МИР (Модернизация. Инновации. Развитие). 2024. Т. 15. № 3

ISSN 2079-4665, E-ISSN 2411-796X https://www.mir-nayka.com

Научная статья

УДК 303.6 + 004.6

JEL: C83, D8, C55

https://doi.org/10.18184/2079-4665.2024.15.3.404-420

Методология извлечения нарративов из больших массивов данных социальных сетей

Петров Евгений Юрьевич [1], Саркисова Анна Юрьевна [2], Дунаева Дарья Олеговна [3], Воронов Александр Сергеевич [4], Мягков Михаил Георгиевич [5]

[1] Национальный исследовательский Томский государственный университет; Томск, Россия

[2-5] Московский государственный университет им. М. В. Ломоносова; Москва, Россия

[1] petrov@data.tsu.ru, https://orcid.org/0000-0002-7140-7882

[2] ovanju@gmail.com, http://orcid.org/0000-0001-5674-0962

[3] darya.dunaewa@gmail.com, http://orcid.org/0000-0002-6622-9882

[4] voronov@spa.msu.ru, http://orcid.org/0000-0003-0058-9217

[5] myagkov@skoltech.ru, http://orcid.org/0000-0002-8419-6404

Аннотация

Цель статьи – представить опыт разработки и апробации методологии извлечения системы нарративов о социально значимом событии из больших массивов аутентичных данных социальных сетей (на примере нарративов о вакцинации от COVID-19 в публикациях пользователей российской социальной сети «ВКонтакте» периода пандемии).

Методы. Использовались методы автоматизированного анализа данных с применением инструментов аналитической платформы PolyAnalyst: тематическое моделирование (методом PLSA), алгоритмы индексирования текста с этапом идентификации предложений, кластеризация, агрегация данных, нормализация данных, расчет количественного индекса («индекса популярности»). Осуществлялись расчет меры близости ключевых слов с использованием языка программирования Python, частичная ручная разметка и валидация данных.

Результаты работы. 4,5 миллиона сообщений, релевантных теме вакцинации от COVID-19, опубликованных пользователями «ВКонтакте» за период с 01.01.2020 по 01.03.2023, сведены к 237-ми устойчивым нарративам. Для каждого нарратива был рассчитан индекс популярности. Наиболее популярным, например, оказался следующий нарратив: «Работодатели оказывают давление, принуждая вакцинироваться» (его поддержка – 76118 текстов). В результате исследования получен датасет, включающий 237 нарративов, содержательный анализ которого не является предметом настоящей статьи и планируется авторами в ближайшей перспективе. Датасет демонстрирует полноту охвата тематики отношения к вакцинации.

Выводы. Разработанный инструментарий имеет универсальный характер: методология может быть адаптирована под любую актуальную тематику, требуя только корректировки входных параметров тематического моделирования. Полученный датасет планируется ввести в научный оборот в качестве актуального материала для изучения общественного мнения о вакцинации в России. С учетом глобального значения пандемии и вакцинационных мероприятий, результаты вносят вклад в международные исследования по теме общественного мнения и коммуникации в условиях кризисов, могут служить основой для дальнейших исследований и практических действий, направленных на улучшение качества общественных коммуникаций и принятия решений на всех уровнях управления.

Ключевые слова: нарратив, автоматизированное извлечение нарративов, тематическое моделирование, PolyAnalyst, социальные сети, общественное мнение о вакцинации, большие данные

Благодарность. Исследование выполнено при финансовой поддержке РНФ, проект 23-28-01025 «Исследование нарративов в социальных медиа с применением технологии анализа больших данных (на примере нарративов о вакцинации от COVID-19)».

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Для цитирования: *Петров Е. Ю., Саркисова А. Ю., Дунаева Д. О., Воронов А. С., Мягков М. Г.* Методология извлечения нарративов из больших массивов данных социальных сетей // МИР (Модернизация. Инновации. Развитие). 2024. Т. 15. № 3. С. 404–420

EDN: https://elibrary.ru/mknigm. https://doi.org/10.18184/2079-4665.2024.15.3.404-420

Original article

Methodology for extracting narratives from social media big data

Evgeny Yu. Petrov [1], Anna Yu. Sarkisova [2], Daria O. Dunaeva [3], Aleksandr S. Voronov [4], Mikhail G. Myagkov [5]

[1] National Research Tomsk State University, Tomsk, Russia

[2-5] Lomonosov Moscow State University, Moscow, Russia

[1] petrov@data.tsu.ru, https://orcid.org/0000-0002-7140-7882

[2] ovanju@gmail.com, http://orcid.org/0000-0001-5674-0962

[3] darya.dunaewa@gmail.com, http://orcid.org/0000-0002-6622-9882

[4] voronov@spa.msu.ru, http://orcid.org/0000-0003-0058-9217

[5] myagkov@skoltech.ru, http://orcid.org/0000-0002-8419-6404

Abstract

Purpose: of the article is to present the experience in developing and testing the methodology for extracting a system of narratives on a socially significant phenomenon from authentic social network big data (using the example of narratives about COVID-19 vaccination in the Russian social network VKontakte during the pandemic).

Methods: of automated data analysis were used by the tools of the PolyAnalyst analytical platform: topic modeling (PLSA method), text indexing algorithms with the sentence identification stage, clustering, data aggregation, data normalization, calculation of a quantitative index. The calculation of the measure of proximity of keywords using the Python, partial manual markup and data validation were also carried out.

Results: 4.5 million messages relevant to the topic of COVID-19 vaccination published in VKontakte from 01.01.2020 to 01.03.2023 were reduced to 237 stable narratives. A popularity index was calculated for each narrative. For example, the following narrative turned out to be the most popular: "Employers put pressure on people to get vaccinated" (it was supported by 76,118 texts). As a result of the study, a dataset was obtained, including 237 narratives.

Conclusions and Relevance: the developed toolkit is universal: the methodology can be adapted to any relevant topic, requiring only adjustments to the input parameters of thematic modeling. The obtained dataset is planned to be introduced into scientific circulation as an up-to-date material for studying public opinion on vaccination in Russia. The results contribute to international research on public opinion and communication in crises and can serve as a basis for practical actions aimed at improving the quality of public communications and decision-making at all levels of government.

Keywords: narrative, automated narrative mining, topic modeling, PolyAnalyst, social media, public opinion on vaccination, big data

Acknowledgments. The study was carried out with the financial support of the Russian Science Foundation, project 23-28-01025 "Study of narratives in social media using big data analysis technology (using narratives about COVID-19 vaccination as an example)".

Conflict of Interest. The authors declare that there is no Conflict of Interest.

For citation: Petrov E. Yu., Sarkisova A. Yu., Dunaeva D. O., Voronov A. S., Myagkov M. G. Methodology for extracting narratives from social media big data. *MIR (Modernizatsiia. Innovatsii. Razvitie) = MIR (Modernization. Innovation. Research)*. 2024; 15(3):404–420. (In Russ.)

EDN: https://elibrary.ru/mknigm. https://doi.org/10.18184/2079-4665.2024.15.3.404-420

Введение

Социальные сети сегодня представляют собой, возможно, главную площадку массовой социальной коммуникации, являясь одновременно как зеркалом общественного мнения по огромному количеству социально значимых вопросов и источником данных для принятия управленческих решений, так и ресурсом для разного рода информационных атак, лоббирования интересов групп, манипулятивного воздействия на население. Большие массивы данных и отсутствие в соцсетях поисковых систем затрудняют анализ информации и актуализируют разработку методов, направленных на преодоление фрагментарности пользовательских текстовых данных.

Одной из перспективных задач анализа контента социальных сетей представляется задача автоматизированного поиска и извлечения нарративов, тематически связанных с определенным «событием». Обилие и хаотичность информации в условиях современного этапа развития информационно-коммуникационных технологий диктуют необходимость поиска инструментальных средств обобщения и систематизации данных. Нарратив как повествовательная структура, концентрирующая антропоцентрический опыт и формирующая объяснительные модели процессов и явлений с субъективной точки зрения, является инструментом конструирования версии события, его причин и следствий.

Цель настоящего исследования – разработать и описать методологию автоматизированного извлечения нарративов о вакцинации от COVID-19, представленных в российской социальной сети «ВКонтакте» в период пандемии.

Разработка методологии извлечения системы нарративов об общественно значимом событии осуществляется в перспективе создать альтернативу классическим опросным методам, с помощью которых в традиционной социологии обычно определяются мнения людей. В сравнении с опросами, система будет обладать большей степенью автоматизации, репрезентативностью, полнотой охвата мнений различных социально-демографических и географических групп, а также полнотой охвата исследуемой тематики, включая основные сферы жизни людей: бытовую, политическую, экономическую, социальную, культурную.

