13. *Miller E.G., Viola P.A.* Ambiguity and constraint in mathematical expression recognition, *in AAAI-98/IAAI-98 Proceedings, July 26-30, 1998, Madison, Wisconsin: AAAI, 1998,* pp. 784-791.
14. *Ong Kai Bin, Yew Kwang Hooi, Said Jadid Abdul Kadir, Haruhiro Fujita and Luqman Hakim Rosli.* Enhanced Symbol Recognition based on Advanced Data Augmentation for Engineering Diagrams, *International Journal of Advanced Computer Science and Applications (IJACSA),* 2022, 13 (5). Available at: http://dx.doi.org/10.14569/IJACSA.2022.0130563.
15. *Bhanbhro H., Yew K.H., Kusakunniran W., Amur Z.* A Symbol Recognition System for Single-Line Diagrams Developed Using a Deep-Learning Approach, *Applied Sciences,* 2023, 13, pp. 8816. Available at: https: // doi.org/10.3390/app13158816.
16. *Moreno-García, C.F.; Elyan, E.; Jayne, C.* Heuristics-Based Detection to Improve Text/Graphics Segmentation in Complex Engineering Drawings, *In Proceedings of the Engineering Applications of Neural Networks: 18th International Conference (EANN 2017), Athens, Greece, 25–27 August 2017,* pp. 87-98.
17. *Pratt W.K.* Digital image processing. New York: Wiley, 1991, 698 p.
18. *Muthukrishnan R, Radha M.* Contour selection algorithms for image segmentation, *International Journal of Computer Science & Information Technology (IJCSIT),* 2014,Vol. 3, No. 6, pp. 259-267.
19. *Poynter Ya.* Programmiruem s PyTorch: Sozdanie prilozheniy glubokogo obucheniya [Programming PyTorch for Deep Learning]. Saint Petersburg: Piter, 2020, 256 p.
20. *Liu Yuxi (Hayden).* PyTorch 1.x Reinforcement Learning Cookbook. Over 60 recipes to design, develop, and deploy self-learning AI models using Python. Birmingham–Mumbai: Packt, 2019, 527 p.

Безуглов Дмитрий Анатольевич – Ростовский филиал Российской таможенной академии; e -mail: bezuglovda@mail.ru; г. Ростов-на-Дону, Россия; д.т.н.; профессор.

Мищенко Марина Сергеевна – Южный федеральный университет; e-mail: yourhuckleberrybtw@mail.ru; г. Ростов-на-Дону, Россия; студент.

Мищенко Сергей Евгеньевич – ФГУП «Ростовский научно-исследовательский институт радиосвязи»; e-mail: mihome@yandex.ru; г. Ростов-на-Дону, Россия; д.т.н.; профессор.

Bezuglov Dmitry Anatolyevich – Rostov branch of the Russian Customs Academy; e-mail: bezuglovda@mail.ru; Rostov-on-Don, Russia; dr. of eng. sc.; professor.

Mishchenko Marina Sergeevna – Southern Federal University; e -mail: yourhuckleberrybtw@mail.ru; Rostov-on-Don, Russia; student.

Mishchenko Sergey Evgenievich – FSUE Rostov Scientific Research Institute of Radio Communications; e-mail: mihome@yandex.ru; Rostov-on-Don, Russia; dr. of eng. sc.; professor.

УДК 004.67 DOI 10.18522/2311-3103-2025-3-144-159

А.Г. Бондаренко, А.Г. Кравец

ИДЕНТИФИКАЦИЯ КЛЮЧЕВЫХ ТЕХНОЛОГИЙ НА ОСНОВЕ СБОРА И АНАЛИЗА ДАННЫХ ИЗ ОТКРЫТЫХ РУССКОЯЗЫЧНЫХ ИСТОЧНИКОВ

Данная статья посвящена разработке и апробации нового подхода к сбору, обработке и анализу открытых данных на русском языке для идентификации ключевых технологических направлений. Для решения задачи формирования и последующего анализа структурированных датасетов разработаны и программно реализованы методы веб-скрейпинга, обработки естественного языка и анализа временных рядов. Описанный в статье подход впервые применен для извлечения и структурирования информации из научных статей, новостных ресурсов и патентной документации на русском языке. В результате анализа полученного датасета научных публикаций выделены 30 наиболее часто упоминаемых биграмм и столько же триграмм технологических терминов. На основе анализа частотности биграмм и триграмм выделены ключевые технологические термины, которые затем использованы для комплексной фильтрации по ключевым технологиям. Комплексная фильтрация позволила осуществить поиск русскоязычных патентов и их сбор для дальнейшего анализа. В результате предварительной обработки полученной патентной информации сформированы временные ряды патентной активности. Программная система идентификации ключевых технологий реализована на JavaScript и Python с использованием библиотек Selenium

и BeautifulSoup для веб-скрейпинга, NLTK и Scikit-learn для обработки и анализа текстовых данных. Исследование динамики развития ключевых технологий во времени позволило выявить периоды интенсивной патентной деятельности и снижения интереса к той или иной технологии. Результаты, изложенные в статье, создают основу для дальнейшей разработки методов машинного обучения с целью прогнозирования технологического развития и выявления перспективных направлений прикладных исследований.

Веб-скрейпинг; анализ текста; обработка естественного языка; ключевые термины; биграммы; триграммы; патентная активность; временные ряды; прогнозирование технологического развития; открытые данные.

A.G. Bondarenko, A.G. Kravets

IDENTIFICATION OF KEY TECHNOLOGIES BASED ON COLLECTION AND ANALYSIS OF DATA FROM OPEN RUSSIAN-LANGUAGE SOURCES

This article is devoted to the development and approbation of a new approach to the collection, processing and analysis of open data in the Russian language for identification of key technological trends. To solve the problem of formation and subsequent analysis of structured datasets methods of web scraping, natural language processing and analysis of time-series have been developed and implemented via programming. The approach described in the article has been applied for the first time in order to extract and structure information from scientific articles, news resources and patent documentation in the Russian language for the first time. As a result of analyzing the obtained dataset of scientific publications, 30 most frequently mentioned bigrams and the same number of trigrams of technological terms have been identified. Based on the frequency analysis of bigrams and trigrams, key technological terms were identified which then were used for complex filtration on key technologies. Complex filtration enabled to fulfill the search of patents in Russian and their collection for further analysis. As a result of preprocessing of the obtained patent data time series of patent activity have been formed. The programme system of key technological identification has been implemented in JavaScript and Python using Selenium and BeautifulSoup libraries for web scraping, NLTK and Scikit-learn for text data processing and analysis. The study focused on the dynamics of the development of key technologies over time has allowed to identify periods of intensive patent activity and declining interest in this or that kind of technology. The results presented in the article provide a basis for further development of machine learning methods for the purpose of predicting technological development and identifying promising areas of applied research.

Web scraping; text analysis; natural language processing; key terms; bigrams; trigrams; patent activity; time series; predicting of technological development; open data.

Введение. Прогнозирование развития технологий становится все более важной задачей в условиях стремительного роста количества инноваций, глобализации науки и новых направлений технологического лидерства [1, 2]. Анализ актуальных инструментов прогнозирования технологического развития [3] позволил выявить ряд несовершенств и проблем существующих подходов. Прежде всего, они касаются качества прогноза, а именно недостаточно высокой точности и ошибок прогноза. Как российские, так и зарубежные ученые подчеркивают сложность идентификации «прорывных» технологий на основе анализа открытых данных [4, 5]. Современные подходы к прогнозированию опираются на использование больших объемов данных из разнообразных источников, таких как научные статьи, патенты, новости и социальные сети [6, 7]. Однако, несмотря на доступность этих данных, остается актуальной проблема их эффективного сбора, анализа и интерпретации для целей прогнозирования [8, 9]. При этом необходимо отметить явный дефицит качественных наборов данных (датасетов) на русском языке, несмотря на многочисленные попытки реализации таких проектов [10, 11].

