DOI 10.37882/2223–2966.2022.12–2.38

РАЗРАБОТКИ ЦЕНТРА КОМПЕТЕНЦИЙ НТИ ПО ТЕХНОЛОГИЯМ ХРАНЕНИЯ И АНАЛИЗА БОЛЬШИХ ДАННЫХ НА БАЗЕ МГУ ИМЕНИ М.В. ЛОМОНОСОВА В ОБЛАСТИ ТЕХНОЛОГИЙ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА

DEVELOPMENTS OF THE BIG DATA STORAGE AND ANALYSIS CENTER ON THE BASIS OF THE LOMONOSOV MOSCOW STATE UNIVERSITY IN THE FIELD OF NATURAL LANGUAGE PROCESSING TECHNOLOGIES

E. Shitov
D. Rakov
I. Tereshenko
E. Kovrova
T. Voronin

Summary. The article is devoted to the analysis of the developments of the Big Data Storage and Analysis Center on the basis of the Lomonosov Moscow State University in terms of natural language processing from the point of their practical potential for solving a wide range of problems, incl. from the point of the prospects for the digital transformation of the national economy of the Russian Federation and the role of natural language processing technologies in this process. The developments of the Big Data Storage and Analysis Center in the field of natural language processing are presented, the analysis of these developments is carried out in terms of the algorithm of their work, technologies in the basis, positive effects and advantages of the implementation, the current UGT. The problems, that can be solved by the practical application of these developments, incl. as part of the digital transformation of the economy of the Russian Federation, are highlighted.

Keywords: artificial intelligence, natural language processing, machine learning, text recognition, NLP, BERT, OCR.

Шитов Егор Александрович
Ведущий специалист Центра компетенций НТИ по направлению «Технологии хранения и анализа больших данных», МГУ имени М.В. Ломоносова, Москва
egor.shitov@digital.msu.ru
Раков Дмитрий Александрович
Ведущий специалист Центра компетенций НТИ по направлению «Технологии хранения и анализа больших данных», МГУ имени М.В. Ломоносова, Москва
rakov.d@digital.msu.ru
Терещенко Игорь Александрович
Ведущий специалист Центра компетенций НТИ по направлению «Технологии хранения и анализа больших данных», МГУ имени М.В. Ломоносова, Москва
igor.tereshchenko@digital.msu.ru
Коврова Екатерина Сергеевна
Ведущий специалист Центра компетенций НТИ по направлению «Технологии хранения и анализа больших данных», МГУ имени М.В. Ломоносова, Москва
lubomirova.ek@digital.msu.ru
Воронин Тимофей Валерьевич
Ведущий специалист Центра компетенций НТИ по направлению «Технологии хранения и анализа больших данных», МГУ имени М.В. Ломоносова, Москва
voronin@digital.msu.ru

Аннотация. Настоящая статья посвящена анализу разработок Центра компетенций НТИ по направлению «Технологии хранения и анализа больших данных» на базе МГУ имени М.В. Ломоносова в части обработки естественного языка с точки зрения их практического потенциала для решения широкого круга задач, в т.ч. с точки зрения перспектив цифровой трансформации национальной экономики Российской Федерации и роли технологий обработки естественного языка в данном процессе. Представлены разработки Центра НТИ в сфере обработки естественного языка, проведен анализ указанных разработок с точки зрения алгоритма их работы, технологий в основе, положительных эффектов и преимуществ внедрения, текущего УГТ. Выделены проблемы, которые может решить практическое применение данных разработок, в т.ч. в рамках цифровой трансформации экономики Российской Федерации.

Ключевые слова: искусственный интеллект, обработка естественного языка, машинное обучение, распознавание текстов, NLP, BERT, OCR.

