УДК 681.3
DOI: 10.26102/2310-6018/2026.59.8.002

Разработка алгоритма распознавания архивных текстов на основе интеграции OCR и больших языковых моделей

В.В. Ковалева✉, П.Ю. Гусев, С.С. Сепкин

Воронежский государственный технический университет, Воронеж, Российская Федерация

Резюме. Статья посвящена повышению качества распознавания архивных текстов, представленных в виде метрических книг Российской империи конца XIX – начала XX века, являющихся уникальным историческим источником для генеалогии и демографии. Стандартные системы оптического распознавания символов, обученные на современных печатных текстах, демонстрируют на подобных документах точность не более 30 % из-за комплекса факторов: ветхость бумажных носителей, выцветание чернил, наличие дореформенной орфографии (буквы ѣ, і, ѳ, ѵ) и смешение кириллического и латинского начертаний. В работе проведён сравнительный анализ двух подходов к улучшению распознавания: дообучение модели PaddleOCR на исторических данных и использование в связке с большой языковой моделью GigaChat-MAX без дообучения. Для постобработки разработан специализированный запрос для языковой модели, включающий правила замены символов, словарь частотных терминов и правила восстановления дореформенной орфографии. Эксперименты на выборке из 50 разворотов метрической книги показали, что дообучение повышает точность до 48 %, а интеграция с языковой моделью – до 65 %, при этом коэффициент ошибок на символ снижается с 70 % до 35 %. Языковая модель исправляет более половины всех ошибок, особенно эффективно заменяя латиницу на кириллицу и восстанавливая исторические окончания. В результате исследования предложен алгоритм распознавания архивных рукописных документов, отличающийся применением комбинации большой языковой модели и модели OCR, и обеспечивающий снижение затрат на разметку данных и вычислительные ресурсы. Практическая значимость: предложенный подход позволяет автоматизировать распознавание метрических книг с точностью 65 %, сокращая ручной труд по транскрипции более чем вдвое.

Ключевые слова: метрические книги, OCR, PaddleOCR, GigaChat-MAX, LLM, дореформенная орфография, CER, WER, TF-IDF.

Для цитирования: Ковалева В.В., Гусев П.Ю, Сепкин С.С. Разработка алгоритма распознавания архивных текстов на основе интеграции OCR и больших языковых моделей. *Моделирование, оптимизация и информационные технологии.* 2026;14(8). URL: https://moitvivt.ru/ru/journal/article?id=2581 DOI: 10.26102/2310-6018/2026.59.8.002

Development of an algorithm for recognizing archival texts based on the integration of OCR and large language models

V.V. Kovaleva✉, P.Yu. Gusev, S.S. Sepkin

Voronezh State Technical University, Voronezh, the Russian Federation

Abstract. The paper addresses the problem of improving text recognition quality for metric books of the Russian Empire (late 19th – early 20th centuries), which are a valuable historical source for genealogy and demography. Conventional optical character recognition (OCR) systems trained on modern printed texts achieve an accuracy of no more than 30 % on such documents due to a combination of factors: paper deterioration, ink fading, pre-reform orthography (letters ѣ, і, ѳ, ѵ) and mixing of Cyrillic and Latin glyphs. The study presents a comparative analysis of two approaches: fine-tuning the PaddleOCR model on historical data versus using a pipeline of baseline PaddleOCR with the large language model

GigaChat-MAX without fine-tuning. For post-processing, a specialised prompt was developed, including character substitution rules, a dictionary of frequent terms, and rules for restoring pre-reform spelling. Experiments on a sample of 50 spreads from a metric book showed that fine-tuning increases accuracy to 48 %, while the integration with the language model achieves 65 %, with the character error rate reduced from 70 % to 35 %. The language model corrects more than half of all errors, being particularly effective in replacing Latin letters with Cyrillic and restoring historical endings. The scientific novelty lies in the first direct comparison of the two strategies and the demonstration of the advantage of post-processing with a language model without the need for labelled data and computational costs. The practical significance is that the proposed approach enables automated recognition of metric books with 65 % accuracy, reducing manual transcription effort by more than half.

Keywords: metric books, OCR, PaddleOCR, GigaChat-MAX, LLM, pre-reform orthography, CER, WER, TF-IDF.

