

Научная статья
УДК 004.89
doi:10.37614/2949-1215.2022.13.2.001

ИНФОРМАЦИОННО-АНАЛИТИЧЕСКАЯ СИСТЕМА ПОДДЕРЖКИ УПРАВЛЕНИЯ РЕГИОНАЛЬНЫМ РАЗВИТИЕМ НА ОСНОВЕ ОТКРЫТЫХ БОЛЬШИХ ДАННЫХ СОЦИАЛЬНЫХ МЕДИА: КОНЦЕПЦИЯ РАЗРАБОТКИ И ПРАКТИКА РЕАЛИЗАЦИИ

А. М. Федоров^{1✉}, И. О. Датьев², М. Г. Шишаев³, С. С. Федотов⁴, И. Г. Вишняков⁵

^{1–5}Институт информатики и математического моделирования имени В. А. Путилова
Кольского научного центра Российской академии наук, Апатиты, Россия

¹fedorov@iimm.ru, <https://orcid.org/0000-0002-2862-7994>

²datyev@iimm.ru, <https://orcid.org/0000-0002-8372-8704>

³shishaev@iimm.ru, <https://orcid.org/0000-0001-7070-7878>

⁴fedotov@iimm.ru

⁵vishnyakov@iimm.ru

Аннотация

Рассматриваются возможности использования больших открытых данных онлайн-социальных сетей для информационно-аналитической поддержки задач управления региональным развитием. Вводятся понятия социальных объектов и процессов, которые связаны с социально-экономической сферой региона и находят отражение в представляемых в социальных сетях данных. Приводится обзор современных средств и технологий анализа данных социальных сетей для решения государственных, муниципальных и бизнес-задач. Отмечается потенциал расширения функциональных возможностей рассмотренных систем на основе применения формирующего искусственного интеллекта (ИИ). Представлен опыт формирования необходимой системной инфраструктуры и разработки компонентов программного фреймворка для работы с большими данными социальных сетей, применяемых в рамках задач информационно-аналитической поддержки управления региональным развитием на примере Мурманской области.

Ключевые слова:

региональное развитие, социальные объекты и процессы, онлайн-социальные сети, открытые большие данные, формирующий искусственный интеллект, фреймворк информационно-аналитической системы, практика реализации

Благодарности:

исследование выполнено в рамках государственного задания Института информатики и математического моделирования имени В. А. Путилова Кольского научного центра Российской академии наук от Министерства науки и высшего образования Российской Федерации, тема научно-исследовательской работы «Методология создания информационно-аналитических систем поддержки управления региональным развитием, основанных на формирующем искусственном интеллекте и больших данных» (регистрационный номер 122022800551-0).

Для цитирования:

Информационно-аналитическая система поддержки управления региональным развитием на основе открытых больших данных социальных медиа: концепция разработки и практика реализации / А. М. Федоров и др. // Труды Кольского научного центра РАН. Серия: Технические науки. 2022. Т. 13, № 2. С. 5–22. doi:10.37614/2949-1215.2022.13.2.001

Original article

INFORMATION AND ANALYTICAL SYSTEM FOR REGIONAL DEVELOPMENT MANAGEMENT SUPPORT BASED ON OPEN BIG DATA OF SOCIAL MEDIA: DEVELOPMENT CONCEPT AND IMPLEMENTATION PRACTICE

A. M. Fedorov^{1✉}, I. O. Datyev², M. G. Shishaev³, S. S. Fedotov⁴, I. G. Vishnyakov⁵

^{1–5}Putilov Institute for Informatics and Mathematical Modeling of the Kola Science Centre
of the Russian Academy of Sciences, Apatity, Russia