В условиях цифровизации современного мира одним из ключевых ресурсов становятся цифровые технологии. Согласно Глобальному индексу инноваций 2021 г. Российская Федерация занимает 45-е место в мире и 29-е в Европе по уровню цифровой трансформации [10]. Одной из причин заметного отставания Российской Федерации от мировых цифровых лидеров является неравномерное проникновение цифровых технологий в различные отрасли экономики. Для сокращения данного отставания разработаны и реализуются программы, направленные на повышение уровня цифровой трансформации страны. При этом особое внимание уделяется «сквозным» цифровым технологиям, в частности, технологиям искусственного интеллекта, среди которых выделяют технологии компьютерного зрения, распознавания и синтеза речи, обработки естественного языка (NLP), создания рекомендательных систем и интеллектуальных бизнес-систем поддержки принятия решений [2]. Настоящая статья посвящена разработкам в области технологий обработки естественного языка, которые имеют стратегическое значение не только с точки зрения технического прогресса и инноваций, но и для дальнейшего развития цифровой экономики Российской Федерации.

Целью работы является анализ разработок Центра компетенций НТИ по большим данным на базе МГУ имени М.В. Ломоносова (далее — Центр НТИ) в части распознавания текстов и обработки естественного языка с точки зрения их практического потенциала для решения широкого круга задач в сфере анализа документов и управления большими данными.

Для выполнения обозначенной цели необходимо:
- представить разработки Центра НТИ в сфере распознавания текстов и обработки естественного языка;
- проанализировать указанные разработки с точки зрения алгоритма их работы, технологий в основе, эффектов и преимуществ внедрения, текущего УГТ;
- показать, какие проблемы, в т.ч. в сфере анализа документов и управления большими данными, может решить практическое внедрение данных разработок.

В 2020–2022 гг. в сфере внимания Центра НТИ находились следующие разработки в сфере распознавания текстов и обработки естественного языка:
- умный правовой помощник для предпринимателей (далее — «Умный помощник»);
- нейронная сеть, способная извлекать из судебных актов ключевые данные;
- сервис автоматизированного анализа документов.

Данные разработки входят в число результатов проекта «Средства интеллектуального анализа больших массивов текстов», поддерживаемого Министерством науки и высшего образования Российской Федерации по Договору МГУ имени М.В. Ломоносова с Фондом поддержки проектов Национальной технологической инициативы от 15.12.2021 № 70–2021–00252.

1. Умный помощник

В соответствии с постановлением Правительства Москвы от 22.02.2012 № 66-ПП «О Штабе по защите прав и законных интересов субъектов инвестиционной и предпринимательской деятельности в городе Москве, а также иных рабочих органах Правительства Москвы в сфере инвестиционной и предпринимательской деятельности» [7] в городе Москве создан и функционирует Штаб по защите прав и законных интересов субъектов инвестиционной и предпринимательской деятельности города Москвы (далее — Штаб) [5] и Проектный офис по улучшению инвестиционного климата в городе Москве.

Целями деятельности Штаба являются [7]:
- создание благоприятных условий для ведения инвестиционной и предпринимательской деятельности;
- развитие и защита конкуренции;
- обеспечение гарантий государственной защиты прав и законных интересов субъектов инвестиционной и предпринимательской деятельности в городе Москве.

В качестве ключевых задач Штаба в сфере инвестиционной и предпринимательской деятельности в городе Москве можно выделить следующие (включая, но не ограничиваясь) [6]:
- развитие предпринимательской деятельности;
- выработка антикоррупционных мер в данной сфере;
- устранение административных барьеров;
- содействие сокращению избыточного вмешательства органов исполнительной власти (далее — ОИВ) города Москвы в деятельность хозяйствующих субъектов;
- координация работы межведомственных комиссий при префектурах административных округов города Москвы по устранению административных барьеров при развитии предпринимательства;
- оказание правовой поддержки предпринимателям в целях защиты их прав и законных интересов.

Основными причинами обращений предпринимателей в Штаб являются две проблемы:

- набор типовых ошибок в комплектах документов, подаваемых предпринимателями в ОИВ;
- административные барьеры со стороны ОИВ.