Актуальность исследования обусловлена необходимостью разработки новых подходов к анализу технологических тенденций, основанных на обработке открытых данных на русском языке [12]. Основной целью исследования является разработка и апробация нового подхода к сбору, обработке и анализу открытых данных для идентификации ключевых технологических направлений. Сбор, обработка и анализ открытых данных с использованием методов веб-скрейпинга и анализа текстовых данных позволит выявить ключевые технологические термины и тенденции, а также сформировать перечень технологий для дальнейшего анализа патентной активности.

Подходы к анализу патентных баз данных для идентификации и прогнозирования технологических тенденций широко обсуждаются современными исследователями [13, 14]. Все большее внимание в публикациях уделяется разработке и совершенствованию методов машинного обучения для анализа патентов [15]. Наилучших результатов по точности моделей достигают проекты, связанные с идентификацией вакантных и перспективных технологий в отдельных предметных областях [16, 17]. Также задача идентификации технологических тенденций актуальна для отдельных высокотехнологичных компаний [18]. Однако такие подходы существенно снижают возможности идентификации «прорывных» результатов междисциплинарных исследований. Для нивелирования этого риска применимы методы конструирования будущих событий [19], но они уступают по своим показателям методам машинного обучения в случае использования качественных датасетов.

Ключевым аспектом представленного в данной статье исследования является реализация методов интеллектуального анализа текстовых данных для выявления значимых технологических терминов и последующее формирование временных рядов патентной активности. Разработанный подход позволяет визуализировать динамику развития технологий и выявить периоды интенсификации инновационной деятельности, что необходимо для более глубокого понимания текущего состояния и тенденций в исследуемых технологических областях.

Статья структурирована следующим образом. Раздел 2 посвящен детальному описанию процессов сбора данных из открытых источников, включая использование методов веб-скрейпинга и извлечения информации из различных типов ресурсов. Раздел 3 фокусируется на этапе обработки собранных данных, включающем в себя предобработку текстовой информации, выделение ключевых терминов, а также формирование биграмм и триграмм для последующего анализа. Раздел 4 охватывает процесс формирования временных рядов на основе патентной активности выделенных технологий, а также создание визуализаций для анализа динамики их развития. Наконец, Раздел 5, содержит основные выводы, полученные в результате проведенного исследования, и намечены перспективы дальнейшего развития работы.

Сбор данных из открытых источников. Процедура сбора данных обычно реализуется с помощью веб-краулинга [20] или парсинга веб-страниц и хранилищ документов [21]. Однако эти подходы ограничивают возможности анализа собранных данных семантикой изначальных поисковых запросов. Поэтому, в рамках исследования реализован метод веб-скрейпинга (рис. 1) для сбора данных о технологиях из открытых русскоязычных источников, включая научные статьи, новостные ресурсы и патенты.

Результатом этого процесса является формирование структурированных датасетов. Полученные данные, после предварительной обработки, представляются в виде наборов ключевых терминов (биграмм и триграмм) и их частотности. Это позволяет определить перечень технологий, которые в дальнейшем используются для фильтрации патентной информации.

Сбор данных из научных статей. Для сбора данных из научных статей выполняется поиск ссылок на публикации. Полученные URL-адреса сохраняются в текстовый файл, где каждая ссылка отделена переносом строки.

Для извлечения структурированных данных из этих веб-страниц разработан специализированный веб-скрейпер. Этот веб-скрейпер предназначен для сбора данных с сайта eLibrary. Он обеспечивает авторизацию на сайте посредством Selenium, с использованием логина и пароля, и работает в headless-режиме браузера. Ссылки на страницы статей загружаются из созданного текстового файла. Далее, Selenium загружает HTML-код каждой страницы, а BeautifulSoup извлекает необходимые данные: название статьи, год публикации, аннотацию и ключевые слова. Ключевые слова извлекаются из таблиц с определенной структурой.

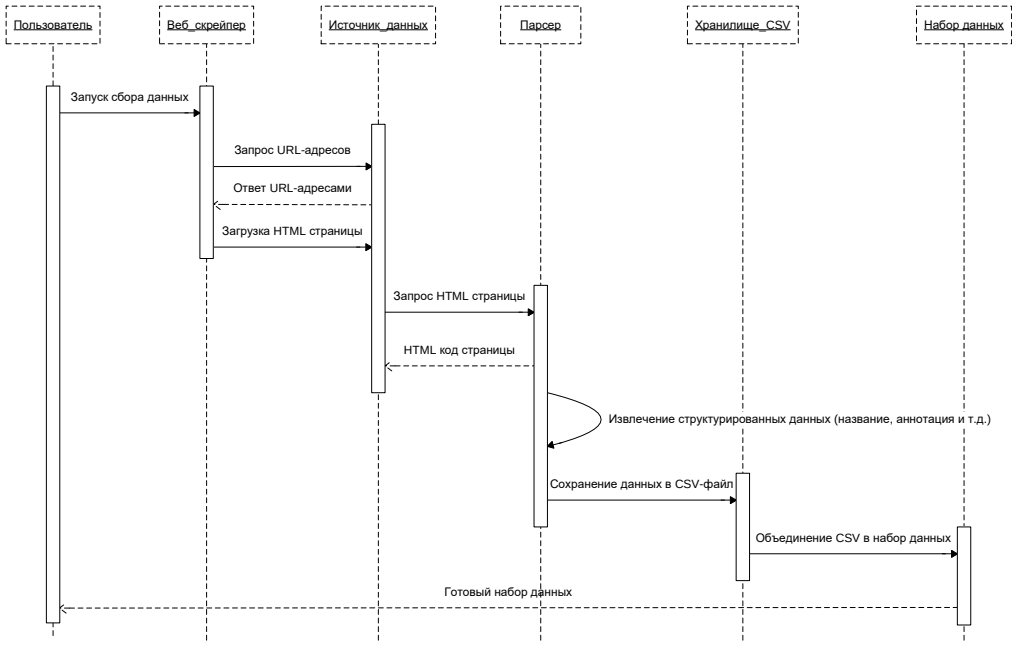

Рис. 1. Диаграмма последовательности метода веб-скрейпинга

В дальнейшем, отдельные CSV-файлы объединяются в единый файл для проведения дальнейшего анализа, с сохранением информации о годах публикации (как временных рядов) и полных текстах документов.