For citation: Kovaleva V.V., Gusev P.Yu., Sepkin S.S. Development of an algorithm for recognizing archival texts based on the integration of OCR and large language models. *Modeling, Optimization and Information Technology.* 2026;14(8). (In Russ.). URL: https://moitvivt.ru/ru/journal/article?id=2581 DOI: 10.26102/2310-6018/2026.59.8.002

Введение

Архивные документы в рамках данной работы представлены метрическими книгами Российской империи конца XIX – начала XX века, которые являются уникальным историческим источником, содержащим записи о рождениях, бракосочетаниях и смертях. Пример архивного документа представлен на Рисунке 1. Несмотря на активную цифровизацию архивных фондов, значительная часть таких документов по-прежнему существует только в виде графических копий. Это существенно ограничивает возможности поиска и анализа содержащихся в них сведений.

При работе с материалами Государственного архива проанализированы страницы метрической книги села Покровское за 1890–1895 гг. Стандартные средства распознавания текста плохо справляются с подобными документами ввиду особенностей дореформенной орфографии, состояния бумаги, качества печати и сложной табличной структуры страниц.

Согласно опубликованным исследованиям, современные OCR-системы способны обеспечивать точность распознавания более 95 % на современных печатных текстах[1]. Однако наблюдается снижение этого показателя при обработке исторических документов [1, 2], что распространяется на исследуемую выборку. Ошибки могут присутствовать даже в часто встречающихся словах и выражениях.

Основным ключом к решению проблемы исследования является использование большой языковой модели GigaChat-MAX для постобработки результатов OCR[2]. База обучения большой языковой модели включает русскоязычные тексты, в том числе и исторические. Преимуществом данного подхода является отсутствие необходимости дообучения OCR-модели и ручной разметки изображений, что снижает затраты на выделяемые ресурсы.

[1] PaddlePaddle. *PaddleOCR.* GitHub. URL: https://github.com/PaddlePaddle/PaddleOCR (дата обращения: 23.03.2026).

[2] *Руководства GigaChat.* Документация для разработчиков от SberDevices и Сбера. URL: https://developers.sber.ru/docs/ru/gigachat/guides/main (дата обращения: 23.03.2026).

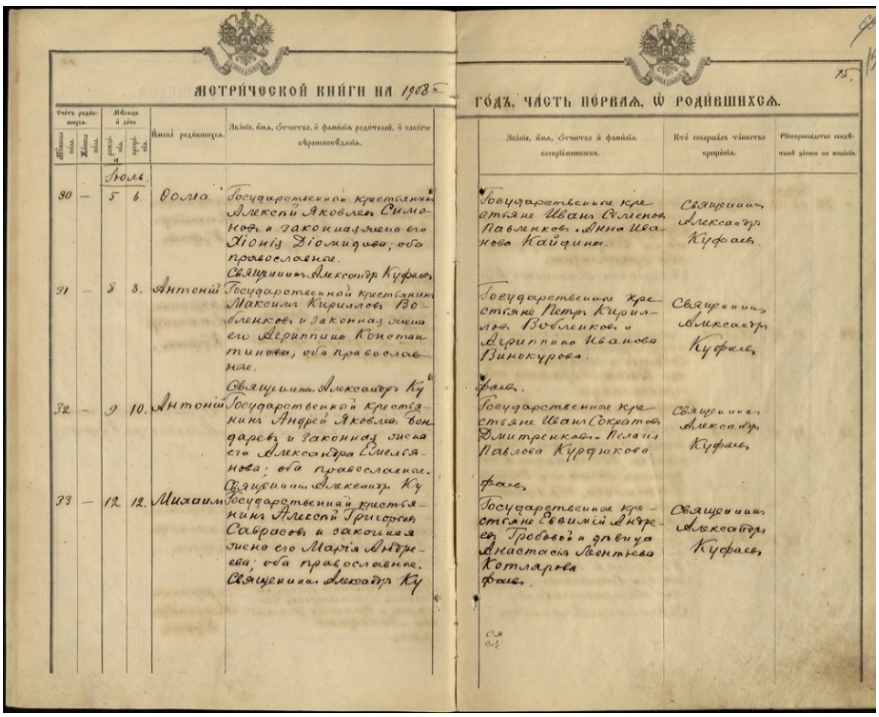

Рисунок 1 – Экземпляр страницы метрической книги
Figure 1 – Copy of a page from the register of births

Материалы и методы

Для оценки качества распознавания используются две метрики: CER (Character Error Rate) и WER (Word Error Rate). Обе основаны на редакционном расстоянии Левенштейна – минимальном количестве односимвольных операций (вставок, удалений, замен), необходимых для преобразования одной строки в другую [1].