Для решения данных проблем и снижения количества обращений в Штаб Центра НТИ разработал решение «Умный помощник», который на основании переписки с предпринимателем формирует ответ на правовые вопросы.

«Умный помощник» умеет распознавать входящий запрос, запрашивать необходимый перечень документов с определением их типов и ключевых данных, проверять полноту комплекта документов и наличие потенциальных проблем, автоматически формирует проект ответного документа с возможностью оценки правовых рисков.

Этапы алгоритма работы «Умного помощника»:
- определение темы обращения;
- наличие отказа от ФОИВ или общая проверка (распознавание поводов для отказа);
- формирование перечня необходимых документов (с помощью заранее сформированных и загруженных юристами в помощник чек-листов);
- распознавание графических документов;
- проверка комплектности пакета документов/ выявленных проблем;
- формирование рекомендаций для устранения недостатков/ автоматическое заполнение заявления.

Проблемы, которые помогает смягчить/ нивелировать внедрение решения:
- массовые отказы со стороны ОИВ из-за неполного комплекта документов/ комплекта с ошибками, поданного предпринимателями;
- высокая нагрузка сотрудников ОИВ, вынужденных заниматься неполными/ неправильно оформленными комплектами документов;
- низкая операционная эффективность сотрудников ОИВ;
- отсутствие быстрой и квалифицированной поддержки административно-юридического характера для предпринимателей;
- риск возникновения административных барьеров/ нарушений законодательства со стороны ОИВ;
- замедленное развитие бизнеса.

Эффекты от внедрения решения [1]:
- проверка документов на полноту и правильность, а также на наличие типовых ошибок, что позволяет предпринимателям избежать отказа или приостановления решения по поданному заявлению;
- снижение нагрузки сотрудников Штаба, которым остается ручная обработка нетиповых запросов;
- повышение их операционной эффективности до 80%;
- обеспечение предпринимателей быстрой и квалифицированной поддержкой;
- уменьшение риска возникновения нарушений законодательства и связанных с этим расходов;
- стимулирование развитие бизнеса.

В основе решения лежат:
- масштабируемая архитектура на реляционных базах, новые кейсы добавляются на уровне «конструктора»;
- полноценный бек-энд, позволяющий разместить решение на любой платформе, включая мессенджеры;
- машинное обучение на каждом этапе (NLP-практики для классификации и верификации документов, а также для их графического распознавания).

Преимуществами решения с технической точки зрения являются:
- классификация визуально неотличимых документов (устав, договор, накладная и т.п.);
- отделение подписи от рукописного текста (в части распознавания графических атрибутов);
- масштабируемость.

В настоящее время решение находится на стадии УГТ 4: создан прототип, который прошел проверку на примере коммуникации Росреестра и предпринимателей, получивших от ведомства отказы или приостановления запросов. За разработку прототипа команда Центра НТИ получила премию Мэра Москвы «Лидеры цифровой трансформации» по треку «Искусственный интеллект в городе» [3]. Прототип нуждается в дальнейшей доработке до стадии пилотирования для дальнейшего использования в сфере деятельности Штаба и Департамента предпринимательства и инновационного развития города Москвы.

2. Нейронная сеть для извлечения ключевой информации из судебных актов

При решении задачи по формированию базы знаний и информации в части правоприменительной практики в отношении субъектов малого и среднего предпринимательства (далее — МСП) на хакатоне Audithon 2021 Счетной палаты Российской Федерации командой Центра НТИ был проведен «анализ влияния размеров штрафов, которые назначались предпринимателям за их деятельность без регистрации, на ко-

личество субъектов МСП и их оборот в отдельных регионах Российской Федерации (Москве, Республике Башкортостан, Ульяновской области)» [4]. Для решения данной задачи была создана нейронная сеть, которая с помощью алгоритмов распознавания выделила ключевые элементы из 1,5 тысяч загруженных в нее судебных решений, в т.ч.:

- истец;
- ответчик;
- статья, по которой вынесено решение;
- дата вынесения решения;
- суть решения (резолютивная часть);
- размер штрафа;
- субъект/ город, в котором рассматривалось дело.

