 297 essays by clinical residents of Burnasyan FMBC of FMBA on the topic «Why am I not a pathologist?». We used neural networks to get text embeddings and traditional algorithms like n-grams and word clouds. Results. Neural networks successfully computed similarity between essays and helped to identify potentially plagiarized ones. However, the usage of these methods for clustering has proven to be ineffective due to the very specific topic of the question. Traditional methods allowed us to deeply understand the semantic content of texts, identifying main themes and emphases within the context of study. Conclusion. Despite the active development and high efficiency of neural network methods, traditional approaches to text data analysis still play a significant role and provide valuable insights. Combining these methods can lead to a more comprehensive and multifaceted understanding of the content in analyzed texts, as demonstrated by the findings of this study.

**Keywords: text data analysis, pathological anatomy, neural networks, machine learning, pedagogy**

## Введение

В настоящее время многие исследователи применяют компьютерный анализ текстов для решения различных задач при работе с неструктурированными текстами, в том числе в медицине [1-3]. Исследуются возможности методов обработки естественного языка для решения широкого спектра задач: скрининга заболеваний [4], определения диагноза [5, 6], улучшения наблюдения за пациентами [7], изучения жалоб пациентов, исправления ошибок в тексте [8] и т.д. Разработаны модели для работы с русскоязычными медицинскими текстами [9].

Несмотря на значительный интерес к алгоритмам извлечения информации из текста на нейросетей [4, 8], сохраняют свою актуальность и более традиционные методы, требующие предварительной обработки данных с последующим анализом n-грамм, применением логистической регрессии и др. [5]. При этом большинство авторов комбинируют традиционные и нейросетевые алгоритмы [10–13].

Несмотря на возросшее количество публикаций, посвященных этой теме в последние годы, методы машинного анализа текста остаются относительно недооцененным в профессиональной медицинской среде [14, 15]. При этом самой текстовой информации становится больше, а возможности ее получения за счет распространения социальных сетей и цифровизации здравоохранения становятся менее ресурсозатратными [15].

Исходя из этого, освоение методов машинного анализа текста представляет собой перспективное направление, особенно в условиях растущего объема доступной текстовой информации.

**Цель исследования**: привлечь внимание российских исследователей к этим инструментам. Для этого авторы хотели бы представить вниманию коллег собственный опыт и таким способом продемонстрировать относительную простоту этих методов и возможности их применения.

## Материалы и методы исследования

В рамках данного исследования проанализированы эссе клинических ординаторов первого года ФГБУ ГНЦ ФМБЦ им. А.И. Бурназяна, проходивших обучение на цикле по патологии в 2021–2022 гг., на тему «Почему я не патологоанатом?». Всего было получено и проанализировано 297 эссе, среднее значение количества слов в эссе составило 422, минимальное и максимальное значение соответственно 57 и 1063 слова, медиана – 424 слова.

В настоящей работе представлен комплексный анализ текстовых данных, который состоит из двух основных частей. В первой части реализована численная характеризация текстов с использованием методов нейросетевого анализа без предварительной обработки текстов. Вторая часть исследования фокусируется на лингвистическом анализе текстов, который включает в себя выявление n-грамм, создание облаков слов и последующий семантический анализ этих данных с целью формирования представления о содержательных аспектах исследуемых текстов. Описание методологических подходов, а также примеры и выводы, сформулированные на основе проведенного анализа, изложены ниже.

## Результаты исследования и их обсуждение

Авторы провели сравнение текстов на предмет антиплагиата с использованием современных методов анализа текста на основе модели глубокого обучения word2vec [16], позволяющей переводить строковые данные в числовые представления. Такое преобразование позволяет применить арифметические операции к текстовым данным [17]. Для получения численных представлений использовалась предобученная модель семейства MPNet [18], поскольку в сравнении с аналогами она демонстрирует высокую эффективность при работе с текстами. Численные векторы, полученные через использование алгоритма нейросетевого глубокого обучения, служат отправной точкой для ряда последующих аналитических методов, таких как кластеризация и анализ на антиплагиат. Эти векторы, сгенерированные с высокой точностью и скоростью, могут быть легко интегрированы в различные системы анализа данных, обеспечивая тем самым повышенную точность и снижение временных затрат на обработку больших текстовых массивов.