CER вычисляется по формуле:

$$CER = \frac{S+D+I}{N} \cdot 100\ \%, \quad (1)$$

где S – количество заменённых символов, D – пропущенных, I – вставленных, N – общее количество символов в эталоне.

Точность на уровне символов:

$$Accuracy = (1 - CER) \cdot 100\ \%. \quad (2)$$

Пример расчёта CER для строки метрической книги: эталон «Метрической книги на 1890 год» (24 символа), результат OCR «МСтРИчеСКой кнИГИ Нл /9ОГОД» (24 символа). Расстояние Левенштейна = 17, CER = 17/24 = 70,8 %, Accuracy_CER = 29,2 %.

WER вычисляется по аналогичной формуле, но оперирует словами:

$$WER = (S_w + D_w + I_w) \,/\, N_w \cdot 100\ \%. \quad (3)$$

В этом примере эталонный текст состоит из 4 слов, а результат содержит 3 слова с ошибками, поэтому значение WER равно 75 %, а точность для слов составляет 25 %.

Если сравнивать методы оценки, то показатель CER точнее фиксирует небольшие отклонения в написании. Это важно для текстов с дореформенной орфографией, так как в них любой символ может менять смысл. Но показатель WER нагляднее показывает, как теряется смысл целых слов. Для базового процесса OCR разница между CER (70 %) и WER (85 %) равна 15 п. п. Это значение указывает на количество слов, которые содержат ошибку только в одном знаке [3].

Для проведения эксперимента исследователи взяли части метрических книг, которые были созданы в конце XIX – начале XX века. Эти документы хранятся в Государственном архиве (метрическая книга села Покровское, 1890–1895 гг.). В выборку вошли 50 разворотов книги, где находятся записи о родившихся людях. И эталонный текст имеет длину 15000 символов, что составляет примерно 250 слов. При этом 85 % строк содержат буквы старого алфавита, а в 70 % строк кириллические знаки перемешаны с латинскими.

Для распознавания текста применена модель PaddleOCR вместе с готовой восточнославянской моделью eslav_PP-OCRv5_mobile_rec. В архитектуру PaddleOCR входит компонент DBNet, который находит области с текстом, и компонент SVTR, который распознает знаки через механизм внимания [4][3]. По итогам базовой обработки точность для символов равна 30 % (CER = 70 %), а точность для слов равна 15 % (WER = 85 %).

Задача распознавания метрических книг предполагает сегментацию изображения на ячейки. Особенностью метрической книги является табличное представление информации, где каждая ячейка может нести строго определённые данные: имя родившегося, сведения о родителях, восприемниках и священнике. Без применения указанного инструмента результат имеет сниженное качество, а модель распознаёт всю таблицу как единый фрагмент информации.

Для решения задачи сегментации применена библиотека OpenCV [2]. Алгоритм включал следующие этапы: перевод изображения в градации серого, повышение контрастности с помощью адаптивной гистограммной эквализации, бинаризацию и морфологическую обработку для выделения структурных элементов таблицы. Вертикальные линии и границы столбцов детектировались с помощью преобразования Хафа. Горизонтальные границы строк определялись на основе расположения цифровых маркеров в столбце с датами, что позволило компенсировать отсутствие или прерывистость горизонтальных линий в исходных документах.

Метод постобработки представляет собой модуль передачи результата распознавания OCR на вход большой языковой модели [5]. Основная задача GigaChat-MAX заключается в функции валидатора для выявления неверных и подозрительных слов в результате OCR. Выбор GigaChat-MAX обусловлен его специализацией на русском языке, высокой контекстной длиной (8192 токенов) и наличием API. Учитывая, что генеративная модель не способна самостоятельно справиться с исправлением ошибок, процесс делегируется модулю TF-IDF-коррекции.