Этапы алгоритма работы модели:

- распознавание входящего графического документа;
- выделение в нем ключевых элементов;
- формирование исходящего документа с разметкой.

Проблемы, которые помогает смягчить/ нивелировать внедрение решения:

- существенные временные затраты на поиск нужной информации в решении суда (особенно при необходимости массовой обработки десятков и сотен решений);
- отсутствие статистики/ неполные данные по показателям, имеющимся в судебных решениях;
- наличие скрытых взаимосвязей между выделенными показателями и статистикой правоприменительной практики;
- отсутствие моделей влияния жесткости политики субъекта Российской Федерации на показатели развития МСП.

Эффекты от внедрения модели:

- быстрое выделение ключевой информации в решении суда;
- возможность формирования статистических баз данных по распознаваемым параметрам (например, по назначенным штрафам и прочим санкциям в отношении субъектов МСП и физических лиц, по количеству уникальных ответчиков и проч.);
- выявление корреляций между найденными показателями и статистикой правоприменительной практики, в т.ч. полученной на основе анализа судебных актов;
- моделирование влияния жесткости политики субъекта Российской Федерации на показатели развития МСП.

В основе модели лежат:

- NER-технологии для анализа и структуризации текстов на естественном языке;
- обучение нейронной сети на основе BERT-архитектуры;
- механизмы графического распознавания текстов (OCR).

Преимуществами модели с технической точки зрения являются:

- возможность автоматизированной обработки судебных актов;
- возможность проведения дальнейшего обучения нейронной сети для улучшения машинного понимания судебных актов;
- масштабируемость.

В настоящее время решение находится на стадии УГТ 4: создан прототип решения, который прошел тестирование в условиях хакатона Audithon 2021.

3. Сервис автоматизированного анализа документов

Согласно постановлению Правительства Москвы № 741-ПП от 04.10.2017 «Об утверждении порядков предоставления субсидий из бюджета города Москвы в целях государственной поддержки субъектов малого и среднего предпринимательства и признании утратившим силу постановления Правительства Москвы от 15.09.2015 № 587-ПП» [8] для подачи заявки на субсидию предпринимателю необходимо прикрепить к заявке набор заверенных документов. С помощью алгоритмов распознавания сервис, разработанный командой Центра НТИ [9], анализирует, были ли поданы необходимые документы и нет ли в них ошибок, сравнивая каждый документ с «идеальным» документом, необходимым для успешного рассмотрения заявки. При наличии ошибок или неполного комплекта документов сервис сформирует соответствующие рекомендации.

Этапы алгоритма работы сервиса:

- определение темы обращения;
- общая проверка (распознавание поводов для отказа);
- распознавание графических документов;
- проверка выявленных проблем;
- формирование рекомендаций для устранения недостатков (зеленый цвет — все хорошо, в случае отсутствия печати/ подписи и проч. недочеты будут выделены соответствующими маркерами).

Проблемы, которые помогает смягчить/ нивелировать внедрение решения:

Таблица 1. Сводная информация по разработкам Центра компетенций НТИ по технологиям хранения и анализа больших данных на базе МГУ имени М.В. Ломоносова в сфере распознавания текстов и обработки естественного языка