### *Проверка эссе на наличие плагиата*

Для выявления плагиата применялась общеизвестная (стандартная) методика, описанная во множестве работ, опубликованных в последние годы [19, 20]. Метод подразумевает сравнение векторов друг с другом по косинусному расстоянию. Диапазон метрики варьируется от 0 до 1, где 0 – это не имеющие ничего общего данные, а 1 – это одинаковые данные. Пороговое значение, по которому выделяются похожие тексты, в рамках данной работы было определено эмпирически и составило 0,95. На рисунке 1 приведена тепловая карта, построенная на основе данных о схожести численных представлений эссе, где, в частности, видно, что у текстов под номерами 150 и 194 значение схожести близко к 1. На рисунке 2 приводятся фрагменты текстов этих эссе.

### *Кластеризация*

Численно представленные векторами тексты могут быть проанализированы стандартными статистическими методами. В данной работе был использован метод кластеризации с целью создания групп сходных объектов (кластеров). Для кластеризации использовались алгоритмы агломеративной кластеризации и KMeans [21].

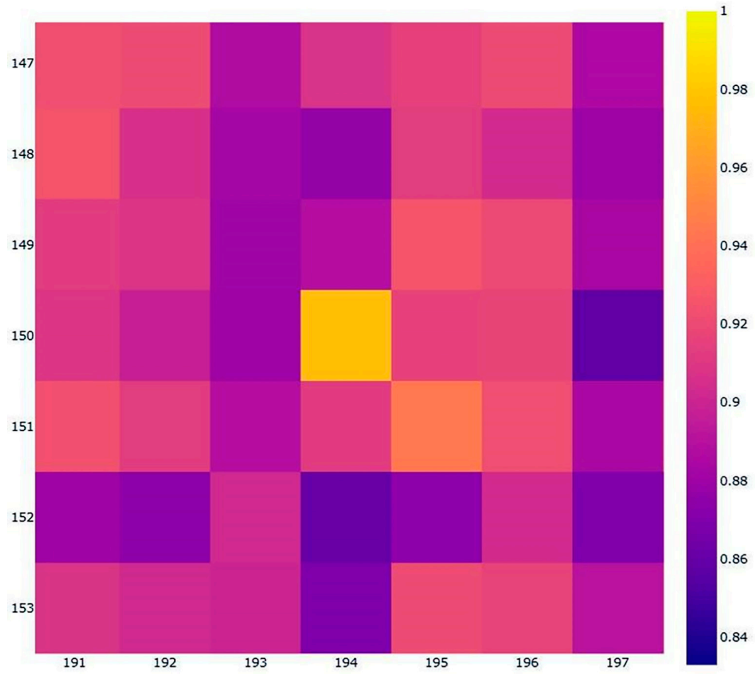


*Рис. 1. Фрагмент тепловой карты с высоким значением косинусного расстояния на пересечении наблюдений 194 и 150*

| Эссе студента №150 | Эссе студента №194 |
|---|---|
| Что касается меня, я всегда хотела помогать живым людям, хотела вносить свой вклад в их здоровье. Я люблю живое общение, похвалу, благодарность, как и все эгоцентричные натуры. Также отвращает запах при вскрытиях: особенно содержимого внутренних органов. Еще будучи студенткой, во время практики, когда я посещала вскрытия, мне становилось не по себе. Один раз я даже упала в обморок. Тогда я для себя четко решила, что патологическая анатомия не для меня. | Что касается меня, я эмоционально нестабильна. Люблю живое общение, похвалу, благодарность, как и все эгоцентричные натуры. Также отвращает запах при вскрытиях: особенно содержимого желудка. Немаловажна оплата и карьерный рост . По этим причинам профессия мне не подходит. |