Компоновка системного языкового запроса (промпта) включает следующие разделы: ролевая установка, словарь частотных терминов, правила дореформенной орфографии, типовые фразы метрических книг. Дополнительным преимуществом GigaChat является возможность генерировать ответ в формате JSON, что упрощает интеграцию с другими модулями системы.

Основным инструментом коррекции в данном случае выступает локальный векторизатор на основе TF-IDF. Он является базовым инструментом библиотеки Scikit-learn, не требующим подключения внешних API. В основе метода лежит подход, при котором символьные n-граммы позволяют сохранить фонетическую и структурную близость слов даже при наличии опечаток.

Предварительным этапом векторизации являлся сбор словаря верных слов в контексте метрической книги: фамилии, имена, названия сословий, термины. Этот словарь покрывает 90 % всех слов в архивном наборе исследуемых документов.

[3] eslav_PP-OCRv5_mobile_rec. *PaddlePaddle.* Hugging Face. URL: https://huggingface.co/PaddlePaddle/eslav_PP-OCRv5_mobile_rec (дата обращения: 23.03.2026).

Процесс векторизации заключается в следующем. Каждое слово из словаря разбивается на n-граммы – последовательности из 2, 3 и 4 символов. Например, фамилия «Бондарев» представляется в виде набора фрагментов: «Бо», «он», «нда», «Бон», «онд», «нда», «Бонд», «онда», «ндарев». Каждая n-грамма формирует отдельный признак, в результате чего слово описывается вектором размерностью 500 признаков. Для ошибочного слова выполняется аналогичная процедура, после чего с помощью косинусной меры близости вычисляется расстояние между его вектором и векторами всех слов из словаря. Замена производится при превышении порога сходства 0,55.

Ограничением метода является его низкая эффективность при точности распознавания слова ниже 40–45 %, когда большинство символов распознаны неверно. В таких случаях формируемые n-граммы содержат ошибки, и векторизатор не может найти корректное соответствие в словаре.

Результаты

Результаты применения метода к 50 строкам метрической книги и 50 строкам современного русского текста оказались следующими. Для метрических книг базовый OCR показал точность 30 %, после LLM – 65 %. Прирост составил 35 п. п. – 117 %. Из 1500 символов эталонного текста после LLM правильно распознано 975, LLM исправила 525 – более половины всех ошибок.

На Рисунке 2 показан пример распознавания современного текста. Базовый OCR допустил незначительные ошибки в пунктуации и букве «ё». LLM исправила все ошибки, достигнув идеальной точности.

Рисунок 3 – Пример распознавания метрической книги
Figure 3 – Example of recognition of a register book

На Рисунке 3 приведен пример распознавания метрической книги. Базовый OCR допустил множественные ошибки: замены латиницы на кириллицу, искажения дореформенной орфографии, неправильное распознание имён и сословий. LLM восстановила текст до своего эталонного вида, исправив все типы ошибок.

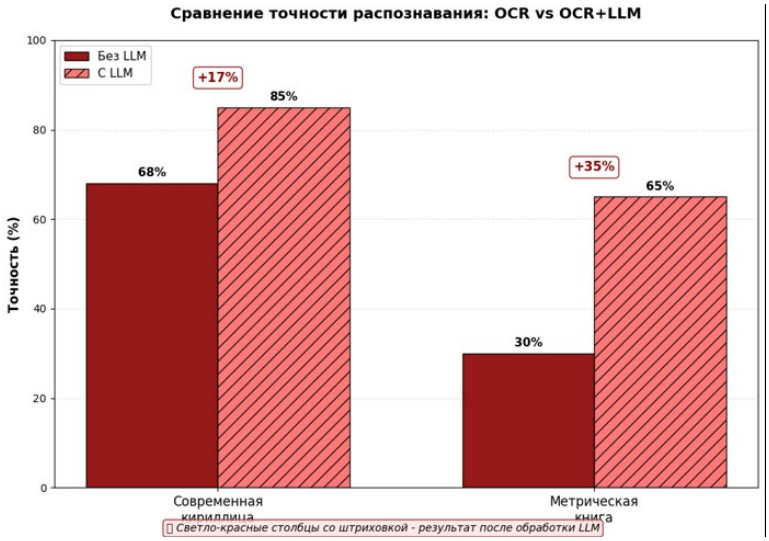

Рисунок 4 – Сравнение прироста точности для двух типов текста
Figure 4 – Comparison of accuracy gain for two types of text