Актуальность исследования обусловлена:

1) универсальностью разработанных инструментальных средств, позволяющей в дальнейшем адаптировать их под любую актуальную тематику, скорректировав входные параметры тематического моделирования;
2) массовым характером «события» вакцинации от COVID-19, его значением для большого количества людей, связью с широким кругом социальных, политических, экономических, религиозных, культурных проблем;
3) вкладом в понимание специфики оперирования коллективным сознанием нарративами, сводящими сложные социальные процессы и фрагментарные, разрозненные факты к удобным для восприятия упрощенным повествовательным структурам, отличающимся логичностью, аксиологичностью и интерсубъективностью; учет данного феномена представляется важным для системы практических действий, направленных на улучшение качества общественных коммуникаций и принятия решений на всех уровнях управления.

Обзор литературы и исследований

Задача автоматизированной детекции нарративов, в том числе с привлечением технологий анализа больших данных, на сегодняшний день является объектом внимания исследователей в разных научных областях, однако имеющиеся работы характеризуются сильной неоднородностью в трактовке термина «нарратив» и актуализируемых методах.

Так, под «нарративами» в зарубежном академическом дискурсе часто принято понимать научные взгляды по одной и той же теме, научные исследования в какой-либо узконаправленной области. Для составления обзоров литературы, за которыми закрепилось наименование "narrative review" (описательные обзоры) (например, [1–3]), активно применяются автоматизированные методы анализа данных.

В работах, сфокусированных на материалах, не относящихся к научным текстам, имеют место, например, такие подходы к автоматизированному извлечению нарративов, как идентификация типовых элементов событийной структуры (в отвлечение от содержания) [4]; поиск и кластеризация повествовательных текстов о событии [5]; приравнивание нарратива к теме и тематическое моделирование текстов [6]; автоматизированное реферирование тематической коллекции текстов и генерация нарративных резюме [7] и др. Таким образом, специфика данных работ подтверждает слабую конвенциональность термина «нарратив», низкую степень пересечения исследовательских задач и методов.

В контексте настоящего исследования нарратив выступает единицей анализа общественного мнения и настроений, циркулирующих в цифровом пространстве относительно некоторого «события» (в нашем случае, вакцинации от COVID-19). Наиболее ценный материал в данном случае представляют социальные сети.

Заявленный исследовательский вектор отчасти коррелирует с разрабатываемыми в компьютерных науках технологиями «извлечения мнений» (opinion mining), которые чаще всего оказываются сфокусированы на задачи маркетинга, анализ поведения и предпочтений потребителей [8], в меньшей степени – на изучение политических предпочтений, прогнозирование политических событий на основе того, как ведут себя граждане и что они обсуждают в социальных сетях [9]. Данные исследования обычно базируются на использовании моделей машинного обучения, однако категория «нарратива» в них, как правило, не актуализируется.

На основе обзорного анализа хрестоматийных и современных научных источников в области

нарратологии, представленного в предыдущей работе авторов [1], в контексте задачи автоматизированного извлечения нарративов из социальных сетей операциональным определением «нарратива» авторы предлагают считать «тематически релевантное высказывание, эксплицированное в форме предикативного простого или сложного предложения, содержащее суждение (утверждение или отрицание), обобщающее высказанные в социальной сети мнения большого количества пользователей и содержащее аксиологическую позицию (оценку, личное отношение к событию)» [2]. Таким образом, нарратив может иметь, например, следующий вид: «Коллективный иммунитет – единственный способ борьбы с коронавирусной инфекцией»; «Вакцина от коронавируса опасна, так как не прошла достаточных испытаний»; «Вакцина от коронавирусной инфекции – это биологическое оружие»; «Привитые болеют ровно так же, как и непривитые» и т.п.

Категория «нарратива» используется для решения близких задач в ряде зарубежных работ: [10–12] и др. Так, в [13] решается задача картирования доминирующих тематических аспектов (нарративов) по заявленной проблеме в контенте социальной сети. В этих работах методы и подходы к извлечению нарративов также существенно разнятся и имеют в каждом случае ряд ограничений.

Таким образом, настоящее исследование вносит вклад в разработку современных подходов к автоматизированному извлечению нарративов из больших данных социальных медиа, при этом актуализируется и вовлекается аутентичный материал российской социальной сети, а также используется конкретизированное операциональное определение нарратива, учитывающее как содержательное наполнение данного понятия, влекущего за собой большую междисциплинарную традицию, так и его оптимальную формальную экспликацию в качестве минимальной коммуникативной единицы – высказывания, формализованного в виде повествовательного предложения.

Исследования последних лет отражают высокую практическую значимость актуализации новых методов и подходов в области сбора и анализа данных. Аналитика данных, в том числе использование больших данных, искусственного интеллекта и методов машинного обучения, занимает все большее место в науке и практике управления [14]. Среди типов данных, применяемых как для выработки управленческих решений, так и для улучшения продуктов и бизнес-процессов, активно используются данные социальных сетей [15]. Подчеркивается, что актуализация новых технологий на основе данных отражает эволюцию управления в информационную эпоху [16].

Вышесказанное в полной мере относится к сфере здравоохранения. Конвергенция здравоохранения и аналитики больших данных открыла новые возможности для оптимизации стратегий медицинской коммуникации, кампаний общественного здравоохранения, разработки индивидуальных вариантов коммуникации и инициатив по вовлечению пациентов [17].

События пандемии и политики массовой вакцинации во многих государствах, в том числе в России, вызвали большой резонанс со стороны населения в социальных сетях. Дискурс вакцинации отличается резкой поляризацией мнений, в медиапространстве прочно закрепились провакцерские и антивакцерские нарративы [18–21]. Тема вакцинации от COVID-19, таким образом, представляет собой удобный материал для анализа порожденных ею нарративов в социальной сети: большое количество пользовательских сообщений, эмоциональность и оценочность текстов, разброс мнений и версий относительно вакцинации дают богатую эмпирию для разрабатываемого инструментария анализа.

Материалы и методы

В качестве источника данных выбрана российская социальная сеть «ВКонтакте» [3], ввиду ее популярности среди российских пользователей [4] и открытого API, делающего возможным доступ к данным и их автоматизированный сбор.

Предварительно был осуществлен сбор данных, то есть подготовка тематической коллекции: извлечение из «ВКонтакте» пользовательских сообщений, посвященных вакцинации от COVID-19. За период с 01.01.2020 по 01.03.2023 авторами извлечено с помощью автоматизированных методов сбора данных 4,5 млн сообщений, релевант-

[1] *Саркисова А.Ю., Дунаева Д.О., Петров Е.Ю.* О концептуальном и операциональном определениях понятия «нарратив» (к проблеме автоматизированного извлечения нарративов из больших массивов данных) // Государственное управление. Электронный вестник. 2024. № 104. С. 77–94. EDN: https://elibrary.ru/dlkukh. https://doi.org/10.55959/MSU2070-1381-104-2024-77-94

[2] Там же, С. 87.

[3] ВКонтакте. URL: https://vk.com/ (дата обращения: 05.09.2024).

[4] Аудитория восьми крупнейших соцсетей в России в 2023 году: исследования и цифры // PPC World. 16.05.2023. URL: https://ppc.world/articles/auditoriya-vosmi-krupneyshih-socsetey-v-rossii-issledovaniya-i-cifry/#Vk (дата обращения: 05.09.2024).

ных тематике вакцинации от COVID-19 [5]. Данная тематическая коллекция текстов составила материал исследования.

Для извлечения нарративов были использованы различные методики с применением инструментов аналитической платформы PolyAnalyst [6]: тематическое моделирование (методом PLSA), алгоритмы индексирования текста с этапом идентификации предложений, кластеризация, агрегация данных, расчет меры близости ключевых слов, частичная ручная разметка и валидация данных, а также нормализация данных и расчет количественного индекса («индекса популярности»). Задачи и специфика их применения охарактеризованы ниже, в описании авторской методологии извлечения системы нарративов из текстов пользователей социальной сети.

Результаты исследования

Главный результат проведенного исследования – разработанная методология извлечения нарративов из больших массивов данных социальной сети. Она включает в себя ряд последовательных этапов.

Тематическое моделирование публикаций пользователей

Полученная коллекция 4,5 млн текстов была разделена на тематические кластеры посредством алгоритма тематического моделирования с использованием аналитической платформы PolyAnalyst [7]. Реализация тематического моделирования выполнена с помощью вероятностного латентно-семантического анализа (англ. PLSA – Probabilistic latent semantic analysis). Данный метод позволяет пользователям представить документы в виде числовых векторов в пространстве слов. Совместная встречаемость слов позволяет получить данные о тематике коллекции документов. Одной из основных проблем является определение числа кластеров. В данном случае задача решалась эмпирическим путем. Указывались различные диапазоны значений количества кластеров, и система определяла наиболее подходящее. Далее темы просматривались вручную, при необходимости диапазоны корректировались.

В результате применения PLSA к извлеченным 4,5 млн текстов о вакцинации получено 138 кластеров. Каждому кластеру был присвоен уникальный числовой ID, который далее на протяжении исследования не менялся. Номер кластера (ID) одновременно присваивается каждому тексту, в соответствии с кластером, в который он попадает. Имя кластера представляет собой строку, где через точку с запятой приводятся все предикторы (слова или фразы), на основе которых тот или иной кластер был выделен. Порядок предикторов не имеет значения, учитывается только сам факт их присутствия или отсутствия в записи, на основе чего определяется ее принадлежность к кластеру. В результатах тематического моделирования кластеры расположены по убыванию количества входящих в кластер сообщений. Самый крупный кластер включает 88055 текстов, самый мелкий – 4948 текстов.