*Рис. 2. Фрагменты текстов эссе № 150 и № 194*

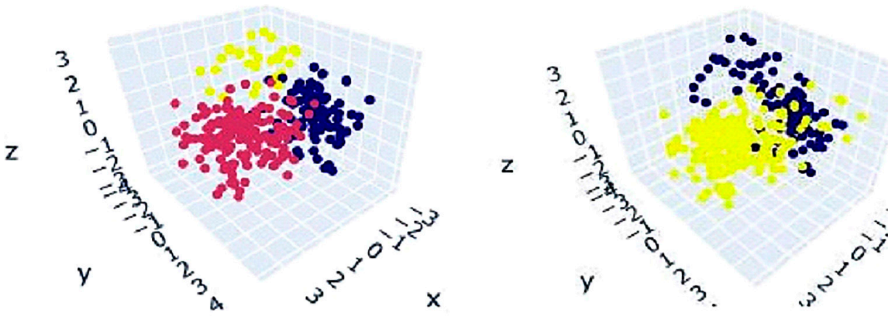


*Рис. 3. Кластеризация отдельных эссе. Слева представлена кластеризация по алгоритму KMeans, справа – по агломеративному алгоритму*

Кластеризация осуществлялась в двух вариантах: на основе отдельных эссе студентов и на основе текстов, объединенных по специальностям. В полученных по результатам анализа кластерах не удалось выявить сколько-нибудь значимую связь со специальностями авторов, как для отдельных эссе, так и при анализе их групп. На рисунке 3 представлены результаты кластеризации, спроецированные на трехмерное пространство с помощью метода главных компонент.

*Анализ текст по n-граммам*
*Предварительная обработка корпуса данных*

Предобработка текста включала в себя следующие элементы: лемматизация, удаление стоп-слов и небуквенных символов; операция лемматизации с помощью библиотеки pymystem3; удаление стоп-слов с помощью библиотеки nltk, расширенной с учетом специфики данных следующими терминами: «специальность», «профессия», «патологоанатом», «врач», «человек», «работа», «медицинский», «университет», «патологический», «анатомия». Удаление небуквенных символов произведено с помощью стандартной библиотеки для регулярных выражений в Python. По результатам предварительной обработки получены «очищенные» тексты, избавленные от нерелевантных стоп-слов, вспомогательных слов и небуквенных символов. Эта оптимизированная форма текста облегчает проведение последующих этапов анализа и улучшает качество семантического анализа.