Название разработки	Технологии в основе	Эффекты	УГТ
Умный помощник	• масштабируемая архитектура на реляционных базах; • бек-энд для размещения на любой платформе, включая мессенджеры; • NLP-практики для классификации и верификации документов, а также для их графического распознавания	• снижение числа отказов или приостановлений решений по поданному заявлению; • снижение нагрузки на сотрудников Штаба; • обеспечение предпринимателей быстрой и квалифицированной поддержкой; • уменьшение риска возникновения нарушений законодательства и связанных с этим расходов; • повышение операционной эффективности до 80%; • стимулирование развитие бизнеса	4
Нейронная сеть для извлечения ключевой информации из судебных актов	• NER-технологии обработки текстов на естественном языке; • обучение нейронной сети на основе BERT-архитектуры; • механизмы графического распознавания текстов (OCR)	• выделение ключевой информации в решении суда; • возможность формирования статистических баз данных по распознаваемым параметрам; • моделирование влияния жесткости политики субъекта Российской Федерации на показатели развития МСП	4
Сервис автоматизированного анализа документов	• принципы transfer learning; • передовые модели машинного обучения нейронных сетей (BERT, YOLO, Inception); • механизмы графического распознавания текстов (OCR); • обработка текстов на естественном языке (NLP)	• снижение числа отказов или приостановлений решений по заявлениям на предоставление субсидии; • обеспечение предпринимателей быстрой и квалифицированной поддержкой; • увеличение доли предоставленных субсидий; • уменьшение риска возникновения нарушений законодательства и связанных с этим расходов; • стимулирование развитие бизнеса	4

Источник: составлено авторами

- массовые отказы со стороны ОИВ в предоставлении субсидии из-за ошибок в документах, поданных предпринимателями;
- высокая нагрузка сотрудников ОИВ, вынужденных заниматься неполными/ неправильно оформленными комплектами документов;
- отсутствие быстрой и квалифицированной поддержки административно-юридического характера для предпринимателей;
- низкая доля предоставленных субсидий относительно общего числа поданных заявок;
- риск возникновения административных барьеров/ нарушений законодательства со стороны ОИВ;
- замедленное развитие бизнеса.

Эффекты от внедрения сервиса:
- проверка документов на наличие типовых ошибок, что позволяет предпринимателям избежать отказа или приостановления решения по поданному заявлению на предоставление субсидии;
- снижение нагрузки на сотрудников ОИВ, обрабатывающих заявления на предоставление субсидии;
- обеспечение предпринимателей быстрой и квалифицированной поддержкой;
- увеличение доли предоставленных субсидий (конечная цель: одна заявка — одна субсидия);
- уменьшение риска возникновения нарушений законодательства и связанных с этим расходов;
- стимулирование развитие бизнеса.

В основе сервиса лежат:
- принцип transfer learning;
- передовые модели машинного обучения нейронных сетей (BERT, YOLO, Inception);
- механизмы графического распознавания текстов (OCR);
- обработка текстов на естественном языке (NLP).

Преимуществами сервиса с технической точки зрения являются:
- отделение подписей и печатей от рукописного текста (в части распознавания графических атрибутов);
- масштабируемость.

В настоящее время решение находится на стадии УГТ 4: создан прототип решения, который проходит проверку на портале i.moscow Московского инновационного кластера.

Сводная информация по всем трем разработкам Центра НТИ представлена в таблице 1.

Представленные разработки имеют потенциал применения в сфере государственного управления, автоматизации, структуризации и оперативного анализа данных и будут способствовать развитию цифровой экономики Российской Федерации. Таким образом, важной задачей становится их доработка, анализ наилучшего опыта их практического применения и дальнейшее внедрение, и масштабирование в различных отраслях.

Благодарности

Исследование выполнено при финансовой поддержке в рамках реализации программы Центров компетенций Национальной технологической инициативы на базе Московского государственного университета имени М.В. Ломоносова (договор о предоставлении средств юридическому лицу, индивидуальному предпринимателю на безвозмездной и безвозвратной основе в форме гранта, источником финансового обеспечения которых полностью или частично является субсидия, предоставленная из федерального бюджета № 70–2021–00252 от 15.12.2021).