Собранные данные сохраняются в единый датасет, содержащий всю необходимую информацию о научных статьях (рис. 2).

	title	year	abstract	keywords
0	ЦИФРОВИЗАЦИЯ ПРОИЗВОДСТВА: ТЕОРЕТИЧЕСКАЯ СУЩНОСТЬ ...	2018	В современных условиях развитие экономики связываю...	ЭКОНОМИЧЕСКАЯ СИСТЕМА, ПРОМЫШЛЕННОЕ ПРОИЗВОДСТВО, ...
1	ОБ УСТАНОВКАХ КОГНИТИВНОЙ НАУКИ И АКТУАЛЬНЫХ ПРОБЛ...	2004	Начальные этапы становления когнитивной науки были...	Ключевые слова не найдены
2	ЦИФРОВАЯ СОЦИАЛИЗАЦИЯ В КУЛЬТУРНО-ИСТОРИЧЕСКОЙ ПАР...	2018	В настоящей статье раскрываются методология, метод...	КУЛЬТУРНО-ИСТОРИЧЕСКАЯ ПСИХОЛОГИЯ, ЦИФРОВЫЕ ТЕХНОЛ...
...	...	...	...	...
647	НОВЫЕ ПРОЕКТЫ ОСВОЕНИЯ РОССИЙСКОЙ АРКТИКИ: ПРОСТРА...	2020	В статье обобщаются результаты анализа 23 современ...	ПРОЕКТЫ РЕСУРСНОГО ОСВОЕНИЯ АРКТИКИ, ПРОСТРАНСТВЕН...
648	LEGALTECHИ ЮРИСТЫ БУДУЩЕГО	2017	Роботизация юридической профессии стала одной из н...	Ключевые слова не найдены
649	ОБУЧЕНИЕ ЦИФРОВЫМ НАВЫКАМ РАБОТНИКОВ КОНТРАКТНЫХ С...	2019	Цифровизация затронула все сферы жизнедеятельности...	Ключевые слова не найдены

Рис. 2. Датасет по собранным статьям

Датасет имеет следующую структуру:

- title – название статьи;
- year – год публикации статьи;
- abstract – аннотация статьи;
- keywords – ключевые слова.

В результате получено 650 научных публикаций на русском языке, связанных с развитием технологий.

Сбор данных из новостных ресурсов. Для расширения базы данных исследования, включающей информацию о технологических трендах выполняется сбор URL-адресов новостных источников. Этот этап реализуется посредством JavaScript-скрипта, предназначенного для автоматического сбора ссылок на новостные веб-сайты.

Принцип сбора данных с новостных ресурсов аналогичен процессу, применяемому для научных статей. Разработанный Python-скрипт использует библиотеки Selenium и BeautifulSoup для анализа структуры веб-страниц и извлечения необходимых данных из списка полученных URL-адресов. Скрипт автоматически собирает информацию о категории новости, времени публикации, заголовке и полном тексте новостной статьи. Полученные структурированные данные затем агрегируются и сохраняются в CSV-файл для дальнейшей обработки и анализа (рис. 3).

	category	publication_time	title	text	url
0	Город	2024-11-19T10:30:05+03:00	Ракова объяснила внедрение искусственного интеллек...	Столичные власти активно внедряют цифровизацию в о...	https://www.rbc.ru/rbcfreenews/673c37ea9a7947bc84c...
1	Технологии и медиа	2024-11-19T10:21:15+03:00	Ученые назвали год максимума солнечной активности ...	Максимальная активность Солнца в текущем цикле наи...	https://www.rbc.ru/rbcfreenews/673c34559a794763747...
2	Технологии и медиа	2024-11-19T07:40:41+03:00	Bloomberg узнал, что Минюст США потребует от Googl...	Высокопоставленные сотрудники антимонопольного упр...	https://www.rbc.ru/rbcfreenews/673c01989a7947e6bc2...
...	...	...	...	...	...
197	Как защититься от мошенников	2024-10-07T20:12:41+03:00	«Лаборатория Касперского» предложила сообщать об у...	Информировать россиян об утечках персональных данн...	https://www.rbc.ru/rbcfreenews/670414619a794795ac1...
198	Задержание Павла Дурова	2024-10-07T10:54:58+03:00	Сеул попросил Париж помочь с расследованием о дипф...	Полиция Сеула обратилась к властям Франции с прось...	https://www.rbc.ru/rbcfreenews/6703896c9a79477da05...
199	Технологии и медиа	2024-10-07T10:47:50+03:00	ВГТРК сообщила о «беспрецедентной» хакерской атаке...	В ночь на 7 октября онлайн-сервисы ВГТРК подвергли...	https://www.rbc.ru/rbcfreenews/670391c89a794705c58...

Рис. 3. Датасет по собранным новостям

Датасет имеет следующую структуру:

- category – категория новости;
- publication_time – дата и время публикации новости;
- title – название новости;
- text – содержание новости;
- url – ссылка на публикацию.

В итоге получено 200 новостных источников, связанных с развитием технологий в России.

Сбор данных из патентов. Для проведения анализа технологических трендов в качестве источника патентной информации используется датасет, который состоит из 89 отдельных файлов. Каждый файл содержит данные о патентах, относящихся к определенной технологической области.

В целях проведения интеллектуального анализа и выявления ключевых технологических направлений, данные патентов объединяются в единый структурированный датасет, представленный на рис. 4.

Анализ объединенного датасета патентов демонстрирует недостаточность текстовых данных для проведения полноценного анализа. В связи с этим, на текущем этапе принято решение о дополнении датасета аннотациями к патентам. Для этого разработан процесс автоматизированного сбора описаний патентов посредством парсинга URL-ссылок, содержащихся в исходном датасете. На первом этапе извлекаются все доступные

ссылки на патентные описания. Затем эти ссылки передаются специально разработанному парсеру, который использует библиотеки Selenium и BeautifulSoup для извлечения аннотаций из веб-страниц.

	id	title	assignee	inventor/author	priority date	filing/creation date	publication date	grant date	result link	representative figure link
0	RU-2486412-C1	Отопительная...	Данфосс А/С	Ян Эрик ТОРС...	2010-11-10	2011-11-09	2013-06-27	2013-06-27	https://patents...	https://patenti...
1	RU-105973-U1	Односекцион...	Керми Гмбх	Роджер ШЕНБ...	2007-07-31	2007-10-22	2011-06-27	2011-06-27	https://patents...	nan
2	RU-2719170-C2	Устройство от...	Киунгдонг На...	Чанг Хеой ХЕ...	2015-06-22	2016-05-04	2020-04-17	2020-04-17	https://patents...	https://patenti...
...	...	...	...	...	...	...	...	...	...	...
67456	RU-2557151-C2	Аппарат охла...	Л'Эр Ликид, С...	Патрис КАВАНЬ	2010-07-09	2011-06-22	2015-07-20	2015-07-20	https://patents...	nan
67457	RU-2530898-C2	Способ перер...	Ге Йенбахер Г...	Франц ПОКШ...	2009-10-02	2010-10-01	2014-10-20	2014-10-20	https://patents...	nan
67458	RU-2766594-C1	Установка для...	Общество С О...	Денис Алекса...	2020-12-22	2020-12-22	2022-03-15	2022-03-15	https://patents...	https://patenti...

Рис. 4. Датасет по собранным патентам

В результате собрано 67458 аннотаций к патентам. После завершения парсинга аннотации объединяются с исходным датасетом в единый структурированный датасет. Содержание результирующего датасета представлено на рисунке 5.

Структура датасета:

- id – уникальный номер патента;
- title – название патента;
- assignee – правообладатель;
- inventor/author – авторы патента;
- prioritydate – дата подачи заявки на патент;
- filing/creationdate – дата создания записи о патенте;
- publicationdate – дата удовлетворения заявки;
- grantdate – дата публикации патента;
- resultlink – ссылка на патент;
- representativefigurelink – ссылка на репрезентативный рисунок;
- abstact – аннотация патента.

	id	title	assignee	inventor/author	priority date	filing/creation date	publication date	grant date	result link	representative figure link	abstract
0	RU-2486412-...	Отопительн...	Данфосс А/С	Ян Эрик ТОР...	2010-11-10	2011-11-09	2013-06-27	2013-06-27	https://paten...	https://paten...	Данное изоб...
1	RU-105973-U1	Односекцио...	Керми Гмбх	Роджер ШЕ...	2007-07-31	2007-10-22	2011-06-27	2011-06-27	https://paten...	nan	1. По меньш...
2	RU-2719170-...	Устройство ...	Киунгдонг Н...	Чанг Хеой Х...	2015-06-22	2016-05-04	2020-04-17	2020-04-17	https://paten...	https://paten...	Изобретени...
...	...	...	...	...	...	...	...	...	...	...	...
67456	RU-2557151-...	Аппарат охл...	Л'Эр Ликид, ...	Патрис КАВ...	2010-07-09	2011-06-22	2015-07-20	2015-07-20	https://paten...	nan	Изобретени...
67457	RU-2530898-...	Способ пере...	Ге Йенбахер...	Франц ПОК...	2009-10-02	2010-10-01	2014-10-20	2014-10-20	https://paten...	nan	Изобретени...
67458	RU-2766594-...	Установка д...	Общество С ...	Денис Алекс...	2020-12-22	2020-12-22	2022-03-15	2022-03-15	https://paten...	https://paten...	Изобретени...