Диаграмма на Рисунке 4 демонстрирует, что метод дает значительный прирост показателя точности для метрических книг, нежели для современной кириллицы. Это обусловлено тем, что система распознавания акцентирует внимание на смешении латиницы и кириллицы, дореформенной орфографии, специфике лексики. Улучшения современного русского языка значительно ниже, так как эти проблемы не свойственны ему.

Таким образом, предложен алгоритм распознавания архивных текстов, представленных в виде метрических книг, представленный на Рисунке 5 и сочетающий применение OCR и больших языковых моделей.

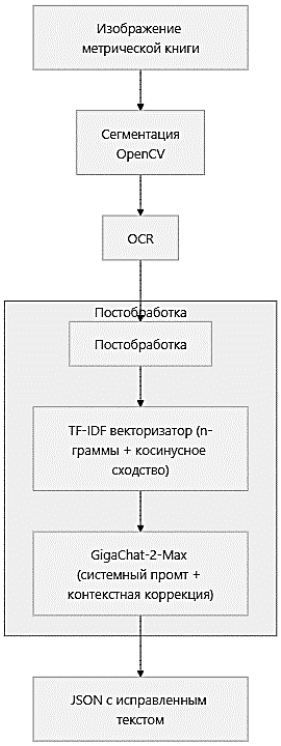

Рисунок 5 – Алгоритм гибридного подхода
Figure 5 – Hybrid approach algorithm

Обсуждение

Сравнивая результаты, можно утверждать, что использование гибридного подхода позволило значительно повысить точность распознавания по сравнению с методом дообучения OCR-модели. Дельта точности 35 п. п. и дельта точности 18 п. п. напрямую связана с качеством работы большой языковой модели. Языковая модель не только исправила ошибки, но и устранила системные искажения исторических данных.

По сравнению с подходом дообучения модели, представленный метод является менее затратным по ресурсам и не менее эффективным по метрике точности, что согласуется с известными оценками эффективности fine-tuning для специализированных OCR-задач [3, 5]. Большой проблемой разметки дореформенной письменности является уникальность каждого почерка писаря, а также значительная эволюция русского языка на протяжении веков. Написание одной и той же буквы может различаться и кардинально не совпадать в почерках разных писарей. Исследуемый метод постобработки не требует дополнительных ресурсов на дообучение и может быть применён к любому документу независимо от почерка. Данный метод является масштабируемым и эффективным.

Роль сегментации и TF-IDF в данной работе значительна, что обеспечивает высокое качество благодаря следующим свойствам: сегментация подаёт на вход строго ограниченные фрагменты данных, предотвращая путаницу в разных столбцах и строках, а векторизатор справился с исправлением ошибок без лишних затрат ресурсов. Проведённые тесты показывают, что даже с использованием большой языковой модели результат без сегментации не позволит добиться такого эффекта.

В литературе описаны попытки улучшить OCR на исторических документах с помощью словарей и правил постобработки [6], однако такие методы обычно дают прирост не более 10–15 процентных пунктов и требуют ручной настройки под каждый тип документа. Применение нейросетевых языковых моделей, таких как BERT или GPT, для исправления OCR-ошибок ранее исследовалось на современных текстах [7], но для исторических документов с дореформенной орфографией подобный подход в отечественной практике практически не применялся. Новизна и актуальность данного исследования заключаются в возможности решить задачу без дополнительного обучения, что открывает новые возможности для цифровизации архивных фондов.

Однако данный подход имеет ряд ограничений: система способна вносить исправления слов только в том случае, если их точность не ниже 45 % от исходного результата до коррекции. Это может приводить к тому, что модель будет галлюцинировать. Также нельзя упускать из внимания качество написанного языкового запроса к модели, так как это главная проблема в формулировке задачи валидации: на чём держать фокус внимания при проверке постредактора. Исследование проводилось на одном типе метрических книг (село Покровское, 1890–1895 гг.), и в качестве прототипа использовался только раздел «о рождении». Работа с другими разделами метрических книг требует масштабирования модуля сегментации в соответствии с особенностями расположения структуры информации на листах книги.