¹fedorov@iimm.ru, <https://orcid.org/0000-0002-2862-7994>

²datyev@iimm.ru, <https://orcid.org/0000-0002-8372-8704>

³shishaev@iimm.ru, <https://orcid.org/0000-0001-7070-7878>

⁴fedotov@iimm.ru

⁵vishnyakov@iimm.ru

Abstract

The paper considers the possibilities of using big open data of online social networks for information and analytical support of regional development. The concepts of social objects and processes are introduced, which are associated with the socio-economic sphere of the region and reflected in the data presented in social networks. An overview of modern tools and technologies for analyzing data of social networks for solving state, municipal and business problems is given. The potential for expanding the functionality of the considered systems based on the use of formative artificial intelligence is noted. The experience of forming the necessary system infrastructure and developing components of a software framework for working with big data of social networks used for information and analytical support of regional development, is presented on the example of the Murmansk region.

Keywords:

regional development, social objects and processes, online social networks, open big data, formative artificial intelligence, information and analytical system framework, implementation practice

Acknowledgments:

the study was carried out within the framework of the Putilov Institute for Informatics and Mathematical Modeling of the Kola Science Centre of the Russian Academy of Sciences state assignment of the Ministry of Science and Higher Education of the Russian Federation, research topic "Methodology for creating information and analytical systems to support the management of regional development based on formative artificial intelligence and big data" (registration number of the research topic 122022800551-0).

For citation:

Information and analytical system for regional development management support based on open big data of social media: development concept and implementation practice / A. M. Fedorov et al. // Transactions of the Kola Science Centre of RAS. Series: Engineering Sciences. 2022. Vol. 13, No. 2. P. 5–22. doi:10.37614/2949-1215.2022.13.2.001

Социальные объекты и процессы как фактор регионального развития

Развитие региона — это сложный, многогранный процесс, целью которого является улучшение определенных качественных и количественных показателей региона с учетом имеющейся региональной специфики и потенциалов роста за счет привлечения внутренних и внешних ресурсов [1]. Под региональным развитием в современном социально ориентированном государстве, каким является Российская Федерация, в первую очередь понимается развитие социально-экономической сферы [2].

Органы государственного и муниципального управления в процессе своей деятельности руководствуются разработанными планами регионального развития и отслеживают ход их реализации посредством контроля целевых показателей. Для этого используется стандартный механизм и специальные службы государственной статистики [3]. Этот канал оценки управляющих воздействий является официальным и вполне объективным.

Однако для получения более полной и оперативной картины реальных социально-экономических процессов статистические данные можно дополнить открытыми данными социальных сетей, которые в данном случае представляют собой неформальный и неофициальный канал обратной связи. Для его чистоты и объективности в каждом конкретном случае требуются дополнительные действия по обработке данных и, возможно, их обогащению сведениями из других источников.

В настоящее время к извлечению и анализу больших открытых данных социальных сетей достаточно активно прибегают как бизнес-структуры [4], так и органы государственной [5] и муниципальной [6] властей, в том числе на международном уровне [7].

В отличие от официальных статистических данных, формат и объем которых формируется целевым образом, открытые данные социальных сетей в общем случае необходимо дополнительно обрабатывать, выделяя в них требуемые социальные объекты, процессы и их характеристики.

Под социальными понимаются такие объекты и процессы, которые прямо или косвенно относятся к социально-экономической сфере региона и каким-либо образом могут быть использованы в реализации стратегий управления региональным развитием.

Реальные социальные объекты. Примерами таких объектов являются отдельные граждане и их различные объединения, целые предприятия, организации, их обособленные подразделения и корпорации, а также администрации разного уровня и их отдельные представители.

В свою очередь, примерами социальных процессов являются в общем случае любые события, связанные с социальными объектами.

Объекты социальных сетей. Социальные сети отражают в накапливаемых в них данных состояния социальных объектов и связанных с ними процессов. В социальных сетях единичные

объекты обычно представляются в виде отдельных пользователей, а групповые объекты — в виде тематических сообществ (групп или пабликов). Например, каждый совершеннолетний гражданин потенциально может зарегистрировать аккаунт в социальной сети и стать её пользователем. Свой аккаунт может зарегистрировать предприятие, организация или, например, администрация города.