Ручная разметка кластеров и ее результаты

На следующем этапе была осуществлена ручная разметка полученных 138-ми кластеров. Анализировались ключевые слова, моделирующие содержание кластера (то есть результат машинного тематического моделирования), и, выборочно, примеры текстов, входящих в кластер.

Ручная разметка производилась авторами. Чтобы обеспечить надежность разметки, применялась процедура перекрестной проверки. Каждый автор независимо размечал тексты, после чего результаты сравнивались и анализировались. В случае выявления расхождений обсуждались причины и искались консенсусные решения. Такой подход позволил минимизировать субъективность и повысить качество разметки данных.

Кластеры были разделены на 3 группы: 1) не содержащие оформленного нарратива (информационно-новостные или рекламные сообщения о вакцинации); 2) транслирующие один выраженный нарратив; 3) содержащие несколько разных нарративов.

Итак, первая полученная группа кластеров маркирована как информационно-новостные публикации. Таких кластеров оказалось 27, они включают в себя 711491 сообщение пользователей «ВКонтакте». Примеры сообщений из таких кластеров:

[5] О методологии и процедуре сбора данных см. подробнее: *Саркисова А.Ю., Петров Е.Ю., Дунаева Д.О.* Разработка системы лингвистических маркеров для автоматизированной выгрузки тематических текстовых данных из социальной сети // Государственное управление. Электронный вестник. 2023. № 97. С. 70–84. EDN: https://elibrary.ru/dbvnty. https://doi.org/10.24412/2070-1381-2023-97-70-84

[6] Преимущества выбранного инструмента аналитики данных изложены в работе авторов: *Петров Е.Ю., Саркисова А.Ю.* Ресурс аналитической платформы PolyAnalyst в социогуманитарных научных исследованиях // Открытые данные – 2021: материалы форума, Севастополь, 30 сентября – 2 октября 2021 г. Томск: Изд-во Том. ун-та, 2021. С. 94–104. EDN: https://www.elibrary.ru/mspldg

[7] Информационно-аналитическая платформа PolyAnalyst. URL: https://www.megaputer.ru/; *Ананян С.М., Сазонов Д.С., Слынько Ю.Н., Соломатин Е.Б.* Аналитическая платформа PolyAnalyst: архитектура, функциональность, практика применения. Москва: Горячая линия – Телеком, 2023. 232 с. URL: http://www.techbook.ru/book.php?id_book=1300 (дата обращения: 05.09.2024)

- «На этой неделе в Ульяновскую область поступит второй транш вакцин от COVID-19. 7 декабря Губернатор Сергей Морозов провел штаб по вопросам комплексного развития региона. Среди прочих тем были рассмотрены и направления противодействия распространения коронавируса»;
- «Прививку от коронавируса во Владивостоке можно сделать в 27-ми пунктах вакцинации. На днях мобильные пункты открылись в торговых центрах «Седанка Сити» (ул. Полетаева, 6д) и «Черемушки» (ул. Черемуховая, 15)»;
- «С начала пандемии петербуржцы продолжают ревакцинироваться от коронавируса. Всего привились 3 129 990 человек, 7500 из которых – дети. За четверг, 5 января, в Петербурге ревакцинировались 299 человек. Всего с начала пандемии коронавируса повторную прививку сделали 828 456 жителей».

Данные кластеры исключались из дальнейшего анализа, так как формализованы по структуре и стилю, не отражают оценочной, аксиологической пользовательской позиции, не выражают глубинных смыслов и эмоций. В целом такие тексты не содержат элементов личных историй и антропоцентрического опыта, поэтому не могут быть названы пользовательскими нарративами. Тем не менее, они представляют интерес и имеют значение как отражение гранд-нарратива (термин Ж.-Ф. Лиотара) – официального нарратива о пользе и эффективности вакцинации, поддерживаемого государством. Большинство таких сообщений публикуются от лица сообществ, а не от лица пользователей [8]. Новости о производстве и выборе вакцин, объявления и рекомендации о том, где можно пройти вакцинацию, статистика о растущем количестве вакцинированных, безусловно, работают на проваксерский нарратив. Примеры кластеров данной группы отражены в табл. 1.

Таблица 1

Примеры кластеров с информационно-новостной повесткой

Table 1

Examples of clusters with official information and news

№ кластера	Ключевые слова	Количество сообщений	Содержание кластера (сформулировано авторами «вручную»)
43	тысяча, область, житель, коронавирус, компонент, человек, регион, сделать, получить, вакцинация	37499	Официальные сообщения, новости, статистика о вакцинации в регионе
53	испытание, доброволец, центр, вектор, клинический, исследование, коронавирус, эпиваккорон, имя, разработать	28729	Новости о производстве вакцин
55	пункт, улица, мобильный, вакцинация, работать, торговый, центр, поликлиника, снилс, паспорт	28138	Объявления, где можно сделать прививку. Вакцинация на местах
58	вакцинация, поступить, доза, регион, область, прививочный, партия, работник, пункт, тысяча	27063	Объявления о том, что в определенный регион поступила определенная вакцина. Статистика о вакцинированных по регионам
59	спутник, российский, рфпи, инвестиция, воз, фонд, одобрить, производство, прямой, зарегистрировать	27015	Объявления о регистрации вакцин в мире, о российской вакцине в мире
60	российский, спутник, беларусь, партия, производство, коронавирус, белоруссия, доза, сербия, лукашенко	26778	О поставках российской вакцины в регионы, в Беларусь, в другие страны
66	испытание, компания, клинический, коронавирус, фаза, разработать, эффективность, китайский, доброволец, оксфордский	25730	Новости о производстве и испытании вакцин за рубежом
109	заболеваемость, регион, область, койка, неделя, рост, пациент, инфекция, орви, число	16984	Сводки о заболеваемости по регионам
115	москва, собянин, мэр, москвич, столица, сергей, вакцинация	16223	Новости о политике относительно вакцинации в Москве

Составлено авторами на основе проведенного анализа

Compiled by the authors based on the analysis conducted

[8] Подробнее об этом см.: *Дунаева Д.О., Петров Е.Ю., Саркисова А.Ю.* Поляризация мнений в нарративах о вакцинации от COVID-19 в социальной сети «ВКонтакте» // Социальные практики и управление: проблемное поле социологии. Материалы VI Сибирского социологического форума с международным участием. Новосибирск: Издательство НГУЭУ, 2023. С. 23–28.
EDN: https://elibrary.ru/ghmemo

Во вторую группу вошли 43 кластера, в каждом из которых оказалось возможным выделить один ведущий нарратив. Эти кластеры включили в себя 1 585 984 сообщения пользователей «ВКонтакте». В данной группе оказались, например, все выявленные на текущем этапе «конспирологические» нарративы, нарративы о принудительной вакцинации и праве на выбор, о фармбизнесе и выгодоприобретателях, о важности коллективного иммунитета и другие. Многие из кластеров этой группы включают поляризованные мнения, но содержат центральный содержательный вопрос, вокруг которого ведется дискуссия. Примеры кластеров, объединенных стержневым нарративом, представлены в табл. 2.

Таблица 2

Примеры кластеров, эксплицирующих один ведущий нарратив

Table 2

Examples of clusters with one leading narrative

№ кластера	Ключевые слова	Количество сообщений	Содержание кластера (сформулировано авторами «вручную»)
6	делать, прививка, заставить, работа, добровольно, хотеть, прививок, добровольный, народ, принудительно	76 118	Принудительная вакцинация. Давление работодателя
12	умереть, прививка, ковид, знакомый, лежать, больница, прививок	68 026	Ковид опасен. Бояться нужно ковида, а не прививок
22	мозг, маска, намордник, стадо, баранавирус, баран, носить, человек, прививка	57 292	Народ – стадо баранов, позволяет с собой так обращаться
26	испытание, пройти, стадия	51 137	Вакцина не прошла достаточных испытаний
32	убить, население, война, вирус, оружие, народ, уничтожить, добить, придумать, уничтожение	45 393	Вакцина – биологическое оружие. Теория заговора: цель – сократить население Земли
38	испытание, эксперимент, подопытный, кролик, пройти, экспериментальный, ответственность, подписать, испытать, добровольный	41 206	Не хочу быть подопытным кроликом. Неприемлемость медицинских экспериментов на людях
56	близкий, здоровье, сделать, вакцинироваться, вакцинация, болезнь, жизнь, беречь, защитить, способ	28 133	Нужно формировать коллективный иммунитет. Очень важно, чтобы все вакцинировались
65	миллиард, компания, производство, рубль, произвести, производитель, фармацевтический, доллар, завод, заработать	25 920	Объем продаж вакцин. Пандемия – бизнес-проект. Соревнования производителей вакцин
71	эффективность, спутник, исследование, журнал, российский, опубликовать, результат, данные, фаза, испытание	25 079	Спутник – эффективная вакцина. Ее эффективность подтверждена
82	бог, христос, церковь, антихрист, господь, православный, святой, иисус, грех, печать	22 551	Вакцинация – это сатанизм. Вакцина – печать Антихриста. Прививка не вылечит душу
112	компенсация, страховой, рубль, смерть, осложнение, выплатить, выплата, случай, страховка, суд	16 534	Должны быть предусмотрены компенсации за осложнения от принудительных прививок

Составлено авторами на основе проведенного анализа
Compiled by the authors based on the analysis conducted

Нарратив, эксплицированный в данных кластерах, был сформулирован «вручную» на основе однородной тематики кластера. Таким образом, первые 43 нарратива из тематической коллекции (4,5 млн текстов) были определены на данном этапе.