Достигнутая точность является приемлемой для автоматизации процесса извлечения информации из архивного документа [8, 9]. Это позволит сократить затраты на ручную транскрипцию вдвое, что важно для исследования архивов и генеалогических обществ. Так как метод не требует дообучения, его интеграционная возможность является достаточной, чтобы внедрить его в уже готовую архивную систему без затрат на вычислительную инфраструктуру [10].

Нельзя не упомянуть, что максимально эффективный подход в задаче распознавания – это комбинирование дообучения с постобработкой [11].

Перспективным направлением в данной задаче будет использование ансамблевого подхода OCR с последующей унификацией результатов с помощью LLM. Кроме того, предусмотрена возможность использования локальных моделей, что в случае архивных документов является решающим фактором безопасности.

Заключение

В работе предложен метод повышения качества распознавания метрических книг на основе интеграции OCR и LLM.

Исследуемый метод основан на интеграции OCR и LLM и обеспечивает качественное распознавание метрических книг. Специально разработанный промпт является основой LLM-модуля валидации, от которого в значительной степени зависит успех данного подхода. Его основные обязанности – это соединение разорванных частей слов, а также поиск ошибок в результатах OCR.

Важным этапом работы является сегментация, в основе которой лежит деление изображения на отдельные ячейки (блоки информации), выполненное с помощью библиотеки OpenCV. Локальная векторизация TF-IDF на основе символьных n-грамм является основным инструментом, без которого LLM не способна осуществлять коррекцию.

С помощью предложенного алгоритма удалось повысить точность распознавания с 30 % до 65 %, а также снизить показатели CER и WER. С помощью LLM исправлено 525 из 1500 символов, то есть более трети всех ошибок.

Наибольший успех заметен в замене латиницы на кириллицу. Ошибки искажения окончаний были сокращены с 20 % до 10 %, ошибки пропуска символов – с 15 % до 8 %, ошибки вставки лишних символов – с 10 % до 5 %, смысловые ошибки – с 10 % до 7 %.

Практическая значимость работы: метод позволяет автоматизировать распознавание метрических книг с точностью 65 % без дообучения модели и без затрат на разметку данных, уменьшая трудозатраты на транскрипцию более чем в два раза.

Перспективы дальнейших исследований включают автоматическое извлечение структурированных полей, адаптацию промпта для других исторических документов, локальное развёртывание открытых моделей.

СПИСОК ИСТОЧНИКОВ / REFERENCES

1. Левенштейн В.И. Двоичные коды с исправлением выпадений, вставок и замещений символов. *Доклады Академии наук СССР.* 1965;163(4):845–848.
   Levenshtein V.I. Binary codes capable of correcting deletions, insertions, and reversals. *Doklady Akademii Nauk SSSR.* 1965;163(4):845–848. (In Russ.).
2. Smith R. An overview of the Tesseract OCR engine. In: *Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), 23–26 September 2007, Curitiba, Brazil.* IEEE; 2007. P. 629–633. https://doi.org/10.1109/ICDAR.2007.4376991
3. Li Ch., Liu W., Guo R., et al. *PP-OCRv3: More Attempts for the Improvement of Ultra Lightweight OCR System*. arXiv. URL: https://arxiv.org/abs/2206.03001 [Accessed 23rd March 2026].
4. Курейчик В.В., Родзин С.И., Бова В.В. Методы глубокого обучения для обработки текстов на естественном языке. *Известия ЮФУ. Технические науки.* 2022;(2):189–199. https://doi.org/10.18522/2311-3103-2022-2-189-199
   Kureichik V.V., Rodzin S.I., Bova V.V. Deep learning methods for natural language text processing. *Izvestiya SFedU. Engineering Sciences.* 2022;(2):189–199. (In Russ.). https://doi.org/10.18522/2311-3103-2022-2-189-199