Рис. 5. Единый датасет по патентам с аннотациями

Обработка сформированного датасета и выявление ключевых технологических терминов. Для последующего анализа и выявления ключевых технологических терминов все текстовые данные, извлеченные из различных источников, на текущем этапе подвергаются ряду процедур обработки.

Последовательность действий метода анализа и выявления ключевых технологических терминов представлена на DFD-диаграмме ниже (рис. 6).

На первом этапе из каждого датасета извлекаются текстовые данные:

- для статей: title, abstract, keywords;
- для новостей: title, text;
- для патентов: title, abstract.

Затем (этап 2) все текстовые данные объединяются в один текстовый корпус (corpus) для анализа.

На третьем этапе текст разбивается на слова с использованием библиотеки NLTK (RegexpTokenizer). Убираются стоп-слова (предлоги, союзы и т.п.), а также знаки пунктуации. Также выполняется лемматизация текста, т.е. приведение слов в начальную форму. Это необходимо для корректного выявления ключевых терминов.

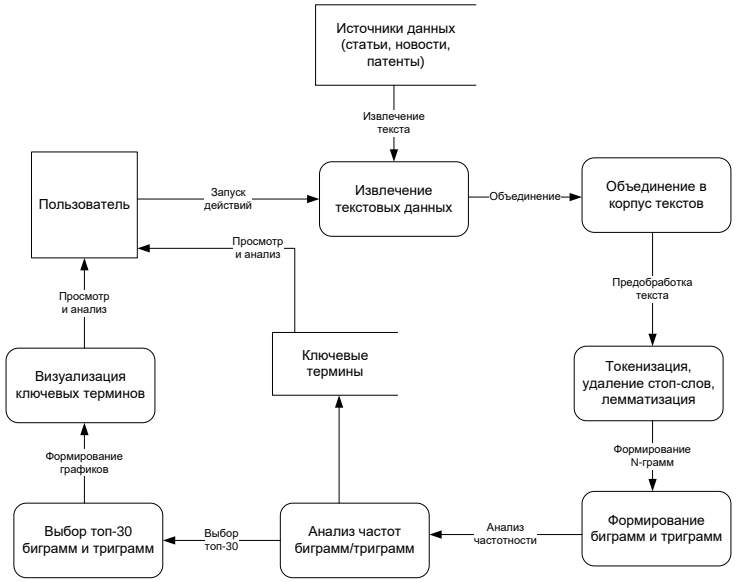

Рис. 6. Диаграмма потоков данных метода анализа и выявления ключевых технологических терминов

На этапе формирования биграмм и триграмм анализ реализован с помощью методов из библиотеки Scikit-learn, а точнее её модуля CountVectorizer из пакета sklearn.feature_extraction.text. Для каждой биграммы/триграммы рассчитана частота появления ключевого термина.

На основании частотности извлекаются 30 наиболее часто встречающихся биграмм и 30 триграмм.

На финальном этапе генерируются горизонтальные гистограммы с биграммами/триграммами и их частотами (рис. 7, 8).

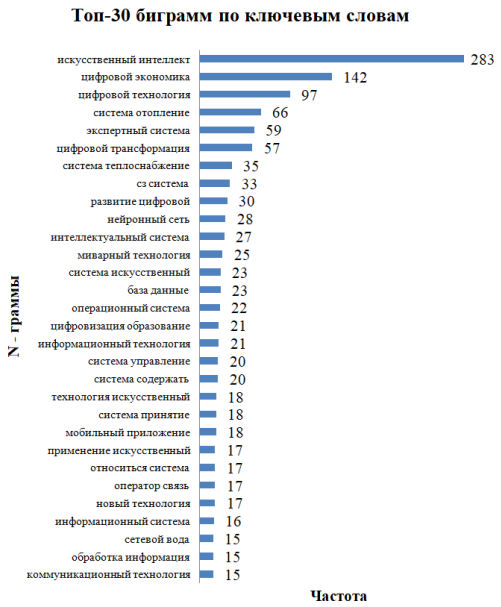

Рис. 7. График биграмм по ключевым терминам

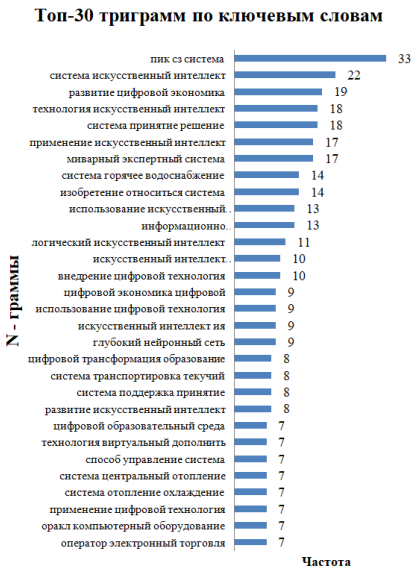

Рис. 8. График триграмм по ключевым терминам

Анализ частотности биграмм и триграмм позволил выявить доминирующие термины. Представленные гистограммы (рис. 7, 8) позволяют идентифицировать ключевые технологии и выявить наиболее значимые сочетания слов, отражающие основные концепции и направления в технологических областях.

Данная информация послужит основой для дальнейшего анализа и формирования перечня технологий, которые будут использованы для фильтрации патентной информации.

Формирование временных рядов ключевых терминов и анализ патентной активности. На основе полученных данных биграмм и триграмм идентифицирован ряд ключевых технологий. Следующим шагом исследования стала разработка метода идентификации ключевых технологий (МИКТ) на основе анализа временных рядов патентной активности (рис. 9).

Формирование временных рядов ключевых терминов. В связи с тем, что на предварительном этапе была выполнена лемматизация слов для идентификации ключевых терминов, в дальнейшем они приводятся к стандартным формам. Технологии, представленные в виде биграмм и триграмм, приведены в табл. 1, в которой также указано количество найденных патентов для каждой технологии.

В табл. 1 наблюдается существенное различие в количестве найденных патентов, содержащих биграммы и триграммы. Это объясняется тем, что многие триграммы, идентифицированные на этапе анализа частотности в текстах научных публикаций, непосредственно не упоминаются в патентной документации, связанной с ключевыми технологиями.