С помощью процедуры верификации можно зафиксировать взаимно однозначную связь между аккаунтом и реальным человеком или организацией. Таким образом другие пользователи социальной сети будут знать, что верифицированный аккаунт точно принадлежит тому, с кем он ассоциирован.

События в социальных сетях отражаются в виде публикаций с упоминанием того или иного объекта. Событиями в социальной сети также можно считать изменение состояний ее внутренних объектов. Например, всплеск числа комментариев в обсуждении какой-либо публикации или увеличение числа её лайков, репостов или просмотров — это событие, связанное с теми объектами, которые присутствуют в этой публикации.

Таким образом, через публикации в социальной сети и реакцию на них других пользователей социальные объекты и процессы реального мира тесно связаны с объектами и процессами в социальных сетях.

В данной работе обозначение наличия такой связи является достаточным. Более глубокий концептуальный подход и детальная классификация объектов и процессов социальных сетей представлены, например, в [8].

Социальные объекты, процессы и большие данные. Социальные объекты и порождаемые ими процессы представляют огромный интерес. Прежде всего это связано с исследованиями современного общества и тенденциями его развития. С другой стороны, понимание структуры и характера социальных объектов и процессов позволяет самым эффективным образом организовывать управление территориями, с которыми связаны эти объекты и процессы.

Исследование социальных объектов и процессов в многообразной терминологии, предлагаемой представителями гуманитарных наук [9], в большинстве случаев можно свести к обработке больших данных, извлекаемых из современных социальных медиа. В данной работе под большими данными понимаются открытые большие данные, так как они являются широкодоступными для большинства представителей социума.

Типовые задачи управления региональным развитием

Практически каждая задача регионального управления прямо или косвенно связана с вопросами, относящимися к социальным объектам и процессам. Получение и аналитическая обработка сведений о таких объектах и процессах из социальных сетей позволяют сделать более качественным принятие управленческих решений. Спектр направлений деятельности региональных властей Мурманской области в решении задач управления региональным развитием включает в себя [10]: здравоохранение, строительство и благоустройство, энергетику и жилищно-коммунальное хозяйство, управление имуществом, транспорт и дорожное хозяйство, организацию процессов и HR, взаимодействие с органами местного самоуправления, ветеринарию, социальную сферу, образование и культуру, экономику, финансы и аналитику, юриспруденцию, туризм, лесное хозяйство, природопользование и экологию, СМИ, связи с общественностью и SMM, информационные технологии и безопасность, государственные закупки и др.

Широко распространенной практикой является стремление социально ответственного крупного бизнеса принимать активное участие в жизни региона своего присутствия и своевременно откликаться на чаяния и нужды живущих там людей. В рамках такой деятельности, помимо непосредственной работы с обращениями граждан через приемные аффилированных общественных организаций, также проводится мониторинг популярных местных групп в социальных сетях. Открыто публикуемые в таких группах материалы получают практически моментальный отклик пользователей, выражаемый в комментариях и количестве лайков, репостов и просмотров. На основании статистического и других более глубоких видов анализа этих данных можно давать оценки и строить прогнозы в отношении общественного мнения по тем или иным текущим вопросам или планируемым к реализации инициативам. Традиционными средствами такая деятельность обычно организуется регулярно,

но редко, и в основном ее приурочивают к тем или иным выборным мероприятиям. Автоматизированный анализ больших открытых данных социальных сетей позволяет сделать этот процесс практически непрерывным, что, несомненно, должно повысить качество управления региональным развитием.