5. Shen Y., Wei J., Niu X., et al. Efficient ultra-lightweight convolutional attention network for embedded identity document recognition system. *Image and Vision Computing.* 2026;168:105930. https://doi.org/10.1016/j.imavis.2026.105930
6. Du Y., Li Ch., Guo R., et al. *PP-OCR: A practical ultra lightweight OCR system.* arXiv. URL: https://arxiv.org/abs/2009.09941 [Accessed 23rd March 2026].
7. Vaswani A., Shazeer N., Parmar N., et al. Attention Is All You Need. In: *Advances in Neural Information Processing Systems 30 (NeurIPS 2017), 04–09 December 2017, Long Beach, CA, USA.* 2017. P. 5998–6008.
8. Волощук А.С. Метрические книги как исторический источник: характеристика, анализ, поиск и доступ к ним. В сборнике: *Документ в современном обществе: коммуникативные модели и технологии: Материалы XVI Всероссийской студенческой научно-практической конференции, 07–08 апреля 2023 года, Екатеринбург, Россия.* Екатеринбург: Уральский федеральный университет имени первого Президента России Б.Н. Ельцина; 2023. С. 256–260.
9. Львович Я.Е., Хакназаров И.В., Лямзин И.С. О возможностях поддержки электронного документооборота в организациях. В сборнике: *Социально-экономическое развитие России: проблемы, тенденции, перспективы: Сборник научных статей участников 24-й Международной научно-практической конференции, 22 мая 2025 года, Курск, Россия*. Курск: Университетская книга; 2025. С. 299–301.
Lvovich Ya.E., Khaknazarov I.V., Lyamzin I.S. On the possibilities of supporting electronic document management in organizations. In: *Socio-economic development of Russia: problems, trends, prospects: Collection of scientific articles of participants of the 24th International Scientific and Practical Conference, 22 May 2025, Kursk, Russia*. Kursk: Universitetskaya kniga; 2025. P. 299–301. (In Russ.).
10. Кузнецов А.В. За пределами тематического моделирования: анализ исторического текста с помощью больших языковых моделей. *Историческая информатика.* 2024;(4):47–65. https://doi.org/10.7256/2585-7797.2024.4.72560
Kuznetsov A.V. Beyond Topic Modeling: Analyzing Historical Text with Large Language Models. *Historical Informatics.* 2024;(4):47–65. (In Russ.). https://doi.org/10.7256/2585-7797.2024.4.72560
11. Бакулин А.Ю., Гусев П.Ю., Львович Я.Е. Оптимизация развития цифровизированной организационной системы обслуживания заказов на основе интеграции с машинным обучением прогностических моделей. *Вестник Российского нового университета. Серия: Сложные системы: модели, анализ и управление.* 2026;(1):68–78. https://doi.org/10.18137/RNU.V9187.26.01.P.68
Bakulin A.Yu., Gusev P.Yu., Lvovich Ya.E. Optimizing the development of a digitalized organizational order servicing system based on integration with machine learning of predictive models. *Vestnik of the Russian New University. Series: Complex Systems: models, analysis and management.* 2026;(1):68–78. (In Russ.). https://doi.org/10.18137/RNU.V9187.26.01.P.68

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Ковалева Виктория Викторовна, студентка, Воронежский государственный технический университет, Воронеж, Российская Федерация.
e-mail: vikakot0004@gmail.com

Victoria V. Kovaleva, Student, Voronezh State Technical University, Voronezh, the Russian Federation.

Гусев Павел Юрьевич, доктор технических наук, доцент, заведующий кафедрой искусственного интеллекта и цифровых технологий, Воронежский государственный технический университет, Воронеж, Российская Федерация.
e-mail: gusevpvl@gmail.com
ORCID: 0000-0002-3752-0152

Pavel Yu. Gusev, Doctor of Engineering Sciences, Docent, Head of the Department of Artificial Intelligence and Digital Technologies, Voronezh State Technical University, Voronezh, the Russian Federation.

Сепкин Сергей Сергеевич, студент, Воронежский государственный технический университет, Воронеж, Российская Федерация.
e-mail: sep001@gmail.com

Sergei S. Sepkin, Student, Voronezh State Technical University, Voronezh, the Russian Federation.

Статья поступила в редакцию 03.07.2026; одобрена после рецензирования 31.07.2026; принята к публикации 10.08.2026.

The article was submitted 03.07.2026; approved after reviewing 31.07.2026; accepted for publication 10.08.2026.