

Секция «Искусственный интеллект и цифровая трансформация в бизнесе и  
государственном управлении»

**Анализ методов и алгоритмов автоматизации работы с научными  
документами с использованием интеллектуальных систем машинного  
обучения**

**Научный руководитель – Березкин Дмитрий Валерьевич**

***Гокарев Вадим Николаевич***

*Аспирант*

Московский государственный технический университет имени Н.Э. Баумана,  
Информатика и системы управления, Москва, Россия

*E-mail: vgokarev@mail.ru*

Актуальность применения ИИ в работе с научными документами обусловлена несколькими факторами. Во-первых, непрерывное увеличение объёма научной информации требует создания автоматизированных систем, способных оперативно обрабатывать данные, что уже нашло отражение в ряде современных исследований и разработок. Так, многие научные коллективы и технологические компании, такие как Google Scholar, Elsevier, Springer Nature, а также ведущие университеты, активно разрабатывают и внедряют алгоритмы машинного обучения для классификации, анализа цитируемости и выявления тематических трендов. Во-вторых, автоматизация рутинных процессов позволяет сократить время на подготовку и обработку информации, что особенно важно для специалистов в области научных исследований, библиотекарей и аналитиков, работающих с большими базами данных [1].

Применение интеллектуальных систем машинного обучения позволяет существенно повысить качество извлечения, структурирования и интерпретации информации, что является основополагающим фактором для научно-исследовательской деятельности. Систематизация и автоматизация процессов анализа научных текстов способствуют не только ускорению научного поиска, но и повышению объективности и воспроизводимости результатов исследований.

Сферы применения подобных решений весьма разнообразны. Они охватывают как научно-исследовательскую деятельность, так и образовательные процессы, управление библиотечными фондами, информационное сопровождение научных проектов и мониторинг инновационных технологий. Применение интеллектуальных систем позволяет обеспечить быстрый и точный поиск релевантной информации, оптимизировать процессы рецензирования и даже способствовать выявлению новых направлений исследований [2, 3]. Таким образом, разработка и внедрение методов автоматизации обработки научных документов становится необходимостью для университетов, исследовательских институтов, издательств и организаций, занимающихся анализом научных трендов.

Применение искусственного интеллекта в работе с научными трудами и документами позволяет перевести большую часть обработки документов в автоматизированный процесс, в который не требуется постоянное вмешательство человека. Это позволяет повысить эффективность функционирования организации и снизить человеческие трудозатраты [4].

На сегодняшний день автоматизация работы с текстовыми данными реализуется посредством множества современных методов. Среди них выделяются технологии распознавания

текста с изображений и сканов, позволяющие переводить печатные документы в цифровой формат, системы интеллектуального поиска, использующие анализ контекста и запросы на естественном языке для выявления релевантной информации, а также алгоритмы интеллектуального сравнения, способные оценивать семантическую близость между текстами и обеспечивать высокую точность анализа.

Однако, несмотря на существующие достижения в области автоматизации, остается ряд сложных технических потребностей, которые требуют особого внимания. К таким потребностям можно отнести: извлечение таблиц, которые не имеют четко выраженных границ ячеек; извлечение графических объектов, таких как рисунки и изображения, и анализ их содержания; распознавание рукописных текстов; устранение дефектов печатного текста, включая бледные отпечатки, надписи поверх текста или другие искажения. Эти задачи требуют применения гибридных технологий, включая машинное обучение, компьютерное зрение и алгоритмы обработки естественного языка, что подчеркивает важность комплексного подхода к автоматизации документооборота.

Для повышения эффективности обработки научных документов и снижения трудозатрат сотрудников организаций предлагаются как внедрение новых методов, так и модернизация уже существующих технологий. Современные подходы активно используют нейросетевые модели, которые обеспечивают высокую точность и скорость выполнения задач, связанных с анализом и обработкой текстовой информации.

Среди новых методов можно выделить генерацию краткого содержания документов, реализуемую с помощью модели MBART – «mbart\_ru\_sum\_gazeta» [5], которая способна автоматически создавать содержательные резюме больших текстов, значительно ускоряя процесс ознакомления с документами. Также, еще одним важным направлением является выделение именованных сущностей, например, названий организаций, дат, мест и других ключевых данных с использованием модели SpaCy – «ru\_core\_news\_lg».

Кроме того, модернизация существующих технологий включает адаптацию таких моделей к специфическим требованиям организаций, настройку параметров обработки данных под конкретные сценарии использования и интеграцию их с уже функционирующими системами документооборота. Такой подход не только снижает нагрузку на сотрудников, но и способствует улучшению качества управления информацией, повышая общую производительность рабочих процессов.

Гибридный подход предполагает объединение нескольких методов для достижения максимальной эффективности. Простейшие правила обработки, основанные на экспертных оценках, могут поддерживать работу высокоуровневой нейросети. Такой подход может быть реализован через механизмы, такие как «мягкое голосование (soft voting)», где решения различных моделей комбинируются для получения финального результата.

Современные нейросетевые модели, такие как BERT или GPT способны учитывать контекст, семантику и скрытые связи между словами, что значительно улучшает точность классификации. Гибридный подход, объединяющий экспертные оценки, традиционные алгоритмы и нейросетевые модели, представляет собой мощное решение для автоматизации классификации документов. Внедрение такого подхода должно обеспечить высокую точность и производительность, снижает трудозатраты и повышает качество работы организаций [6].

Таким образом, применение комбинированного подхода автоматизации процессов работы с научными документами, который включает модель представления кластеров документов, гибридный подход с применением экспертных оценок и многоходовую нейронную сеть должен обеспечить значительный прирост к точности обработки. Использование описанного подхода открывает возможности для автоматизации анализа больших объемов данных, минимизации ошибок, связанных с ручным вводом, и повышения прозрачности научной работы.

### Источники и литература

- 1) Torre-López J. de la, Artificial intelligence to automate the systematic review of scientific literature / Torre-López J. de la, Ramírez A., Romero J.R. // Computing. 2023. Т. 105. № 10. С. 2171–2194.
- 2) Мирошниченко М. А. Тенденции Развития Технологий Искусственного Интеллекта В Управлении Документами Организации / М. А. Мирошниченко, А. А. Абдуллаева, М. А. Вовк // ЕГИ. 2024. №1 (51).
- 3) Кубарский А.В. Технологии искусственного интеллекта при построении самообучающейся системы электронного документооборота // Эпомен. 2021. № 61. С. 49–55.
- 4) Белов И.И. Автоматизация функций систем электронного документооборота посредством применения технологий искусственного интеллекта / И. И. Белов // Вестник архивиста. – 2022. – № 3. – С. 772-783.
- 5) Russian spaCy Models Documentation. – [Electronic source] URL: <https://spacy.io/models/ru> (access date: 11.2024).
- 6) Mehdy A.K.M.N., Multi-Input Multi-Output Transformer-Based Hybrid Neural Network for Multi-Class Privacy Disclosure Detection / Mehdy A.K.M.N., Mehrpouyan H. A // Advances in Machine Learning, Data Mining and Computing. : Academy and Industry Research Collaboration Center (AIRCC), 2021. С. 221–241.