Другим примером является заинтересованность бизнеса и представителей властей разного уровня не только в мониторинге обобщенной картины состояния региональных сообществ, но и в возможности оперативной реакции на возникающие в социальных сетях события. В этом случае активное использование коммуникационных возможностей социальных сетей позволяет значительно уменьшить время реакции ответственных лиц на публикации и обсуждения острых вопросов, имеющих широкий общественный резонанс. Известно, что среди пользователей социальных сетей зачастую имеется очень большой процент так называемых технических аккаунтов (ботов, дублирующих аккаунтов и т. п.). Как и в любой другой интернет-дискуссии, в социальных сетях важно иметь представление о собеседниках. Даже простые открытые данные об аккаунтах комментаторов могут дать представление о том, кто из них и с какой степенью вероятности является реальным человеком, а кто — нет. Получение оперативной статистики об активности комментаторов в сообществах и об открытых данных их пользовательских профилей позволяет сделать интернет-дискуссию более продуктивной, а включающее в себя это средство коммуникации региональное управление — более качественным.

Два предыдущих примера демонстрируют потребности представителей властей и бизнеса в использовании пассивных и активных (в том числе интерактивных) средств мониторинга общественного мнения социума региона, за разностороннее развитие которого эти власти несут ответственность. С другой стороны, и представители общественности — отдельные граждане или общественные организации — заинтересованы в доступных и прозрачных средствах отображения деятельности властей и бизнеса. Традиционно обратная связь от властей к социуму реализуется посредством проведения отчетных мероприятий и публикацией в СМИ соответствующих документов. Возможности социальных сетей и дополнительные информационно-аналитические средства анализа их данных позволяют вести мониторинг публичной деятельности ответственных лиц, администраций и организаций. Учитывая то, что социальные сети в большинстве случаев являются неформальным средством коммуникации, использование процедуры обогащения их данных данными официальных источников (например, официальных сайтов, статистических отчетов и т. п.) позволяет сделать такой мониторинг более объемным и качественным. Таким образом, полученные из открытых источников данные о заявленных планах властей и бизнеса можно через некоторое время сравнить с отчетными данными и сделать выводы об эффективности проводимого управления региональным развитием.

Обзор современных систем анализа социальных медиа

Уже сегодня инструментальные средства, построенные на базе современных информационных технологий, позволяют обрабатывать большие данные, генерируемые пользователями социальных медиа. Интернет-ресурс G2-Business Software Reviews [11] содержит перечень, включающий 221 систему мониторинга социальных медиа. Существующие решения нацелены на коммерческое использование (что, впрочем, следует из названия самого ресурса) представителями бизнеса преимущественно в маркетинге, иногда встречаются упоминания о применении в таких областях как: связи с общественностью (PR, пиар), автомобилестроение, недвижимость, информационные технологии и сервисы, Интернет, высшее образование, безопасность, здоровье. Редко (около 10 процентов случаев) в описании систем мониторинга также встречаются указания о возможности использования в управлении некоммерческими организациями. В качестве наиболее популярных систем, по версии G2-Business Software Reviews, называются: Hootsuite, Zoho Social, Sprout Social, Meltwater, Semrush, Agorapulse, Brandwatch Consumer Intelligence, Reputation, Falcon.io (теперь Brandwatch Social Media Management), Sendible, Brand24, eclincher, YouScan, Sprinklr Social Engagement and Sales, Lucidya — Social Media Analytics, Dataminr.

Отечественные разработки также представлены преимущественно системами для коммерческого использования. Компания Brand Analytics предлагает линейку продуктов мониторинга социальных медиа. Одноименная система Brand Analytics собирает данные социальных сетей, видеохостингов,