Оставшиеся 68 кластеров (2 115 884 сообщения), составляющие третью группу, были оценены в ходе разметки как содержащие более одного нарратива, сформулировать которые на настоящем этапе оказалось затруднительно, поэтому они нуждаются в дальнейшем анализе. В табл. 3 представлены примеры таких кластеров.

Углубленный анализ кластеров третьей группы

Для семантической конкретизации нарративов, рассеянных в содержательно и оценочно неоднородных кластерах третьей группы, была применена следующая методика.

Прежде всего было необходимо разделить все тексты на отдельные предложения.

Метод разбивки документа на предложения имеет ряд преимуществ. Мы работаем с текстами на естественном языке, которые состоят из ограниченного числа абзацев и предложений [22].

Таблица 3

Примеры кластеров, включающих разные нарративы

Table 3

Examples of clusters with different narratives

№ кластера	Ключевые слова	Количество сообщений	Содержание кластера (сформулировано авторами «вручную»)
1	умереть, прививка, заболеть, гарантия, дать, смерть, ковид, статистика, осложнение, шанс	88 055	Поляризованный кластер. Споры о том, что опаснее, ковид или вакцина. Дает ли прививка гарантии. Запросы статистики
2	прививка, болеть, ковид, делать, сделать, переболеть, заболеть, муж, перенести, грипп	85 958	Поляризованный кластер. «Я буду делать прививку, потому что...» / «Я не буду делать прививку, потому что...». «Кто-то сделал прививку – и у него...» / «Кто-то не сделал прививку – и у него...». Апелляция к фактам, опыт вакцинации, личные мнения
3	привитый, легкий, болеть, заболеть, прививка, форма, заразить, перенести, защитить, болезнь	85 285	Разделение населения на привитых и непривитых. Почему привитые тоже болеют, умирают и т.д. Болеют ли привитые легче непривитых
5	спутник, пфайзер, признать, привиться, воз, страна, ковивак, выбор, одобрить, хороший	78 012	Обсуждение российской вакцины, сравнение с зарубежной. Много смежных реплик на тему: Россия – мир – ВОЗ
30	смертность, статистика, почему, народ, ковид, человек, заболеваемость, привитый, власть, умереть	47 634	О манипуляции статистикой. Доверие / недоверие к статистике
36	оспа, прививка, обязательный, делать, прививок, дитя, эпидемия, ссср, привить, история	43 938	Дискуссия об истории вакцин. Сравнения с вакцинами от других болезней
42	температура, день, укол, прививка, болеть, слабость, боль, рука, подняться, вечер	37 979	В основном – сообщения о своей непосредственной реакции на прививку
46	вирус, карантин, маска, грипп, человек, эпидемия, паника, мера, смертность, заразить	33 828	Ситуация с ковидом в мире. Поведение в условиях пандемии. Гадания, что будет
73	больница, зона, красный, врач, актер, реанимация, умереть, артист, гаркалин, скончаться	24 806	Смерти в больницах (от ковида, от прививки)
135	файл, копия, справка, документ, килобайт, медицинский, сертификат, паспорт, общежитие, карта	10 554	Требования к медицинским документам в школах, общежитиях, поликлиниках

Составлено авторами на основе проведенного анализа

Compiled by the authors based on the analysis conducted

Как правило, авторы используют абзацы и предложения для логико-смыслового членения своих текстов. Одно и то же слово в разных предложениях может нести разные смыслы, иметь разную оценочную коннотацию. Но при написании текста авторы сами организуют структуру документа в логические единицы – отдельные предложения. Высказывание (грамматически выраженное предложением) – минимальная коммуникативная единица речи. Таким образом, в отдельном предложении заключается одна минимальная уникальная, самостоятельная и законченная мысль. (В классической нарратологии также предложение считается минимальной единицей повествования, при этом признается возможность резюмирования рассказа до одного предложения [23, 24].)

Использование такого подхода позволяет учесть дополнительный уровень информации и обрабатывать текст документа не только на уровне символов (слов), но и на уровне семантики. То есть с уровня лексического анализа текстов (актуализированного в тематическом моделировании) подход позволяет перейти на более высокий, синтаксический уровень анализа.

Тексты делились на предложения с помощью алгоритмов индексирования текста с этапом идентификации предложений (распознаванием конца предложения, англ. End of sentence parsing (EOS parsing)) [25]. Данный алгоритм управляет параметрами процесса идентификации предложений и определяет, какие символы или последовательности символов с высокой вероятностью означают конец одного предложения и предшествуют началу следующего. По умолчанию к таким символам относятся точка, восклицательный знак, вопросительный знак и многоточие, помимо этого, в

качестве такого символа добавлено начало новой строки, так как в исследуемых текстах очень часто предложения начинаются с новой строки. Каждый из перечисленных символов помогает алгоритму выявить границы между предложениями.

После разбивки всех сообщений на отдельные предложения получено 11 700 166 предложений. Однако некоторые из этих предложений очень короткие и не несут самостоятельного содержания, поэтому предложения, состоящие менее чем из 50-ти символов (как и на первом этапе предварительной обработки), были удалены. В результате осталось 7 379 666 предложений.

Следующий этап – кластеризация полученных текстов. Как и на предыдущем этапе, в качестве алгоритма кластеризации использовался вероятностный латентно-семантический анализ (PLSA), только на этот раз в качестве входных данных подавались отдельные предложения. В результате кластеризации получено 285 кластеров (рис. 1).

Кластер №	Имя кластера	Поддержка
#1	эффективность, безопасность, вакцина, спутник	60 996
#2	маска, носить, ходить	47 586
#3	сутки, случай, выявить, заражение, новый, коронавирус, зафиксировать, тысяча, последний, зарегистрироват	46 980
#4	прививка, делать, ковид, сделать, подсказать	46 003
#5	спутник, лайт, ковивак, эпиваккорон, вакцина	45 856
#6	народ, власть, враг	44 076
#7	зрение, точка, мышление, знание, мнение, человек, логика, понять, понятие, критический	42 472
#8	пост, комментарий, писать, написать, тема, автор, коммент, мнение, прочитать, читать	41 731
#9	лёгкий, переболеть, форма, перенести, прививка, легко, болеть, ковид, заболеть, тяжело	41 695
#10	прививка, сделать, прививок, последствие, ребёнок, осложнение, делать, здоровье	40 981
#11	организм, вирус, бороться, иммунитет, ослабить	40 845
#12	тяжёлый, защитить, течение, заражение, болезнь, исход, снизить, летальный, риск, заболевание	40 609
#13	привиться, хотеть, народ, вакцина, спутник, колоться	40 221
#14	вакцина, спутник, векторный, основа, создать, эбола, технология, аденовирус, платформа, ковивак	39 813
#15	путин, сделать, коронавирус, владимир, прививка, президент, песков, рассказать, чувствовать	39 636

Составлено авторами в системе PolyAnalyst

Рис. 1. Результаты кластеризации предложений (лидирующие кластеры)

Compiled by the authors in the PolyAnalyst platform

Fig. 1. Results of sentence clustering (leading clusters)

Помимо отнесения текста к тому или иному кластеру, для каждого текста алгоритм рассчитывает метрику вероятности отнесения текста к данному кластеру: она варьируется от 0 до 1. Чем больше показатель, тем текст более релевантен для этого кластера. С целью выделения наиболее релевантных текстов для каждого кластера были оставлены предложения с максимальным значением метрики. Таким образом для каждого кластера оставлены 1–4 предложения. Один кластер полностью удален, так как содержал инструкции к вакцине или согласия на обработку данных с вложенным файлом, например:

«Файл

1_soglasie_na_privivku_i_na_pers_dannye_Gam-KOVID-Vak (5).docx

Файл DOCX, 16 КБ».

Предложения агрегированы в рамках каждого кластера для удобства восприятия – объединены в одном поле с новой строки. Например:

- «И осложнение у нее после прививки, а не от прививки»;
- «Но как от прививки, так и без прививки могут быть осложнения»;
- «От любой прививки могут быть осложнения, причем именно тут эта прививка»;
- «От любой прививки у тебя могут быть осложнения и ОТ ЛЮБОЙ прививки ты можешь крякнуть»;
- «Теперь у кого осложнения после прививки, говорят, что прививка ни при чем».

В ряде случаев кластер оказался представлен одним предложением. Например:

- «От прививки никто не умер, люди умирали и умирают от других заболеваний»;
- «Эффективность не доказана, безвредность не доказана»;
- «Прививка добровольная, так пусть и будет добровольная».