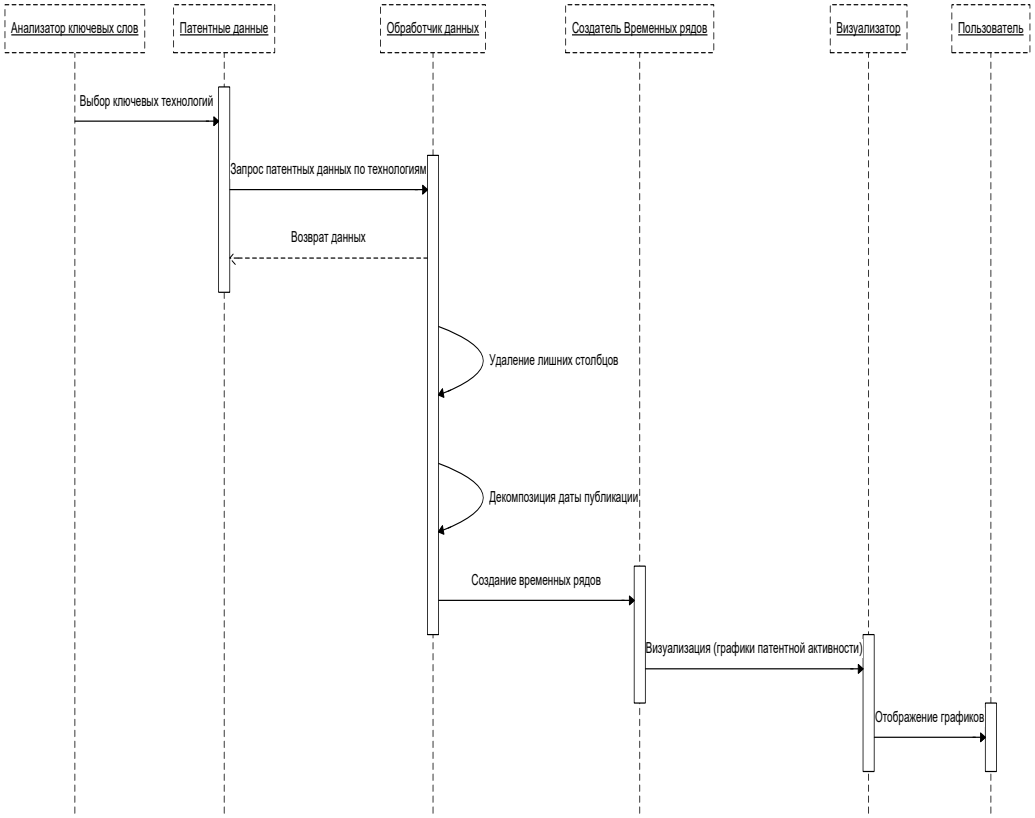

Рис. 9. Диаграмма последовательности метода идентификации ключевых технологий на основе анализа временных рядов патентной активности

Таблица 1

Биграммы и триграммы ключевых технологий и количество найденных патентов

Биграммы	Число патентов	Триграммы	Число патентов
искусственный интеллект	397	системы искусственного интеллекта	1376
цифровая экономика	28	система принятия решений	5755
цифровые технологии	8466	система горячего водоснабжения	1412
система отопления	4383	информационно коммуникационные технологии	87
экспертная система	604	логический искусственный интеллект	51
цифровая трансформация	75	глубокие нейронные сети	104
система теплоснабжения	1780	поддержка принятия решений	422
нейронные сети	1417	система текучей транспортировки	2787
интеллектуальная система	1306	цифровая трансформация образования	36
базы данных	17160	оператор электронной торговли	174
искусственные технологии	2907	система отопления/охлаждения	1520
операционная система	4517	система центрального отопления	536
информационные технологии	5246	технология виртуальной реальности	590
цифровизация образования	16		
система управления	17321		
искусственные системы	5683		
коммуникационные технологии	1234		
обработка информации	17068		
компьютерное оборудование	1366		

По указанным выше технологиям анализируется патентная активность. В рамках МИКТ была применена комплексная фильтрация по наименованию ключевых технологий для поиска русскоязычных патентов и их сбор для дальнейшего анализа. Последую-

щая обработка включает удаление избыточных столбцов из полученных датасетов с сохранением только класса технологии и даты публикации патента. Для анализа временных рядов дата публикации патента декомпозируется на отдельные поля, представляющие месяц и год публикации (табл. 2).

Таблица 2

Итоговый формат датасета временных рядов ключевых технологий (фрагмент)

Название технологии	Год	Месяц	Количество публикаций
Искусственный интеллект	2021	11	5
Искусственный интеллект	2021	12	7
…			
Нейронные сети	2022	6	11

Для демонстрации процесса формирования датасета временных рядов используется пример технологии «Искусственный интеллект». Для остальных датасетов применяется аналогичный принцип, отличие заключается лишь в названии ключевой технологии.

В рамках обработки удаляются все колонки, кроме даты публикации патента, и в первой колонке устанавливается название ключевой технологии. Также дата публикации разделяется на колонки year (год публикации патента) и month (месяц публикации патента). Наконец, добавляется колонка count_publication, отражающая количество патентов за указанный временной период. Месяцы и годы должны следовать последовательно, что означает, что если в каком-либо месяце не было опубликовано ни одного патента, то для него устанавливается значение 0. Результат обработки представлен на рис. 10.

	technology	year	month	count_publication
0	Искусственный интеллект	1980	1	0
1	Искусственный интеллект	1980	2	0
2	Искусственный интеллект	1980	3	0
...	...	...	...	...
537	Искусственный интеллект	2024	10	6
538	Искусственный интеллект	2024	11	2
539	Искусственный интеллект	2024	12	0

Рис. 10. Данные временных рядов технологии «Искусственный интеллект»

Анализ патентной активности. На этапе анализа патентной активности выявлены тенденции развития (подъем Rise или спад Fall) ключевых технологий, которые были отфильтрованы и представлены в виде временных рядов (рис. 11, 12).

Визуализация патентной активности на представленных графиках (рис. 11, 12.) позволяет оценить динамику развития каждой идентифицированной технологии. Например, патентование в области систем управления и обработки информации достигло максимальной активности в 2008 году, после чего последовал значительный спад, что свидетельствует о стабилизации или смене технологических приоритетов. В то же время патентование в областях систем искусственного интеллекта и глубоких нейронных сетей демонстрирует активное развитие в последние годы. Это подтверждает тенденцию на расширение сфер применения методов искусственного интеллекта.

Наблюдаемые пики на графиках отражают периоды интенсивной патентной деятельности, свидетельствующие о повышенном интересе к данной технологии в этот временной промежуток. Эти всплески могут быть связаны с прорывными открытиями, появлением новых применений технологии или же с общей активизацией инновационной деятельности в определенной области.

В свою очередь, падения на графиках могут сигнализировать о снижении интереса к технологии, возможно, в связи с насыщением рынка, появлением более перспективных альтернатив или же в результате смены технологических парадигм. Анализ продолжительности и интенсивности этих пиков и падений позволяет выявить жизненный цикл технологии, определить ее текущее состояние и потенциальные перспективы развития.