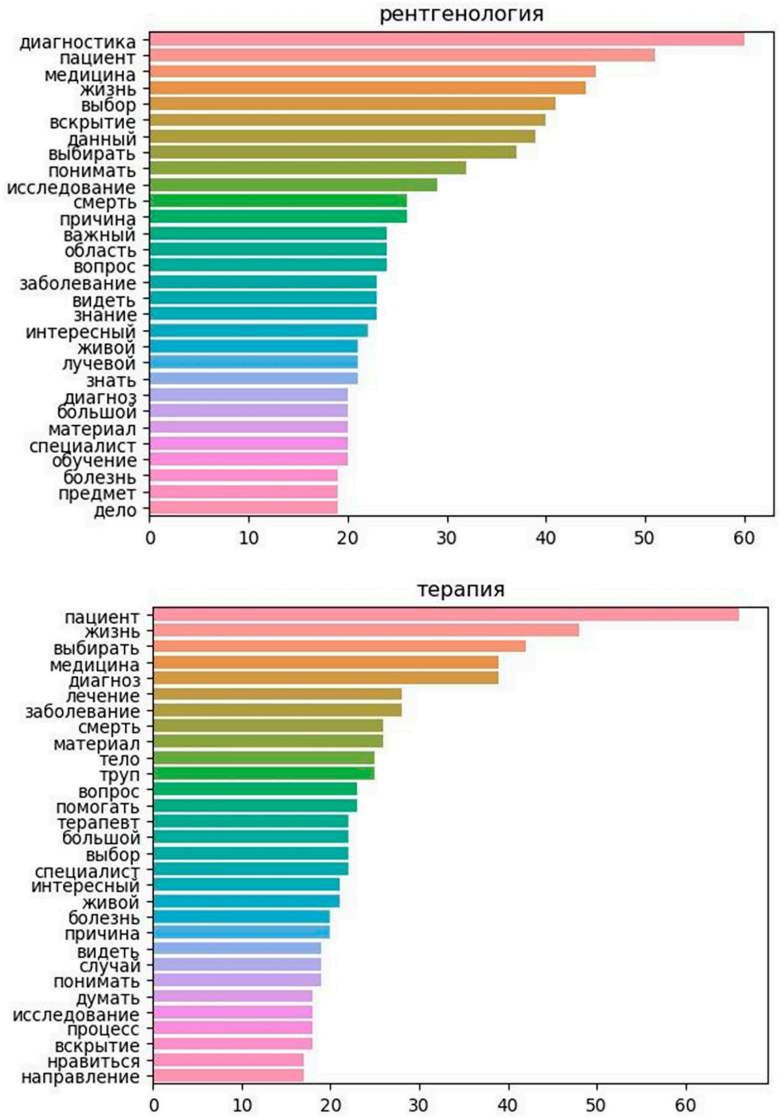


*Рис. 4. Столбчатые диаграммы частот юнграмм в наиболее представительных группах диагностических («Рентгенология», 33 наблюдения) и клинических («Терапия», 27 наблюдений) специальностей*

*Анализ частотности n-грамм*

Анализ частотности n-грамм позволяет изучить распределение и взаимосвязь последовательностей из n слов в тексте, что поможет выявить ключевые понятия и дать представление об общей концепции анализируемого текста [22]. Были использованы три варианта n-грамм: юниграммы, биграммы, триграммы, содержащие соответственно одно, два и три слова. Среди наиболее частых юниграмм выявлены следующие слова: «пациент», «жизнь», «вскрытие», «смерть», «диагноз», «труп», «лечение».

Это позволило сформировать общее представление о содержании текста. Кроме этого, были выявлены различия в частотах слов по группам специальностей. Слова «жизнь», «помощь» и «общение» занимали более высокий ранг в терапевтических и хирургических специальностях, в эссе диагностических специальностей термин «аутопсия» занимал более высокий ранг. Наиболее частые юниграммы для специальностей «Рентгенология» и «Терапия» приведены на рисунке 4.

Анализ биграмм позволил уточнить предположения, полученные при анализе юниграмм. В частности, юниграмма «смерть» с высоким рангом во всех группах при анализе биграмм оказалась составным элементом «причина смерть», т.е. с большей вероятностью связана с темой эссе, чем с личными причинами невыбора специальности. Юниграмма «пациент» распределилась по нескольким биграммам: с высокими частотами «общение пациент» и «жизнь пациент» и более редкие «смерть пациент» и «живой пациент». При этом частоты самых популярных биграмм во всем корпусе данных исчислялись десятками, что группировку эссе по каким-либо признакам и последующий их анализ. Тридцать наиболее частых биграмм во всем корпусе данных приведены на рисунке 5.

Анализ триграмм в значительной степени дублировал наблюдения, обнаруженные при анализе биграмм, при этом акцент в большей степени сместился к психологическим аспектам работы и к необходимости взаимодействовать с биологическим материалом как источником профессионального риска. При этом наиболее распространенные триграммы по своим частотам находились в диапазоне одного-двух десятков, что в еще большей степени ограничивало возможность их применения для межгруппового анализа. Увеличение количества эссе, как и в случае с биграммами, позволило бы обойти это ограничение, однако первичные данные в таком случае должны были бы исчисляться тысячами наблюдений. Наиболее частые триграммы представлены на рисунке 6.

*Облако слов*

Из-за ограниченного объема данных инструмент «облако слов» применяли только к юниграммам. Облако слов – это визуальное представление списка категорий, при котором частота каждой категории выражается размером шрифта или цветом [23].