По форме данные предложения, извлеченные автоматизированными методами, во многих случаях уже соответствуют операциональному определению нарратива, сформулированному выше.

На следующем этапе осуществлялась ручная разметка 284-х кластеров. Экспертно анализировались извлеченные тематики (ключевые слова) и полученные автоматически предложения с наибольшей метрикой вероятности. Результатом разметки стало разделение кластеров на три группы:

1) нерелевантные задачам кластеры (реклама и др.) (27 кластеров);
2) кластеры, эксплицирующие четко выраженный нарратив, релевантный теме вакцинации от COVID 19 (204 кластера);
3) «смешанные» кластеры – включающие как полезные, так и нерелевантные тексты; в них сложно однозначно выделить стержневую формулировку (нарратив) (53 кластера).

Нерелевантные кластеры исключались из дальнейшего анализа.

Для релевантных кластеров были сформулированы определяющие их содержание нарративы. Причем в большинстве случаев формулировка нарратива совпадала со сформированным автоматически предложением с наибольшей метрикой вероятности, например: «Пенсионный возраст подняли, цены подняли, пенсии нищенские, зарплаты такие же, а зато вакцинация»; «Ковид сделан искусственно, но что я могу тебе ответить: маску в помощь, вакцина на подходе»; «Чем больше процент привитых – тем больше процент смертности, к чему бы это»; «Вакцин не хватает, а Россия, как всегда, делится с другими странами»; «Оспа была побеждена с помощью вакцинации»; «Колют тех, кого надо, а избранных – не колют»; «Часть людей не верит вакцинам и не верит вообще ничему» и др. В оставшихся случаях формулировка корректировалась «вручную», с сохранением основного смыслового содержания, например: «Вакцина от вируса не возможна, потому что вирус мутирует»; «Вирус мутирует и надо остановить распространение сейчас, потом будет поздно»; «Пандемия (не) может сравниться с другими эпидемиями XX века»; «Доверять власти нельзя, поэтому нельзя доверять и вакцине» и др.

«Смешанные кластеры» нуждались в очередной дополнительной фильтрации. Была поставлена задача сравнить их содержание с уже выявленными нарративами. Для этого выполнялось сравнение ключевых слов из «релевантных» кластеров и «смешанных».

Было принято решение рассчитать меру близости ключевых слов на основе косинусного расстояния (англ. Cosine similarity) [26]. Это метрика, которая вычисляет косинус угла между векторами. Значения варьируются от 0 до 1, где 0 – отсутствие сходства, 1 – полное сходство [27]. Данная метрика вычисляется следующим образом:

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}. \quad (1)$$

Таким образом, для расчета метрики необходимо представить слова в виде векторов. Эмбеддинги (векторное представление слов) были получены с помощью предобученной модели paraphrase-multilingual-MiniLM-L12-v2 [9], которая используется для семантического поиска. После чего рассчитывалось косинусное расстояние для списков ключевых слов из «релевантных» и «смешанных» кластеров.

В результате 49 кластеров из 53-х были соотнесены с одним из уже выделенных нарративов на основании сходства более 65%. 4 кластера, для которых не удалось найти соответствия, были исключены из анализа.

Итоговую коллекцию нарративов составили 237 единиц (43 выделено на этапе тематического моделирования и 194 – на этапе анализа предложений).

Расчет индекса популярности

Ставилась также цель рассчитать индекс популярности для каждого нарратива, то есть установить, какие из имеющихся 237-ми нарративов чаще и активнее всего возникают в обсуждениях пользователей социальных сетей, а какие реже.

Показательной метрикой является количество сообщений, транслирующих тот или иной нарратив. Однако, в связи с тем, что часть нарративов извлечена на основе полных текстов сообщений, а другая часть – на основе предложений с наибольшей метрикой вероятности, необходима предварительная процедура нормализации данных.

Нормализация выполнялась отдельно для полных сообщений и отдельно для предложений по следующей формуле:

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}, \quad (2)$$

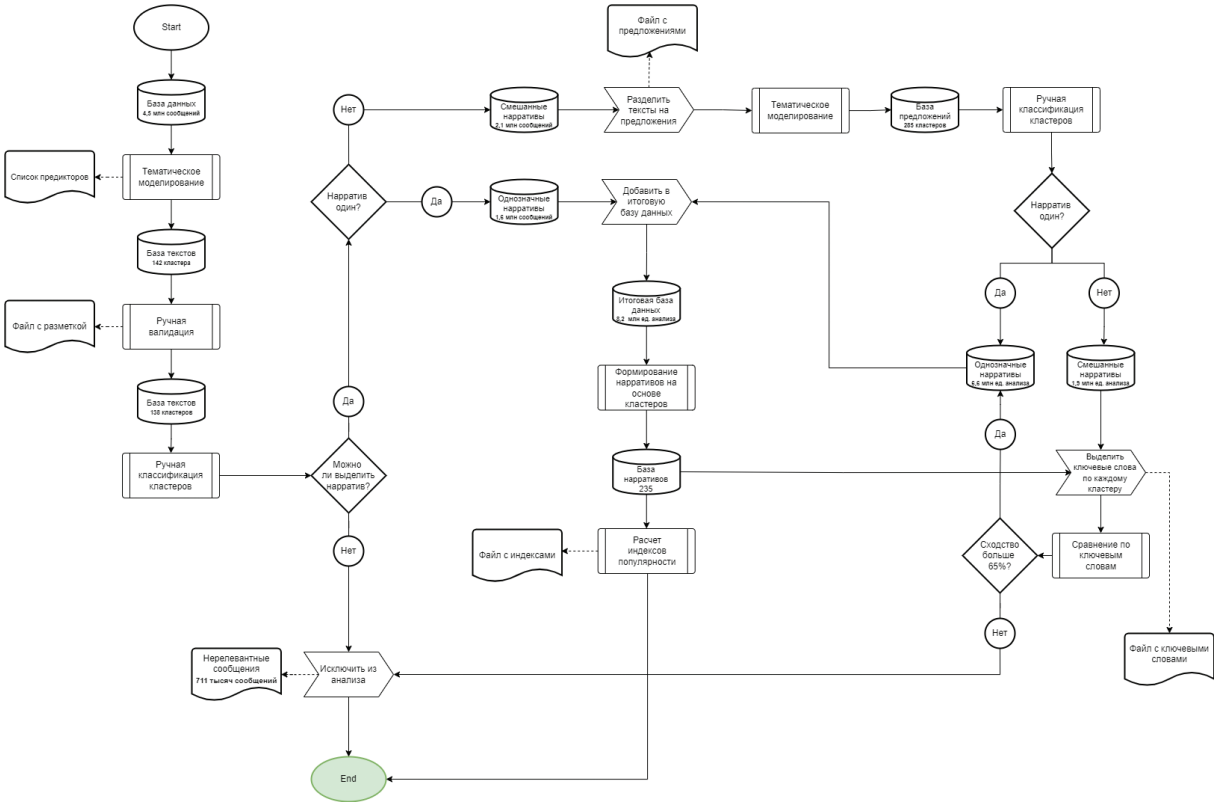
[9] Sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 // Hugging Face. URL: https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 (дата обращения: 05.09.2024).

где z_i – нормализованное значение, x_i – начальное значение, min(x) – минимальное значение, max(x) – максимальное значение.

Полученный индекс меняется в диапазоне от 0 до 1, где 0 – наименее популярный нарратив, 1 – наиболее популярный.

Полученные результаты, их значимость и ограничения

Разработанная методология извлечения нарративов, составившая основной результат исследования, визуализирована на рис. 2.



Составлено авторами

Рис. 2. Разработанная методология извлечения нарративов

Compiled by the authors

Fig. 2. Developed methodology for extracting narratives

Вторым результатом является полученный датасет, включающий 237 нарративов, сопровожденных метаданными: количеством представляющих его текстов, индексом популярности. Содержательный, предметный анализ реестра нарративов не входит в задачи текущего этапа исследования. В табл. 4 представлены 10 наиболее популярных из 237-ми нарративов о вакцинации от COVID-19, циркулирующих в социальной сети «ВКонтакте» в период пандемии.

Практическая значимость расчета индекса популярности нарративов заключается в возможности получения объективных данных, которые могут служить основой для принятия управленческих решений в области общественного здравоохранения, коммуникационных стратегий и социальной политики.

Во-первых, данный индекс позволяет выявить наиболее распространенные и значимые нарративы, что предоставляет аналитическую основу для разработки целенаправленных информационных кампаний. Это особенно важно в условиях, когда необходимо своевременно реагировать на распространение дезинформации и корректировать поведенческие установки населения.

Во-вторых, индекс популярности способствует пониманию динамики общественного мнения, позволяя выявлять изменения в восприятии и отношении к вакцинации на различных этапах пандемии. Это, в свою очередь, может помочь в прогнозировании поведенческих трендов и адаптации стратегий коммуникации, направленных на повышение уровня вакцинации и снижения уровня недоверия среди населения.