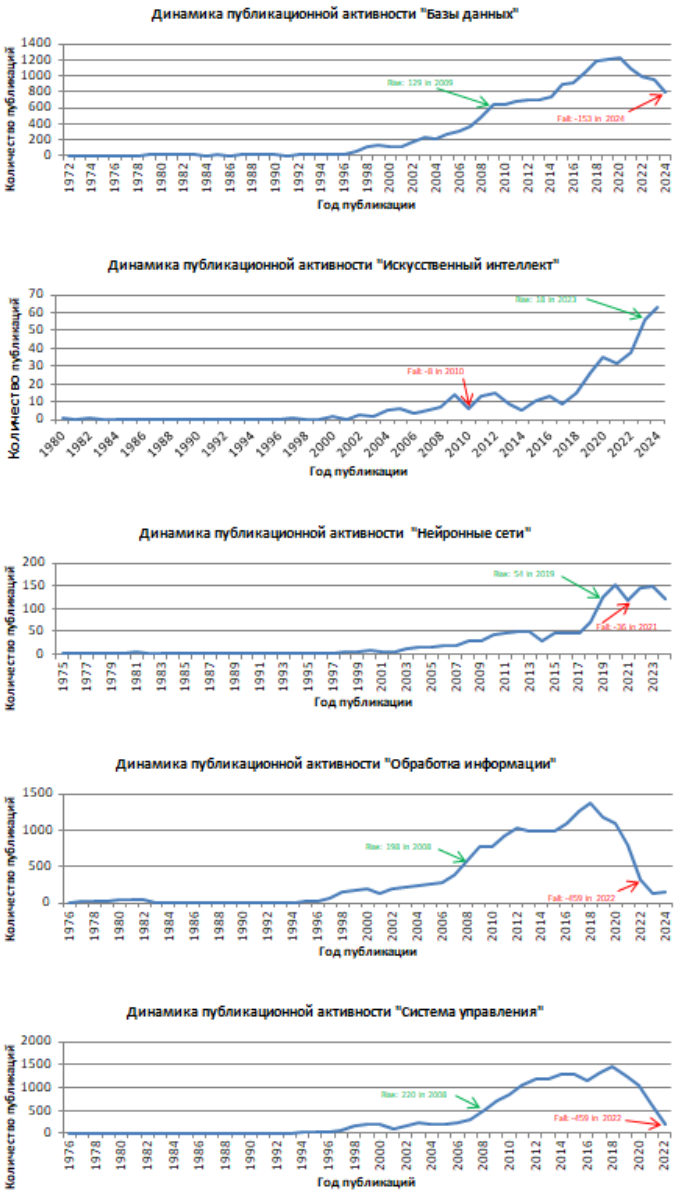

Рис. 11. Графики патентной активности исследуемых технологий-биграмм (фрагмент)

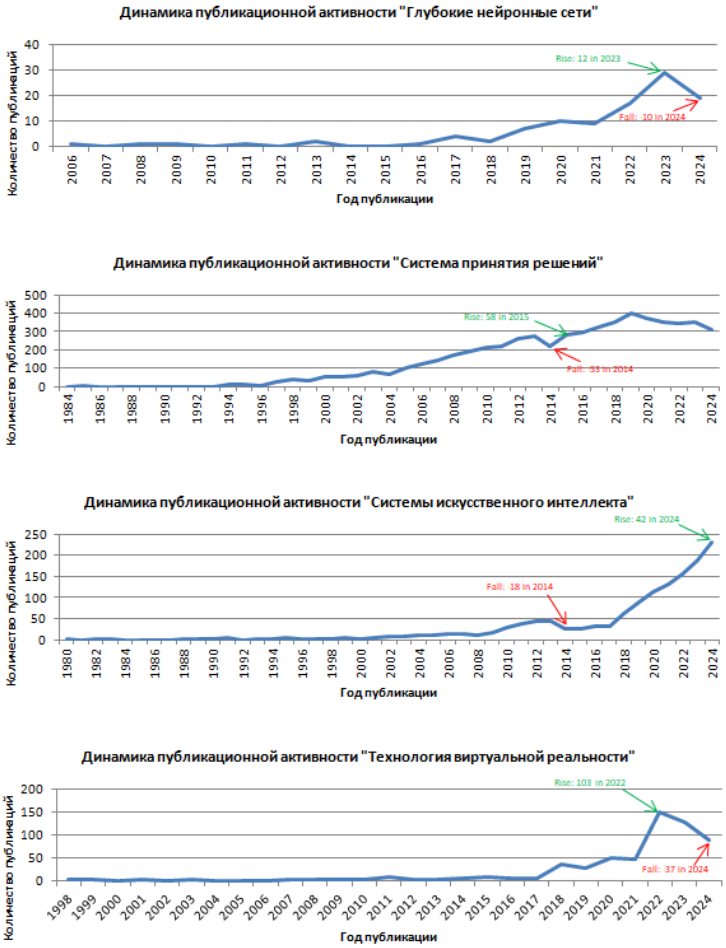

Рис. 12. Графики патентной активности исследуемых технологий-триграмм (фрагмент)

Итоговые результаты, представленные в табл. 3, подтверждают эффективность предложенного подхода к идентификации ключевых технологий.

Таблица 3

Результаты исследования

Технология	Частота встречаемости в текстах	Количество патентов	Год начала активности технологий	Пик активности / прирост числа публикаций	Спад активности/ убыль числа публикаций
Искусственный интеллект	283	397	1980 г.	2023 г. /18	2010 г. /-8
Нейронные сети	28	1417	1975 г.	2019 г. /54	2021 г. /-36
Базы данных	23	17160	1972 г.	2009 г. /149	2024 г. /-153
Система управления	20	17321	1976 г.	2008 г. /220	2022 г. /-459

Окончание табл. 3

Технология	Частота встречаемости в текстах	Количество патентов	Год начала активности технологий	Пик активности / прирост числа публикаций	Спад активности/ убыль числа публикаций
Обработка информации	15	17068	1976 г.	2008 г. /198	2022 г. /-459
Системы искусственного интеллекта	22	1376	1980 г.	2024 г. /42	2014 г. /-18
Система принятия решений	18	5755	1984 г.	2015 г. /58	2014 г. /-53
Глубокие нейронные сети	9	104	2006 г.	2023 г. /12	2024 г. /-10
Технология виртуальной реальности	7	590	1998 г.	2022 г. /108	2024 г. /-37

Заключение. В рамках проведенного исследования был разработан и реализован комплексный подход к сбору, обработке и анализу открытых данных с целью идентификации ключевых технологий. Разработка метода веб-скрейпинга, использование методов обработки естественного языка и анализа временных рядов позволило сформировать структурированные датасеты. На основе анализа частотности биграмм и триграмм были выделены ключевые технологические термины, которые в дальнейшем легли в основу для МИКТ. В рамках исследования проанализированы исключительно русскоязычные документы, что позволяет учитывать специфику отечественного технологического развития.

Выполненный с помощью МИКТ анализ временных рядов патентной активности позволил визуализировать динамику развития каждой исследуемой технологии, выявить периоды интенсивной патентной деятельности и снижения интереса к ним. На основе полученного датасета были сформированы графики патентной активности. Это является важным шагом для дальнейшего анализа и кластеризации, а также для прогнозирования развития ключевых технологий.

Анализ частоты упоминания технологических терминов в текстах и их патентной активности позволил выявить динамику развития различных технологий, определить периоды их интенсивного роста и спада.

В результате проведенного исследования были созданы необходимые основы для дальнейшего применения методов машинного обучения с целью прогнозирования технологического развития, что является следующим шагом в исследовании и выходит за рамки данной статьи.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Безруков А.О., Байдаров Д.Ю., Файков Д.Ю.* Технологическое лидерство государства: концептуальное понимание и механизмы формирования // Экономическое возрождение России. – 2024. – № 1 (79). – С. 75-89. – DOI 10.37930/1990-9780-2024-1-79-75-89. – EDN ZSGGUU.
2. *Елисеев В.А.* Доминанты прогнозирования научно-технологического развития // Автоматизация. Современные технологии. – 2019. – Т. 73, № 10. – С. 461-466.
3. *Бондаренко А.Г., Кравец А.Г.* Инструменты прогнозирования технологического развития на основе данных из открытых источников: систематическое исследование русскоязычных документов // Прикаспийский журнал: управление и высокие технологии. – 2024. – № 3 (67). – С. 49-62.