ЛИТЕРАТУРА

1. В Центре НТИ по большим данным МГУ разрабатываются NLP-решения по распознаванию документов // Официальный сайт Центра компетенций НТИ по технологиям хранения и анализа больших данных на базе МГУ имени М.В. Ломоносова. URL: https://bigdata.msu.ru/news/202/ (дата обращения: 24.11.2022)
2. Дорожная карта развития «сквозной» цифровой технологии «Нейротехнологии и искусственный интеллект» // Официальный сайт Министерства цифрового развития, связи и массовых коммуникаций Российской Федерации. URL: https://digital.gov.ru/uploaded/files/07102019ii.pdf (дата обращения: 24.11.2022)
3. Команда Центра компетенций НТИ по большим данным МГУ выиграла премию мэра Москвы в размере миллиона рублей // Официальный сайт Центра компетенций НТИ по технологиям хранения и анализа больших данных на базе МГУ имени М.В. Ломоносова. URL: https://bigdata.msu.ru/news/143/ (дата обращения: 24.11.2022)
4. Команда Центра компетенций НТИ по большим данным МГУ — победитель хакатона Audithon 2021 Счетной палаты РФ // Официальный сайт Центра компетенций НТИ по технологиям хранения и анализа больших данных на базе МГУ имени М.В. Ломоносова. URL: https://bigdata.msu.ru/news/190/ (дата обращения: 24.11.2022)
5. Официальный сайт Штаба по защите прав и законных интересов субъектов инвестиционной и предпринимательской деятельности города Москвы // URL: http://shtab.mos.ru (дата обращения: 24.11.2022)

6. Положение о Штабе по защите прав и законных интересов субъектов инвестиционной и предпринимательской деятельности в городе Москве (приложение 1 к Постановлению Правительства Москвы от 22.02.2012 № 66-ПП) // Официальный сайт Мэра Москвы. URL: https://www.mos.ru/dipp/documents/normativnye-pravovye-akty-goroda-moskvy/view/243343220/ (дата обращения: 24.11.2022)
7. Постановление Правительства Москвы от 22.02.2012 № 66-ПП «О Штабе по защите прав и законных интересов субъектов инвестиционной и предпринимательской деятельности в городе Москве, а также иных рабочих органах Правительства Москвы в сфере инвестиционной и предпринимательской деятельности» // Официальный сайт Мэра Москвы. URL: https://www.mos.ru/dipp/documents/normativnye-pravovye-akty-goroda-moskvy/view/243343220/ (дата обращения: 24.11.2022)
8. Постановление Правительства Москвы № 741-ПП от 04.10.2017 «Об утверждении порядков предоставления субсидий из бюджета города Москвы в целях государственной поддержки субъектов малого и среднего предпринимательства и признании утратившим силу постановления Правительства Москвы от 15.09.2015 № 587-ПП» // Официальный сайт Мэра Москвы. URL: https://www.mos.ru/authority/documents/doc/37071220/ (дата обращения: 24.11.2022)
9. Сервис по распознаванию документов, созданный при участии ЦК НТИ по большим данным МГУ, поможет бизнесу получить субсидии // Официальный сайт Центра компетенций НТИ по технологиям хранения и анализа больших данных на базе МГУ имени М.В. Ломоносова. URL: https://bigdata.msu.ru/news/191/ (дата обращения: 24.11.2022)
10. Global Innovation Index 2021 // URL: https://www.wipo.int/edocs/pubdocs/en/wipo_pub_gii_2021.pdf (дата обращения: 24.11.2022)

© Шитов Егор Александрович (egor.shitov@digital.msu.ru), Раков Дмитрий Александрович (rakov.d@digital.msu.ru),
Терещенко Игорь Александрович (igor.tereshchenko@digital.msu.ru), Коврова Екатерина Сергеевна (lubomirova.ek@digital.msu.ru),
Воронин Тимофей Валерьевич (voronin@digital.msu.ru).
Журнал «Современная наука: актуальные проблемы теории и практики»

Московский государственный университет имени М.В. Ломоносова