Таблица 4

Наиболее популярные нарративы о вакцинации от COVID-19 в сети «ВКонтакте»

Table 4

The most popular narratives about COVID-19 vaccination on VKontakte

№	Нарратив	Индекс популярности
1	Работодатели оказывают давление, принуждая вакцинироваться	1
2	Вакцинацию рекомендуется пройти в следующих поликлиниках	1
3	Необходимо пройти ревакцинацию	0,92674037
4	Имеющиеся вакцины от ковида не эффективны, так как вирус мало изучен, мутирует	0,903177508
5	Массовая вакцинация – это фармбизнес, поддерживаемый государственной властью	0,901876862
6	Ковид очень опасен: бояться нужно ковида, а не прививок	0,877618306
7	Каждый человек имеет право на выбор: вакцинироваться или нет	0,839415617
8	Ковида нет, «барановирус» придуман, это срежессированный спектакль	0,815383917
9	«Добровольное согласие» – это ложь и издевательство, вакцинация носит принудительный характер	0,806430635
10	Если есть антитела, то прививку делать не нужно / нельзя	0,758200874

Составлено авторами на основе проведенного анализа

Compiled by the authors based on the analysis conducted

Нормализация данных перед расчетом индекса обеспечивает корректность сравнения различных нарративов, учитывая различия в источниках данных и способах их извлечения. Это повышает точность анализа и делает выводы более надежными, что особенно важно для научно-обоснованного подхода к решению проблем, связанных с общественным здоровьем.

В ходе апробации разработанной методологии сделаны следующие наблюдения и выводы относительно ее качества и ограничений.

1. Анализ данных осуществлялся автоматизированными методами, «ручной» анализ был сведен к минимуму, к работе не привлекались команды разметчиков и т.д. Большой объем данных (4,5 млн текстов) не позволял обработать / проверить на разных отдельных этапах материал вручную для улучшения качества классификации. Как следствие, методология демонстрирует возможности и ограничения именно автоматизированных методов (при ограниченных «людских» ресурсах) детекции и извлечения нарративов, посвященных репрезентации социально значимого явления в медиапространстве, представленных в большом массиве текстовых данных. Таким образом, ручная валидация данных на всех этапах исследования носит исключительно выборочный характер.
2. Представлено два метода работы с текстами: на уровне полнотекстовых сообщений и на уровне отдельных предложений. Выбор второго из них изначально (то есть в отношении 4,5 млн сообщений), скорее всего, позволил бы получить более детализированную информацию и выделить большее количество нарративов. Однако при разбивке текстов на предложения очень существенно увеличивается количество единиц анализа, что требует для обработки больше времени и вычислительных ресурсов. В целом при работе с большими объемами данных представляется целесообразным уделять большое внимание их предварительной обработке, стремиться максимально сократить нерелевантную и бесполезную информацию, чтобы обеспечить большую производительность и улучшить качество данных. Исходя из этого, данный подход применяется только к группе кластеров, которые не удалось результативно обработать менее трудозатратным способом.
3. Автоматизированные методы анализа оказались хороши для извлечения популярных нарративов. В социальных сетях большая доля несамостоятельного контента (который в настоящем исследовании удалялся на этапе сбора данных, анализировались только уникальные тексты [10]), репостов, повторов, вбро-

[10] *Саркисова А.Ю., Петров Е.Ю., Дунаева Д.О.* Разработка системы лингвистических маркеров для автоматизированной выгрузки тематических текстовых данных из социальной сети // Государственное управление. Электронный вестник. 2023. № 97. С. 70–84. EDN: https://elibrary.ru/dbvnty. https://doi.org/10.24412/2070-1381-2023-97-70-84

сов. Уникальные записи также не отличаются массовым разнообразием: в основном пользователей волнуют одни и те же вопросы, часто источником новых сообщений становятся сами социальные сети, то есть их чтение пользователями. В ходе исследования практически не выявлено редких, неожиданных, оригинальных нарративов, позволяющих увидеть новый ракурс проблемы. Такие нарративы в сети, безусловно, есть (многие из них можно встретить, читая случайный контент) – но единичные, не распространенные нарративы машинными методами вычленить трудно: извлекаются только статистически значимые мнения, присущие большому количеству пользователей. С другой стороны, принципы массовости, интерсубъективности, распространенности в обществе отвечают задачам исследования и критериям анализа больших данных.

4. Большое количество кластеров о вакцинации, выделенных в результате операций тематического моделирования и кластеризации, носит поляризованный характер. Соответственно, формулировка нарратива в таких случаях может иметь, например, следующий вид: «Я (не) поставил прививку и (не) заболел»; «Пандемия (не) может сравниться с другими эпидемиями XX века»; «ВОЗ признала эффективность вакцины – это (не) аргумент». По сути, такая формулировка включает в себя два противоположных нарратива – проваксерский и антиваксерский. Мнения пользователей – авторов сообщений, наполняющих соответствующий кластер, полярны. Ключевые слова, моделирующие содержание кластера, будут в этом случае совпадать, поэтому тематическое моделирование малопригодно для разделения поляризованных мнений. Разделение таких кластеров требует процедуры сентимент-анализа применительно к объектам «вакцинация», «прививка», «Спутник V» и др. Анализ тональности является одной из наиболее сложных задач для решения автоматизированными методами ввиду сложности машинной дифференциации эмоций и оценок. В данном исследовании эта операция не использовалась, так как пробные итерации показали слабую степень согласованности результатов.

В остальных кластерах можно однозначно говорить, что стержневой нарратив – проваксерский (например: «Нужно как в армии: сказали надевать медицинские маски и пользоваться антисептиками, значит, так надо»; «Активно распространяется недостоверная информация о вакцинации») или антиваксерский («Вакцинация – это вживление чипов, цифровых кодов»; «Нельзя принуждать вакцинироваться в условиях, когда никто ни за что не отвечает»).

Выводы

В исследовании описано решение задачи автоматизированного извлечения нарративов о вакцинации от COVID-19, рассеянных в 4,5 млн текстов, опубликованных пользователями социальной сети «ВКонтакте» в период с 01.01.2020 по 01.03.2023. Результатами исследования стали: 1) методология, применимая полностью или частично к любой социально значимой тематике, отмеченной большим количеством откликов в социальных медиа; 2) датасет из 237-ми нарративов, характеризующих общественное мнение россиян о вакцинации от COVID-19 в период пандемии.

Перспективами исследования являются содержательная типология и семантико-прагматический анализ полученной системы нарративов; апробация методологии на других данных; разработка подходов к сентимент-анализу поляризованных кластеров.

Использованные для исследования технологии сбора данных позволяют рассчитать востребованность того или иного нарратива из реестра в конкретных регионах России или у конкретных социально-демографических групп населения (статистически учесть пол, возраст, образование и другие характеристики, отраженные в полях, которые заполняют на личной странице пользователи «ВКонтакте»). Это открывает возможности альтернативы классическим опросным методам, востребованным в социологии и менеджменте для анализа мнений.

Практическая значимость и применимость результатов данного исследования заключаются в нескольких ключевых аспектах.

Во-первых, разработанная методология автоматизированного извлечения нарративов из социальных сетей предоставляет более объективный и масштабируемый инструмент для анализа общественного мнения по сравнению с традиционными методами, такими как опросы. Этот инструмент позволяет не только оперативно реагировать на изменения в массовых настроениях, но и анализировать мнение широкого спектра социально-демографических и географических групп, что делает его незаменимым для мониторинга общественных настроений и принятия информированных решений в условиях быстро меняющейся информационной среды.

Во-вторых, универсальность предложенной методологии позволяет ее адаптировать для анализа различных социально значимых событий, что расширяет ее применимость и делает ее ценным инструментом для исследователей и практиков в области социологии, политологии, менеджмента, маркетинга и других смежных дисциплин. Напри-

мер, данную методологию можно использовать для анализа нарративов, связанных с политическими кампаниями, экономическими кризисами, социальными движениями и другими важными событиями.

В-третьих, понимание специфики формирования и распространения нарративов в социальных сетях способствует более эффективному управлению общественным мнением. Это особенно важно в условиях поляризации мнений, когда одни нарративы могут способствовать стабилизации общественного дискурса, а другие – усугублять социальные напряжения. Разработка и внедрение таких аналитических инструментов могут помочь в разработке стратегий противодействия дезинформации, продвижения общественно значимых инициатив и улучшения коммуникации между государственными институтами и обществом.