4. *Porter A.L. et al.* Emergence scoring to identify frontier R&D topics and key players // Technol. Forecast. Soc. Change. 2019. – Vol. 146. – P. 628-643. – DOI: 10.1016/j.techfore.2018.04.016.
5. *Кравец А.Г., Нгуен Т.В.* Прогнозирование технологических тенденций на основе анализа разнородных данных // Программные продукты и системы. – 2022. – № 3. – С. 396-412. – DOI: 10.15827/0236-235X.139.396-412.
6. *Nivash J.P., Babu L.D.D.* Analyzing the impact of news trends on research publications and scientific collaboration networks // Concurrency and Computation-Practice & Experience. – 2019. – Vol. 31, No. 14. – P. 10.
7. *Injadat M.N., Salo F., Nassif A.B.* Data mining techniques in social media: A survey. Neurocomputing. – 2016. – Vol. 214. – P. 654-670. – DOI: 10.1016/j.neucom.2016.06.045.
8. *Antons D. et al.* The application of text mining methods in innovation research: current state, evolution patterns, and development priorities // R & D Management. – 2020. – P. 329-351. – DOI: 10.1111/radm.12408.
9. *Zhou Y. et al.* Forecasting emerging technologies using data augmentation and deep learning // Scientometrics. – 2020. – DOI: 10.1007/s11192-020-03351-6.
10. *Каленов Н.Е., Власова С.А.* О реализации многофункциональной web-системы регистрации и учета результатов интеллектуальной деятельности ученых // Программные продукты и системы. – 2021. – № 4. – С. 501-510. – DOI: 10.15827/0236-235X.136.501-510.
11. *Сотников А.Н., Каленов Н.Е., Власова С.А.* Развитие системы «Экспертиза» как инструмента для формирования энциклопедий и наполнения Единого цифрового пространства научных знаний // Программные продукты и системы. – 2022. – № 4. – С. 541-548. – DOI: 10.15827/0236-235X.140.541-548.
12. *Vasiliev S.S., Korobkin D.M., Kravets A.G. et al.* Extraction of Cyber-Physical Systems Inventions' Structural Elements of Russian-Language Patents // Studies in Systems, Decision and Control. – 2020. – Vol. 259. – P. 55-68. – DOI: 10.1007/978-3-030-32579-4_5.
13. *Song K., Kim K., Lee S.* Identifying promising technologies using patents: A retrospective feature analysis and a prospective needs analysis on outlier patents // Technol. Forecast. Soc. Change. – 2018. – DOI: 10.1016/j.techfore.2017.11.008.
14. *Коробкин Д.М., Рублев А.А., Фоменков С.А.* Прогнозирование значимости запатентованных технологий на основе метрик инновационного потенциала // Программная инженерия. – 2024. – Т. 15, № 5. – С. 243-253. – DOI: 10.17587/prin.15.243-253.
15. *Lee C., Kwon O., Kim M., Kwon D.* Early identification of emerging technologies: A machine learning approach using multiple patent indicators // Technol. Forecast. Soc. Change. – 2018. – DOI: 10.1016/j.techfore.2017.10.002.
16. *Yu J. et al.* Identification of vacant and emerging technologies in smart mobility through the GTM-based patent map development // Sustain. – 2020. – DOI: 10.3390/su12229310.
17. *Jun S. et al.* Identification of promising vacant technologies for the development of truck on freight train transportation systems // Appl. Sci. – 2021. – DOI: 10.3390/app11020499.
18. *Yoon B., Park I., Yun D., Park, G.* Exploring promising vacant technology areas in a technology-oriented company based on bibliometric analysis and visualization // Technol. Anal. Strateg. Manag. – 2019. – DOI: 10.1080/09537325.2018.1516864.
19. *Белевцев А.А., Белевцев А.М., Балыбердин В.А.* Методика прогнозирования развития технологических трендов и построения дорожных карт на основе конструирования будущих событий // Известия ЮФУ. Технические науки. – 2023. – № 3(233). – С. 56-64. – DOI: 10.18522/2311-3103-2023-3-56-64.
20. *Вьет Н.Т., Кравец А.Г.* Алгоритм работы веб-краулера для решения задачи сбора данных из открытых интернет источников // Известия Санкт-Петербургского государственного технологического института (технического университета). – 2019. – № 51 (77). – С. 115-119. – DOI: 10.36807/1998-9849-2019-51-77-115-119.
21. *Козина С.А., Кулинченко И.А., Коробкин Д.М., Фоменков С.А.* Концепция и архитектура парсинга и хранения единой базы патентов и научных журнальных публикаций // Моделирование, оптимизация и информационные технологии. – 2024. – Т. 12, № 4. – 15 с. – DOI: 10.26102/2310-6018/2024.47.4.024. – URL: https://moitvivt.ru/ru/journal/pdf?id=1740.

REFERENCES

1. *Bezrukov A.O., Baydarov D.Yu., Faykov D.Yu.* Tekhnologicheskoye liderstvo gosudarstva: kontseptual'noye ponimaniye i mekhanizmy formirovaniya [Technological leadership of the state: conceptual understanding and mechanisms of formation], *Ekonomicheskoye vozrozhdenie Rossii* [Economic Revival of Russia], 2024, No. 1 (79), pp. 75-89. DOI: 10.37930/1990-9780-2024-1-79-75-89.