сервисов карт, публичных каналов, мессенджеров, отзовиков, форумов, блогов, маркетплейсов, онлайн-СМИ, газет, журналов и федеральных телеканалов. Кроме стандартных текстов сообщений, система анализирует тексты на изображениях и расшифровках видео, чекины (сообщения, отправляемые пользователем социальной сети о своём географическом местонахождении), сторис (история с фото или видео длительностью до 15 секунд). Разработчики заявляют о применении машинно-обучаемых алгоритмов для категоризации сообщений, выявления трендов медиаполя, агрессии, распознавания текстов на изображениях, а также автоматического определения тональности упоминаний (сентимент-анализ) [12, 13] с заявленной точностью 85–90 %. В других системах от компании Brand Analytics прослеживается специализация для более узкого круга пользователей и задач с соответствующей вариативностью функциональных возможностей: BA Express (мониторинг социальных медиа для небольших компаний), Brand Visor (медиадашборд для топ-менеджмента), Event Wall (соцмедиадашборд для мероприятий), мониторинг СМИ (мониторинг СМИ с применением ИИ для анализа и оповещения), Actionable Analytics (маркетинговые исследования для выявления инсайтов и трендов).

Компания «Ашманов и партнеры» разработала систему «КРИБРУМ», функциональные возможности и описание которой практически идентичны системе Brand Analytics. «Крибрум находит в социальных медиа упоминания брендов и помогает бизнесу проводить маркетинговые исследования, выявлять информационные атаки и утечки конфиденциальных данных, а также оперативно откликаться на негативные отзывы о продуктах и клиентском сервисе» [14].

Компания «МЕДИАЛОГИЯ» предлагает линейку продуктов PR, SMM, «Инцидент», «Медиа» [15]. Система PR позиционируется как инструмент оперативного мониторинга и анализа СМИ: позволяет получать периодические отчеты, оценивать тональность упоминаний, также разработчики заявляют возможность анализа эффективности инфоповодов и площадок. Система «Медиа» специфицирована для анализа цитируемости и оценки индекса цитируемости СМИ, а также анализа цитируемости каждого сообщения. Система SMM разработана для анализа соцсетей и позволяет оценить мнения потребителей, изучить портрет целевой аудитории, а также измерить эффективность SMM. Система «Инцидент» позиционируется как система управления коммуникациями в соцсетях. Разработчики компании «МЕДИАЛОГИЯ» заявляют оценку тональности своим приоритетным направлением в развитии технологий текстового анализа, а также отмечают использование технологий глубокого обучения — нейронных сетей.

Наряду с вышеупомянутыми, ещё одним масштабным проектом отечественных разработчиков является линейка продуктов анализа социальных медиа от компании «М13» [16]: «Катюша», «Арена», «Арсенал», «Арсенал.Соцмедиа», «Страйк». Система «Катюша» позволяет группировать данные по времени, информационным поводам, тематикам, событиям, персонам и типам СМИ, а также определять тональность, количество публикаций и охват аудитории. Разработчики заявляют, что сообщения появляются в системе уже через 1–10 минут после их публикации на сайте источника. Предусмотрены различные виды оповещений — sms, email, push-уведомления. Система «Арена» ориентирована на потребности и запросы руководителей и позволяет в режиме реального времени отображать только самые важные новости. Системы «Арсенал» и «Арсенал.Соцмедиа» обладают схожими функциональными возможностями, позволяют проводить анализ по заданным тематикам с помощью контекстных запросов, учитывающих в том числе и дополнительные параметры публикаций (время, охват и др.). Результаты мониторинга в виде отчетов экспортируются в удобные для пользователя форматы (в том числе .xlsx и .docx). Система «Страйк», по сути, представляет собой телеграм-бот для мониторинга социальных сетей «Твиттер» и «Телеграм». Основная особенность — отслеживание упоминания бренда в единой ленте без необходимости подписываться на множество каналов и авторов. Уже через 2–3 минуты после опубликования в ленте отображается часть сообщения, содержащая упоминание с указанием автора, социальной сети и времени публикации.

Система Scan Interfax [17] создана для анализа деловой репутации компаний и оценки рисков на основе СМИ и социальных медиа. Спектр решаемых задач соответствует традиционному пиар: «мониторинг упоминаний в СМИ, отзывов в соцсетях, налаживание диалога с изданиями, обнаружение и минимизация репутационных рисков, управление репутацией, нейтрализация негатива, быстрая аналитика, экспресс-оценка инфоповодов».