Список источников

1. *Zhang Q., Gao J., Wu J.T., Cao Z., Zeng D.D.* Data science approaches to confronting the COVID-19 pandemic: a narrative review // Philosophical Transactions. Series A, Mathematical, physical, and engineering sciences. 2021. Vol. 380. P. e20210127. https://doi.org/10.1098/rsta.2021.0127

2. *Bozkurt A., Karakaya K., Turk M., Karakaya Ö., Castellanos-Reyes D.* The impact of COVID-19 on education: A meta-narrative review // TechTrends. 2022. Vol. 66. P. 883–896. https://doi.org/10.1007/s11528-022-00759-0

3. *Mennella C., Maniscalco U., De Pietro G., Esposito M.* Ethical and regulatory challenges of AI technologies in healthcare: A narrative review // Heliyon Volume. 2024. Vol. 10. Iss. 4. P. e26297. https://doi.org/10.1016/j.heliyon.2024.e26297

4. *Kim J., Monroy-Hernandez A.* Storia: Summarizing social media content based on narrative theory using crowdsourcing // CSCW '16: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing (February 27 – March 2, 2016). San Francisco, 2016. P. 1018–1027. https://doi.org/10.1145/2818048.2820072

5. *Рудакова Г.М., Корчевская О.В.* Разработка системы по обработке нарративных данных // ИТНОУ: Информационные технологии в науке, образовании и управлении. 2018. № 5(9). С. 33–38. EDN: https://elibrary.ru/yofcnn

6. *Бойченко А.Е., Жучкова С.В.* Что скрывает русский рэп? Тематическое моделирование текстов русскоязычной хип-хоп сцены // Журнал социологии и социальной антропологии. 2020. Т. 23. № 2. С. 130–165. EDN: https://elibrary.ru/rqypza. https://doi.org/10.31119/jssa.2020.23.2.6

7. *Ghodratnama S., Beheshti A., Zakershahrak M., Sobhanmanesh F.* Intelligent narrative summaries: From indicative to informative summarization // Big Data Research. 2021. Vol. 26. P. 1–13. https://doi.org/10.1016/j.bdr.2021.100257

8. *Messaoudi C., Guessoum Z., Ben Romdhane L.* Opinion mining in online social media: a survey // Social Network Analysis and Mining. 2022. Vol. 12. P. 25. https://doi.org/10.1007/s13278-021-00855-8

9. *Jaidka K.* Chapter 17: Public opinion analytics with social media // In: Research Handbook on Social Media and Society / Ed. *M.M. Skoric, N. Pang.* 2024. P. 224–239. https://doi.org/10.4337/9781800377059.00028

10. *Oghaz T.A., Mutlu E.C., Jasser J., Yousefi N., Garibay I.* Probabilistic model of narratives over topical trends in social media: A discrete time model // Proceedings of the 31st ACM Conference on Hypertext and Social Media (HT '20). New York, 2020. P. 281–290. https://doi.org/10.1145/3372923.3404790

11. *Shahsavari S., Holur P., Wang T., Tangherlini T.R., Roychowdhury V.* Conspiracy in the time of corona: automatic detection of emerging COVID-19 conspiracy theories in social media and the news // Journal of Computational Social Science. 2020. Vol. 3. P. 279–317. https://doi.org/10.1007/s42001-020-00086-5

12. *Sharma K., Zhang Y., Liu Y.* COVID-19 vaccine misinformation campaigns and social media narratives // Proceedings of the International AAAI Conference on Web and Social Media. 2022. Vol. 16. Iss. 1. P. 920–931. https://doi.org/10.1609/icwsm.v16i1.19346

13. *Edinger A., Valdez D., Walsh-Buhi E., Trueblood J.S., Lorenzo-Luaces L., Rutter L.A., Bollen J.* Misinformation and public health messaging in the early stages of the MPOX outbreak: Mapping the Twitter narrative with deep learning // Journal of Medical Internet Research. 2023. Vol. 25. P. e43841. https://doi.org/10.2196/43841

14. *Shafiq W.* Optimizing organizational performance: A data-driven approach in management science // Bulletin of Management Revew. 2024. Vol. 1. Iss. 2. P. 31–40. URL: https://bulletinofmanagement.com/index.php/Journal/article/view/48 (дата обращения: 05.09.2024).

15. *Saura J.R., Ribeiro-Soriano D., Palacios-Marqués D.* Data-driven strategies in operation management: mining user-generated content in Twitter // Annals of Operations Research. 2024. Vol. 333. P. 849–869. https://doi.org/10.1007/s10479-022-04776-3

16. *Sarioguz O., Miser E.* Data-driven decision-making: Revolutionizing management in the information era // Journal of Artificial Intelligence General Science. 2023. Vol. 4. Iss. 1. P. 179–194. https://doi.org/10.60087/jaigs.v4i1.131

17. *Adegoke B.A., Odugbose T., Adeyemi C.* Harnessing big data for tailored health communication: A systematic review of impact and techniques // International Journal of Biology and Pharmacy Research Updates. 2024. Vol. 03. Iss. 02. P. 001–010. https://doi.org/10.53430/ijbpru.2024.3.2.0024

18. *Johnson N.F., Velásquez N., Restrepo N.J., Leahy R., Gabriel N., El Oud S., Zheng M., Manrique P., Wuchty S., Lupu Y.* The online competition between pro-and anti-vaccination views // Nature. 2020. Vol. 582. P. 230–233. https://doi.org/10.1038/s41586-020-2281-1

19. *Germani F., Biller-Andorno N.* The anti-vaccination infodemic on social media: A behavioral analysis // PLoS One. 2021. Vol. 16. Iss. 3. P. e0247642. https://doi.org/10.1371/journal.pone.0247642

20. *Mønsted B., Lehmann S.* Characterizing polarization in online vaccine discourse – A large-scale study // PLoS One. 2022. Vol. 17. Iss. 2. P. e0263746. https://doi.org/10.1371/journal.pone.0263746

21. *Nguyen A., Catalan-Matamoros D.* Anti-vaccine discourse on social media: an exploratory audit of negative tweets about vaccines and their posters // Vaccines. 2022. Vol. 10. Iss. 12. P. 2067. https://doi.org/10.3390/vaccines10122067

22. *Воронцов К.В.* Задачи и методы понимания естественного языка для мониторинга медиа-пространства // В книге: Математические методы распознавания образов: тезисы докладов 20-й Всероссийской конференции с международным участием, г. Москва, 2021 г. Москва: Российская академия наук, 2021. С. 362–367. URL: http://machinelearning.ru/wiki/images/0/02/Mmpr_2021.pdf (дата обращения: 05.09.2024).

23. *Danto A.* Narrative sentences // History and Theory. 1962. Vol. 2. Iss. 2. P. 146–179. URL: https://abuss.narod.ru/Biblio/eng/danto_narrsentences.htm (дата обращения: 05.09.2024).

24. *Genette G.* Narrative discourse: An essay in method. New York: Cornell University Press, 1983. 285 p. URL: https://ia802908.us.archive.org/24/items/NarrativeDiscourseAnEssayInMethod/NarrativeDiscourse-AnEssayInMethod.pdf (дата обращения: 05.09.2024).

25. *Kempen G.* Sentence parsing // In: Language Comprehension: A Biological Perspective. Berlin, Heidelberg: Springer, 1998. P. 213–228. https://doi.org/10.1007/978-3-642-97734-3_7

26. *Гиниятуллин В.М., Салихова М.А., Хлыбов А.В., Чурилов Д.А., Чурилова Е.А.* Оценка семантической близости между критериями оценивания в рабочих программах вуза // Современные наукоемкие технологии. 2021. № 1. С. 12–19. EDN: https://elibrary.ru/rfttvv. https://doi.org/10.17513/snt.38464

27. *Белова К.М., Судаков В.А.* Исследование эффективности методов оценки релевантности текстов // Препринты ИПМ им. М.В. Келдыша. 2020. № 68. 16 с. http://doi.org/10.20948/prepr-2020-68

Статья поступила в редакцию 06.09.2024; одобрена после рецензирования 17.09.2024; принята к публикации 24.09.2024

Об авторах:

Петров Евгений Юрьевич, техник Суперкомпьютерного центра; SPIN-код: 6469-0644, Scopus ID: 57224334888

Саркисова Анна Юрьевна, кандидат филологических наук, доцент, научный сотрудник факультета государственного управления; SPIN-код: 1212-0879, Researcher ID: ABF-4692-2020, Scopus ID: 58125063500

Дунаева Дарья Олеговна, научный сотрудник факультета государственного управления; SPIN-код: 7164-7368, Researcher ID: ADT-1114-2022, Scopus ID: 57328403000

Воронов Александр Сергеевич, доктор экономических наук, доцент, профессор факультета государственного управления; SPIN-код: 4606-5045

Мягков Михаил Георгиевич, PhD, ведущий научный сотрудник факультета государственного управления; Researcher ID: G-6049-2017, Scopus ID: 6602445231

Вклад авторов:

Петров Е. Ю. – разработка методологии, ее практическая апробация и описание; обсуждение результатов.

Саркисова А. Ю. – постановка целей и задач исследования; обсуждение результатов; подготовка основного текста статьи.

Дунаева Д. О. – аналитика данных; обсуждение результатов; формулирование практической значимости исследования.

Воронов А. С. – проведение критического анализа материалов; обсуждение результатов.

Мягков М. Г. – научное руководство; обсуждение результатов.

Авторы прочитали и одобрили окончательный вариант рукописи.

References

1. Zhang Q., Gao J., Wu J.T., Cao Z., Zeng D.D. Data science approaches to confronting the COVID-19 pandemic: a narrative review. *Philosophical Transactions. Series A, Mathematical, physical, and engineering sciences*. 2021; 380:e20210127. https://doi.org/10.1098/rsta.2021.0127 (In Eng.)