2. *Eliseev V.A.* Dominanty prognozirovaniya nauchno-tekhnologicheskogo razvitiya [Dominants of forecasting scientific and technological development], *Avtomatizatsiya. Sovremennye tekhnologii* [Automation. Modern Technologies], 2019, Vol. 73, No. 10, pp. 461-466.
3. *Bondarenko A.G., Kravets A.G.* Instrumenty prognozirovaniya tekhnologicheskogo razvitiya na osnove dannykh iz otkrytykh istochnikov: sistematicheskoye issledovaniye russkoyazychnykh dokumentov [Tools for forecasting technological development based on data from open sources: a systematic study of Russian-language documents], *Prikaspiyskiy zhurnal: upravlenie i vysokie tekhnologii* [Caspian Journal: Management and High Technologies], 2024, No. 3 (67), pp. 49-62.
4. *Porter A.L. et al.* Emergence scoring to identify frontier R&D topics and key players, *Technol. Forecast. Soc. Change*, 2019, Vol. 146, pp. 628-643. DOI: 10.1016/j.techfore.2018.04.016.
5. *Kravets A.G., Nguyen T.V.* Prognozirovaniye tekhnologicheskikh tendentsiy na osnove analiza raznorodnykh dannykh [Forecasting technological trends based on the analysis of heterogeneous data], *Programmnye produkty i sistemy* [Software & Systems], 2022, No. 3, pp. 396-412. DOI: 10.15827/0236-235X.139.396-412.
6. *Nivash J.P., Babu L.D.D.* Analyzing the impact of news trends on research publications and scientific collaboration networks, *Concurrency and Computation-Practice & Experience*, 2019, Vol. 31, No. 14, pp. 10.
7. *Injadat M.N., Salo F., Nassif A.B.* Data mining techniques in social media: A survey, *Neurocomputing,* 2016, Vol. 214, pp. 654-670. DOI: 10.1016/j.neucom.2016.06.045.
8. *Antons D. et al.* The application of text mining methods in innovation research: current state, evolution patterns, and development priorities, *R & D Management*, 2020, pp. 329-351. DOI: 10.1111/radm.12408.
9. *Zhou Y. et al.* Forecasting emerging technologies using data augmentation and deep learning, *Scientometrics,* 2020. DOI: 10.1007/s11192-020-03351-6.
10. *Kalenov N.E., Vlasova S.A.* O realizatsii mnogofunktsional'noy web-sistemy registratsii i ucheta rezul'tatov intellektual'noy deyatel'nosti uchenykh [On the implementation of a multifunctional web-system for registration and accounting of the results of intellectual activity of scientists], *Programmnye produkty i sistemy* [Software & Systems], 2021, No. 4, pp. 501-510. DOI: 10.15827/0236-235X.136.501-510.
11. *Sotnikov A.N., Kalenov N.E., Vlasova S.A.* Razvitiye sistemy «Ekspertiza» kak instrumenta dlya formirovaniya entsiklopediy i napolneniya Edinogo tsifrovogo prostranstva nauchnykh znaniy [Development of the "Expertise" system as a tool for the formation of encyclopedias and filling the Unified Digital Space of Scientific Knowledge], *Programmnye produkty i sistemy* [Software & Systems], 2022, No. 4, pp. 541-548. DOI: 10.15827/0236-235X.140.541-548.
12. *Vasiliev S.S., Korobkin D.M., Kravets A.G. et al.* Extraction of Cyber-Physical Systems Inventions' Structural Elements of Russian-Language Patents, *Studies in Systems, Decision and Control*, 2020, Vol. 259, pp. 55-68. DOI: 10.1007/978-3-030-32579-4_5.
13. *Song K., Kim K., Lee S.* Identifying promising technologies using patents: A retrospective feature analysis and a prospective needs analysis on outlier patents, *Technol. Forecast. Soc. Change,* 2018. DOI: 10.1016/j.techfore.2017.11.008.
14. *Korobkin D.M., Rublev A.A., Fomenkov S.A.* Prognozirovanie znachimosti zapatenovannykh tekhnologiy na osnove metrik innovatsionnogo potentsiala [Forecasting the significance of patented technologies based on metrics of innovative potential], *Programmnaya Inzheneriya* [Software Engineering], 2024, Vol. 15, No. 5, pp. 243-253. DOI: 10.17587/prin.15.243-253.
15. *Lee C., Kwon O., Kim M., Kwon D.* Early identification of emerging technologies: A machine learning approach using multiple patent indicators, *Technol. Forecast. Soc. Change*, 2018. doi:10.1016/j.techfore.2017.10.002.
16. *Yu J. et al.* Identification of vacant and emerging technologies in smart mobility through the GTM-based patent map development, *Sustain,* 2020. DOI: 10.3390/su12229310.
17. *Jun S. et al.* Identification of promising vacant technologies for the development of truck on freight train transportation systems, *Appl. Sci.,* 2021. DOI: 10.3390/app11020499.
18. *Yoon B., Park I., Yun D., Park, G.* Exploring promising vacant technology areas in a technology-oriented company based on bibliometric analysis and visualisation, *Technol. Anal. Strateg.Manag.,* 2019. DOI: 10.1080/09537325.2018.1516864.
19. *Belevtsev A.A., Belevtsev A.M., Balyberdin V.A.* Metodika prognozirovaniya razvitiya tekhnologicheskikh trendov i postroyeniya dorozhnykh kart na osnove konstruirovaniya budushchikh sobytiy [Methodology for forecasting the development of technological trends and building roadmaps based on the construction of future events], *Izvestiya YuFU. Tekhnicheskiye Nauki* [Izvestiya SFedU. Engineering Sciences], 2023, No. 3(233), pp. 56-64. DOI: 10.18522/2311-3103-2023-3-56-64.

20. *Viet N.T., Kravets A.G.* Algoritm raboty web-kraulera dlya resheniya zadachi sbora dannykh iz otkrytykh internet istochnikov [The algorithm of the web crawler for solving the problem of collecting data from open Internet sources], *Izvestiya Sankt-Peterburgskogo gosudarstvennogo tekhnologicheskogo instituta (tekhnicheskogo universiteta)* [Izvestiya of Saint-Petersburg State Technological Institute (Technical University)], 2019, No. 51 (77), pp. 115-119. DOI: 10.36807/1998-9849-2019-51-77-115-119.
21. *Kozina S.A., Kulinchenko I.A., Korobkin D.M., Fomenkov S.A.* Kontseptsiya i arkhitektura parsinga i khraneniya edinoy bazy patentov i nauchnykh zhurnal'nykh publikatsiy [The concept and architecture of parsing and storing a unified database of patents and scientific journal publications], *Modelirovanie, optimizatsiya i informatsionnye tekhnologii* [Modeling, Optimization and Information Technology], 2024, Vol. 12, No. 4, 15 p. DOI: 10.26102/2310-6018/2024.47.4.024. Available at: https://moitvivt.ru/ru/journal/pdf?id=1740.

Бондаренко Артём Геннадьевич – Волгоградский государственный технический университет; e-mail: temdit01@yandex.ru; г. Волгоград, Россия; тел.: +79375596156; кафедра систем автоматизированного проектирования и поискового конструирования; магистрант.

Кравец Алла Григорьевна – Волгоградский государственный технический университет; e-mail: AllaGKravets@yandex.ru; г. Волгоград, Россия; тел.: +79023639186; кафедра систем автоматизированного проектирования и поискового конструирования; д.т.н.; профессор.

Bondarenko Artem Gennadevich – Volgograd State Technical University; e-mail: temdit01@yandex.ru; Volgograd, Russia; phone: +79375596156; the Department of CAD&RD; master student.

Kravets Alla Grigorievna – Volgograd State Technical University; e-mail: AllaGKravets@yandex.ru; Volgograd, Russia; phone: +79023639186; the Department of CAD&RD; dr. of eng. sc.; professor.

УДК 004.89 DOI 10.18522/2311-3103-2025-3-159-171

А.М. Мансур, Ж.Х. Мохаммад, Ю.А. Кравченко

РАЗРАБОТКА ЧАТ-БОТА ДЛЯ КЛАССИФИКАЦИИ И АНАЛИЗА ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ С ИСПОЛЬЗОВАНИЕМ ЛОКАЛЬНЫХ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

Исследуются локальные большие языковые модели (Local large language models, Local LLM) и их применение в задачах классификации текста, а также проводится сравнение их производительности с традиционными методами. Статья предоставляет всесторонний обзор ряда ключевых локальных LLM, уделяя особое внимание их архитектурным преимуществам, характеристикам и областям применения. В частности, рассматриваются модели с различным количеством параметров, их способность адаптироваться к специализированным доменам, а также требования к вычислительным ресурсам при их развертывании на локальном оборудовании. Особый акцент делается на компромиссах между производительностью и эффективностью использования ресурсов. В качестве практического вклада разработан чат-бот, использующий локальные LLM (такие как DeepSeek, Gemma и Llama2 через Ollama) для классификации входящих текстов по заранее заданным категориям, демонстрируя работу этих моделей без использования облачных вычислений. Система реализована с модульной архитектурой, позволяющей легко интегрировать новые модели и сравнивать их эффективность. Вычислительный эксперимент включает оценку точности и скорости вывода локальных LLM в сравнении с более простыми методами, такими как Sentence-BERT, TF-IDF и BoWC, выделяя сценарии, в которых локальные модели превосходят традиционные подходы или уступают им. Тестирование проводилось на основе эталонного набора данных BBC. Результаты показывают, что языковые модели (включая модели с 7 миллиардами параметров) демонстрируют сильную и логически обоснованную классификационную производительность при обработке текстов на естественном языке, однако их результаты не являются идеальными для эталонных наборов данных. В частности, обнаружены случаи, когда все тестируемые модели, включая традиционные методы, ошибочно классифицировали документы, что указывает на возможные проблемы в разметке данных. Полученные результаты указывают на необходимость пересмотра эталонных меток в стандартных наборах данных. Это особенно