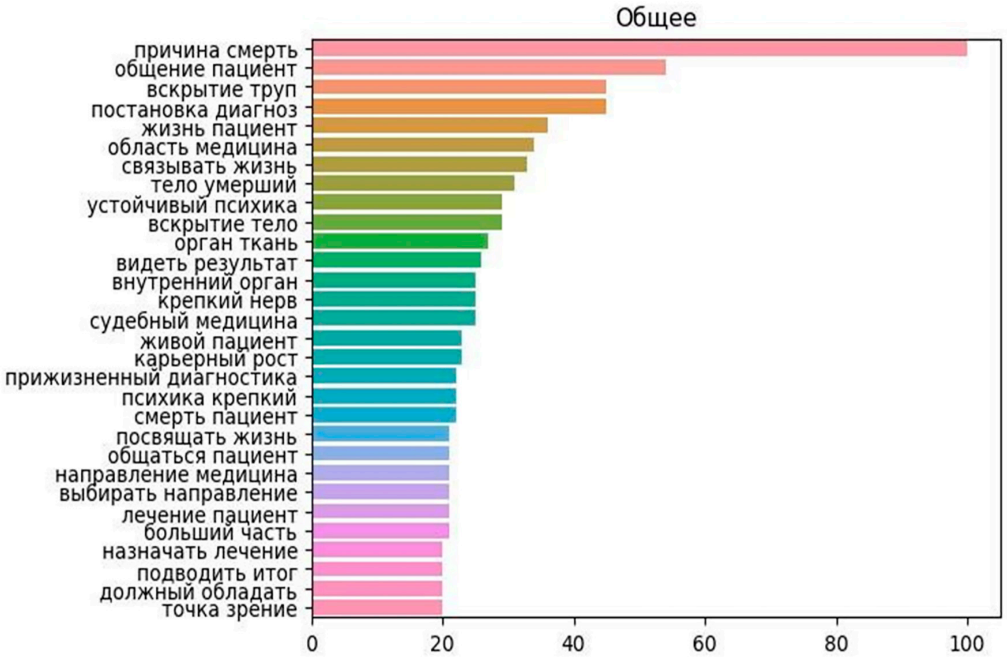


*Рис. 5. Столбчатая диаграмма частот биграмм в корпусе данных*

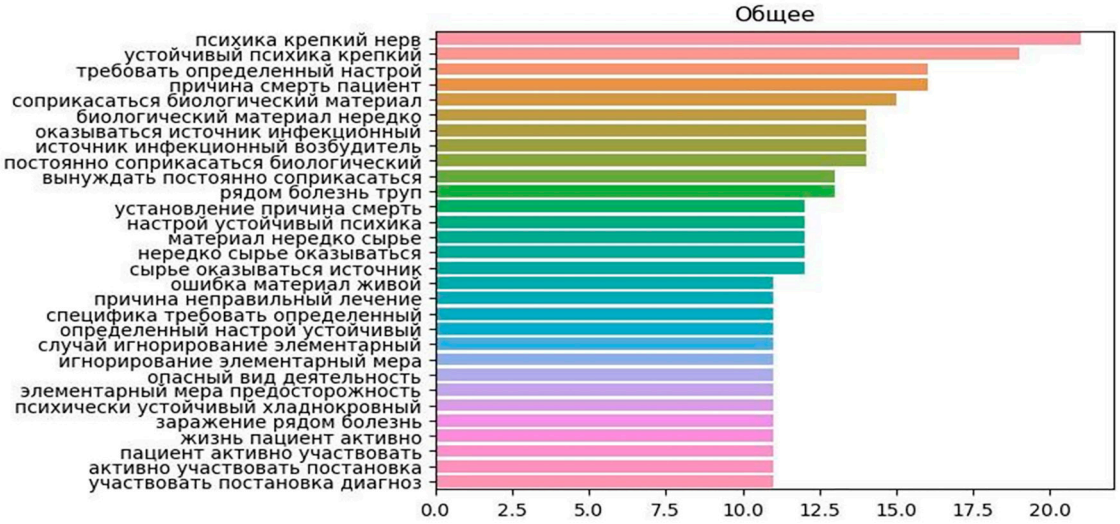


*Рис. 6. Столбчатая диаграмма частот триграмм в корпусе данных*


*Рис. 7. Облако слов, составленное для всего корпуса данных*

Такое представление удобно для быстрого восприятия и для распределения терминов по популярности друг относительно друга, что, по мнению авторов, делает его удобным инструментом для ориентировочного анализа. Для построения облака слов использовалась библиотека wordcloud языка программирования Python.