2. Bozkurt A., Karakaya K., Turk M., Karakaya Ö., Castellanos-Reyes D. The impact of COVID-19 on education: A meta-narrative review. *TechTrends*. 2022; 66:883–896. https://doi.org/10.1007/s11528-022-00759-0 (In Eng.)

3. Mennella C., Maniscalco U., De Pietro G., Esposito M. Ethical and regulatory challenges of AI technologies in healthcare: A narrative review. *Heliyon Volume*. 2024; 10(4):e26297. https://doi.org/10.1016/j.heliyon.2024.e26297 (In Eng.)

4. Kim J., Monroy-Hernandez A. Storia: Summarizing social media content based on narrative theory using crowdsourcing. In: *CSCW '16: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing (February 27 – March 2, 2016).* San Francisco, 2016. P. 1018–1027. https://doi.org/10.1145/2818048.2820072 (In Eng.)

5. Rudakova G.M., Korchevskaya O.V Delopment of a system for processing narrative data. *ITNOU: Information technologies in science, education and management.* 2018; (5(9)):33–38. EDN: https://elibrary.ru/yofcnn (In Russ.)

6. Boichenko A.E., Zhuchkova S.V. What is inside Russian rap? Topic modeling of the texts of the Russian-speaking hip-hop stage. *The Journal of Sociology and Social Anthropology.* 2020; 23(2):130–165. EDN: https://elibrary.ru/rqypza. https://doi.org/10.31119/jssa.2020.23.2.6 (In Russ.)

7. Ghodratnama S., Beheshti A., Zakershahrak M., Sobhanmanesh F. Intelligent narrative summaries: From indicative to informative summarization. *Big Data Research.* 2021; 26:1–13. https://doi.org/10.1016/j.bdr.2021.100257 (In Eng.)

8. Messaoudi C., Guessoum Z., Ben Romdhane L. Opinion mining in online social media: a survey. *Social Network Analysis and Mining.* 2022; 12:25. https://doi.org/10.1007/s13278-021-00855-8 (In Eng.)

9. Jaidka K. Chapter 17: Public opinion analytics with social media. In: *Research Handbook on Social Media and Society* / ed. Skoric M.M., Pang N. 2024. P. 224–239. https://doi.org/10.4337/9781800377059.00028 (In Eng.)

10. Oghaz T.A., Mutlu E.C., Jasser J., Yousefi N., Garibay I. Probabilistic model of narratives over topical trends in social media: A discrete time model. In: *Proceedings of the 31st ACM Conference on Hypertext and Social Media (HT '20).* New York, 2020. P. 281–290. https://doi.org/10.1145/3372923.3404790 (In Eng.)

11. Shahsavari S., Holur P., Wang T., Tangherlini T.R., Roychowdhury V. Conspiracy in the time of corona: Automatic detection of emerging COVID-19 conspiracy theories in social media and the news. *Journal of Computational Social Science.* 2020; 3:279–317. https://doi.org/10.1007/s42001-020-00086-5 (In Eng.)

12. Sharma K., Zhang Y., Liu Y. COVID-19 vaccine misinformation campaigns and social media narratives. In: *Proceedings of the International AAAI Conference on Web and Social Media.* 2022; 16(1):920–931. https://doi.org/10.1609/icwsm.v16i1.19346 (In Eng.)

13. Edinger A., Valdez D., Walsh-Buhi E., Trueblood J.S., Lorenzo-Luaces L., Rutter L.A., Bollen J. Misinformation and public health messaging in the early stages of the MPOX outbreak: Mapping the Twitter narrative with deep learning. *Journal of Medical Internet Research.* 2023; 25:e43841. https://doi.org/10.2196/43841 (In Eng.)

14. Shafiq W. Optimizing organizational performance: A data-driven approach in management science. *Bulletin of Management Revew.* 2024; 1(2):31–40. URL: https://bulletinofmanagement.com/index.php/Journal/article/view/48 (accessed: 05.09.2024) (In Eng.)

15. Saura J.R., Ribeiro-Soriano D., Palacios-Marqués D. Data-driven strategies in operation management: Mining user-generated content in Twitter. *Annals of Operations Research.* 2024; 333:849–869. https://doi.org/10.1007/s10479-022-04776-3 (In Eng.)

16. Sarioguz O., Miser E. Data-driven decision-making: Revolutionizing management in the information era. *Journal of Artificial Intelligence General Science.* 2023; 4(1):179–194. https://doi.org/10.60087/jaigs.v4i1.131 (In Eng.)

17. Adegoke B.A., Odugbose T., Adeyemi C. Harnessing big data for tailored health communication: A systematic review of impact and techniques. *International Journal of Biology and Pharmacy Research Updates.* 2024; 03(02):001–010. https://doi.org/10.53430/ijbpru.2024.3.2.0024 (In Eng.)

18. Johnson N.F., Velásquez N., Restrepo N.J., Leahy R., Gabriel N., El Oud S., Zheng M., Manrique P., Wuchty S., Lupu Y. The online competition between pro-and anti-vaccination views. *Nature.* 2020; 582:230–233. https://doi.org/10.1038/s41586-020-2281-1 (In Eng.)

19. Germani F., Biller-Andorno N. The anti-vaccination infodemic on social media: A behavioral analysis. *PLoS One.* 2021; 16(3):e0247642. https://doi.org/10.1371/journal.pone.0247642 (In Eng.)

20. Mønsted B., Lehmann S. Characterizing polarization in online vaccine discourse – A large-scale study. *PLoS One.* 2022; 17(2):e0263746. https://doi.org/10.1371/journal.pone.0263746 (In Eng.)

21. Nguyen A., Catalan-Matamoros D. Anti-vaccine discourse on social media: an exploratory audit of negative tweets about vaccines and their posters. *Vaccines.* 2022; 10(12):2067. https://doi.org/10.3390/vaccines10122067 (In Eng.)

22. Vorontsov K.V. Problems and approaches of natural language understanding for media monitoring. In: *Mathematical methods of pattern recognition: Book of abstract of the 20th Russian National Conference with International Participation,* Moscow, 2021. Moscow: Russian Academy of Sciences, 2021. P. 362–367. URL: http://machinelearning.ru/wiki/images/0/02/Mmpr_2021.pdf (accessed: 05.09.2024) (In Russ.)

23. Danto A. Narrative sentences. *History and Theory.* 1962; 2(2):146–179. URL: https://abuss.narod.ru/Biblio/eng/danto_narrsentences.htm (accessed: 05.09.2024) (In Eng.)

24. Genette G. Narrative Discourse: An essay in method. New York: Cornell University Press, 1983. 285 p. URL: https://ia802908.us.archive.org/24/items/NarrativeDiscourseAnEssayInMethod/NarrativeDiscourse-AnEssayInMethod.pdf (accessed: 05.09.2024). (In Eng.)

25. Kempen G. Sentence parsing. In: *Language Comprehension: A Biological Perspective.* Berlin, Heidelberg: Springer, 1998. P. 213–228. https://doi.org/10.1007/978-3-642-97734-3_7 (In Eng.)

26. Giniyatullin V.M., Salikhova M.A., Khlybov A.V., Churilov D.A., Churilova E.A. Evaluation of the semantic similarity between assessment criteria in the educational programs of the university. *Modern High Technologies.* 2021; (1):12–19. EDN: https://elibrary.ru/rfttvv. https://doi.org/10.17513/snt.38464 (In Russ.)

27. Belova K.M., Sudakov V.A. Effectiveness of methods for assessing the texts relevance. In: *Preprints of the Keldysh Institute of Applied Mathematics of the Russian Academy of Sciences.* 2020; (68):16. http://doi.org/10.20948/prepr-2020-68 (In Russ.)

The article was submitted 06.09.2024; approved after reviewing 17.09.2024; accepted for publication 24.09.2024

About the authors:

Evgeny Yu. Petrov, Technician of the Supercomputer C enter; SPIN: 6469-0644, Scopus ID: 57224334888

Anna Yu. Sarkisova, Candidate of Philological Sciences, Associate Professor, Research Associate of the School of Public Administration; SPIN: 1212-0879, Researcher ID: ABF-4692-2020, Scopus ID: 58125063500

Daria O. Dunaeva, Research Associate of the School of Public Administration; SPIN: 7164-7368, Researcher ID: ADT-1114-2022, Scopus ID: 57328403000

Aleksandr S. Voronov, Doctor of Economic Sciences, Associate Professor, Professor of the School of Public Administration; SPIN: 4606-5045

Mikhail G. Myagkov, PhD, Leading Researcher of the School of Public Administration; Researcher ID: G-6049-2017, Scopus ID: 6602445231

Contribution of the authors:

Petrov E. Yu. – development of the methodology, its practical testing and description; discussion of the results.

Sarkisova A. Yu. – setting the goals and objectives of the research; discussion of the results; preparation of the main text of the article.

Dunaeva D. O. – data analysis; discussion of the results; formulation of the practical significance of the study.

Voronov A. S. – conducting a critical analysis of the materials; discussion of the results.

Myagkov M. G. – scientific supervision; discussion of the results.

All authors have read and approved the final manuscript.