*Обсуждение*

Проведенный в настоящей работе анализ неструктурированных текстов различными автоматизированными методами позволил сравнить их эффективность для раскрытия семантического содержания эссе.

Нейросетевые алгоритмы были успешно применены для численного представления и поиска схожих по содержанию текстов – анализа на антиплагиат. Были выявлены тексты различной степени схожести, вплоть до полной идентичности. Этот подход позволил во всем массиве текстов быстро и с минимальными затратами ресурсов найти работы, которые с высокой вероятностью были написаны одним автором.

Несмотря на широкое применение метода кластеризации в анализе неструктурированных текстов [24], в рамках данного исследования с его помощью не удалось обнаружить значимых связей. Скорее всего, такой результат связан с тем, что анализируемые тексты принадлежат одной тематической области и были созданы сравнительно однородной группой авторов.

Помимо современных нейросетевых методов, классические подходы сохраняют свою актуальность. Анализ частотности n-грамм является широко используемым в аналитической практике. По результатам данного исследования наибольшую информативность продемонстрировали биграммы. В сравнении с юниграммами и триграммами биграммы обеспечивали наилучшее соотношение между числом упоминаний и семантической насыщенностью, становясь, таким образом, «золотой серединой». Графические методы, такие как столбчатые диаграммы и облака слов, делают этот анализ более наглядным для исследователя.

## Заключение

Применение нейросетевых моделей при анализе текста было наиболее эффективным для численного представления анализируемого текста и поиска плагиата среди эссе. Вероятно, из-за относительно малого объема первичного материала и узконаправленной темы использование результатов нейросетевого анализа для кластеризации оказалось неэффективным.

Применение традиционных методов анализа, требующих предварительной подготовки корпуса данных и базирующихся на вычислении частот n-грамм и построении облака слов, позволило более детально изучить смысловую часть текстов в их общем представлении и выделить особенности для различных специальностей или их групп.

## Список литературы

1. Suissa O., Elmalech A., Zhitomirsky-Geffet M. Text analysis using deep neural networks in digital humanities and information science // Journal of the Association for Information Science and Technology. 2021. DOI: 10.1002/asi.24544.

2. Percha B. Modern Clinical Text Mining: A Guide and Review // Annu Rev Biomed Data Sci. 2021. № 4. P. 165–187. DOI: 10.1146/annurev-biodatasci-030421-030931.

3. Spasic I., Nenadic G. Clinical Text Data in Machine Learning: Systematic Review // JMIR Med Inform. 2020. Vol. 8, № 3. P. e17984. DOI: 10.2196/17984.

4. Rasmy L., Xiang Y., Xie Z., Tao C., Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction // NPJ Digit Med. 2021. Vol. 4(1). P. 86. DOI: 10.1038/s41746-021-00455-y.

5. Zhao SS., Hong C., Cai T., Xu C., Huang J., Ermann J., Goodson JN, Solomon HD., Cai T., Katherine PK. Incorporating natural language processing to improve classification of axial spondyloarthritis using electronic health records // Rheumatology. 2020. Vol. 59(5). P. 1059-1065. DOI: 10.1093/rheumatology/kez375.

6. Wang M., Wei Z., Jia M., Chen L., Ji H. Deep learning model for multiclassification of infectious diseases from unstructured electronic medical records // BMC medical informatics and decision making. 2022. Vol. 22(1). DOI: 10.1186/s12911-022-01776-y.

7. Elbattah M., Arnaud É., Gignon M., Dequen G. The Role of Text Analytics in Healthcare: A Review of Recent Developments and Applications // In Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies. 2021. Vol. 5. P. 825-832.

8. Balabaeva K., Funkner AA., Kovalchuk SV. Automated Spelling Correction for Clinical Text Mining in Russian // MIE, 2020. Vol. 270. P. 43-47. DOI: 10.3233/SHTI200119.

9. Yalunin A., Nesterov A., Umerenkov D. RuBioRoBERTa: a pre-trained biomedical language model for Russian language biomedical text mining. 2022. [Электронный ресурс]. URL: https://arxiv.org/abs/2204.03951 (дата обращения: 15.05.2024).

10. Hendrickx I., Voets T., van Dyk P., Kool RB. Using Text Mining Techniques to Identify Health Care Providers With Patient Safety Problems: Exploratory Study // J Med Internet Res. 2021. Vol. 23(7). P. e19064. DOI: 10.2196/19064.

11. Miotto R., Li L., Kidd B. A., Dudley J. T. Deep patient: An unsupervised representation to predict the future of patients from the electronic health records // Scientific reports. 2016. Vol. 6. P. 26094. DOI: 10.1038/srep26094.

12. Woodman R.J., Mangoni A.A. A comprehensive review of machine learning algorithms and their application in geriatric medicine: present and future // Aging Clin Exp Res. 2023. Vol. 35(11). P. 2363-2397. DOI: 10.1007/s40520-023-02552-2.

13. Mugisha C., Paik I. Comparison of Neural Language Modeling Pipelines for Outcome Prediction From Unstructured Medical Text Notes // IEEE Access, 2022. Vol. 10. P. 16489 – 16498. DOI: 10.1109/ACCESS.2022.3148279.

14. Rajkomar A., Dean J., Kohane I. Machine Learning in Medicine // N Engl J Med, 2019. Vol. 380(14). P. 1347-1358. DOI: 10.1056/NEJMra1814259.

15. Zia A., Aziz M., Popa I., Khan S.A., Hamedani A.F., Asif A.R. Artificial Intelligence-Based Medical Data Mining // Journal of Personalized Medicine. 2022. Vol. 12(9). P. 1359. DOI: 10.3390/jpm12091359.

16. Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space // Proceedings of Workshop at ICLR. 2013. [Электронный ресурс]. URL: https://www.cs.ubc.ca/~amuham01/LING530/papers/mikolov2013efficient.pdf (дата обращения: 15.05.2024).

17. Subakti A., Murfi H., Hariadi N. The performance of BERT as data representation of text clustering // J Big Data, 2022. Vol. 9(1). P. 15. DOI: 10.1186/s40537-022-00564-9.

18. Kaitao S., Xu T., Tao Q., Jianfeng L., TieYan L. Mpnet: Masked and permuted pretraining for language understanding // Curran Associates, Inc. 2020 [Электронный ресурс]. URL: https://proceedings.neurips.cc/paper/2020/file/c3a690be93aa602ee2dc0ccab5b7b67e-Paper.pdf (дата обращения: 12.05.2024).

19. Tim vor der Brück, Marc Pouly. Text Similarity Estimation Based on Word Embeddings and Matrix Norms for Targeted Marketing // In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019. Vol. 1. P. 1827–1836.

20. Saeed A. A. M., Taqa A. Y. Textual Plagiarism Detection Using Embedding Models and Siamese LSTM // International Conference for Natural and Applied Sciences, Baghdad, Iraq. 2022. P. 95-100.

21. Abiodun M., Absalom E., Abualigah L., Abuhaija B., Heming J. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data // Information Sciences. 2023. Vol. 622. P. 178-210.

22. Zhu L., Wennan W., Maoyi H., Maomao C., Yiyun W., Zhiming C. A N-gram based approach to auto-extracting topics from research articles // Journal of Intelligent & Fuzzy Systems. 2022. Vol. 43. P. 1-10.

23. Cooshna-Naik, D. Exploring the Use of Tweets and Word Clouds as Strategies in Educational Research // Journal of Learning for Development. 2022. Vol. 9. P. 89–103.

24. Mehta V., Bawa S., Singh J. WEClustering: word embeddings based text clustering technique for large datasets // Complex Intell Systems. 2021. Vol. 7. P. 3211-3224.