Система Babkee [18] заявлена разработчиками как «бесплатный российский сервис мониторинга упоминаний и анализа соцмедиа, управления репутацией и изучения активности конкурентов». Бесплатно можно получить лишь усеченную версию функциональных возможностей системы. В качестве особенностей упоминается оценка значимости сообщений как зависимость от авторитетности автора, а также ручное определение тональности.

Exlibris — это медиааналитическое агентство (очевидно, использующее систему сбора и анализа данных социальных медиа), предоставляющее услуги по мониторингу и анализу социальных медиа «для брендов, государственных корпораций и структур, малого и среднего бизнеса» [19]. Заявлен мониторинг и анализ социальных медиа, «управление репутацией, работа с блогерами», «брендинг территорий» [19].

Система «ПрессИндекс» [20] анализирует более 100 000 источников: традиционные медиа, «ВКонтакте», «Твиттер», «Ютуб», «Яндекс.Дзен», «Живой журнал», «Телеграм», «Фейсбук» и «Инстаграм»*. Позволяет отслеживать упоминания как в постах, так и в комментариях, автоматически определять тональность, формировать отчеты, осуществлять рассылки и оповещения, а также разграничивать права доступа к системе. Интересными особенностями, заявленными разработчиками, являются ретроспективный мониторинг, автоматическое определение тональности, выявление трендов.

Система «Интегрум» [21] позиционируется разработчиками как «поисково-аналитическая система по компаниям, индивидуальным предпринимателям, физическим лицам, тендерам и контрактам, СМИ и социальным сетям».

Система IQBuzz [22] — сервис мониторинга социальных медиа и онлайн-СМИ, профессиональный инструмент анализа и управления репутацией в Интернете.

Система SemanticForce [23] позволяет проводить «мониторинг онлайн-медиа при помощи глубокого семантического и визуального анализа». В качестве особенности упоминается также наличие омниканального Хелпдеск. Позиционируется разработчиками как «единая платформа мониторинга и аналитики СМИ и сайтов электронной коммерции, а также реагирования...»

На основе проведенного обзора систем мониторинга социальных медиа следует отметить, что в целом рассмотренные системы построены на идентичных базовых принципах, основные различия состоят в специфицировании систем для решения конкретных, преимущественно коммерческих, задач и использовании конкретными группами пользователей. При разработке систем мониторинга и анализа, всё большее количество компаний концентрирует свои усилия на применении технологий ИИ. Рассмотренные инструментальные системы вполне могут использоваться в качестве информационно-аналитических сервисов для решения задач поддержки управления региональным развитием, поскольку многие из этих задач (однако, далеко не все), по сути, являются пиар, то есть нацелены на управление общественным мнением, а также выстраивание взаимоотношений общества и государственных органов [24]. В качестве примера, следует отметить системы «Инцидент» и «Катюша», в использовании которых проявляют заинтересованность государственные структуры, вероятно, по причине их наилучшего соответствия задачам информационной поддержки государственного управления.

Рассмотренные системы, например, не прогнозируют популярность публикуемых в социальных сетях сообщений [25, 26], результат информационного воздействия, другими словами, не способны производить информационное управление [27, 28], не учитывают многие сетевые характеристики автора сообщения и его влияние на аудиторию [29].

Потенциал функциональных возможностей, полезных в решении задач государственного и муниципального управления, реализован не полностью, не говоря об отсутствии комплексного, модульного, кроссплатформенного решения для информационной поддержки управления региона в целом.

Сегодня в мире интенсивно развивается формирующий ИИ, который поглощает информацию из разных источников и на ее основе создает новый контент (музыка, изображения или видео, программный код). Если рассматривать применимость формирующего ИИ в государственном и региональном управлении, то его ценность заключается в возможности анализа разнородных данных

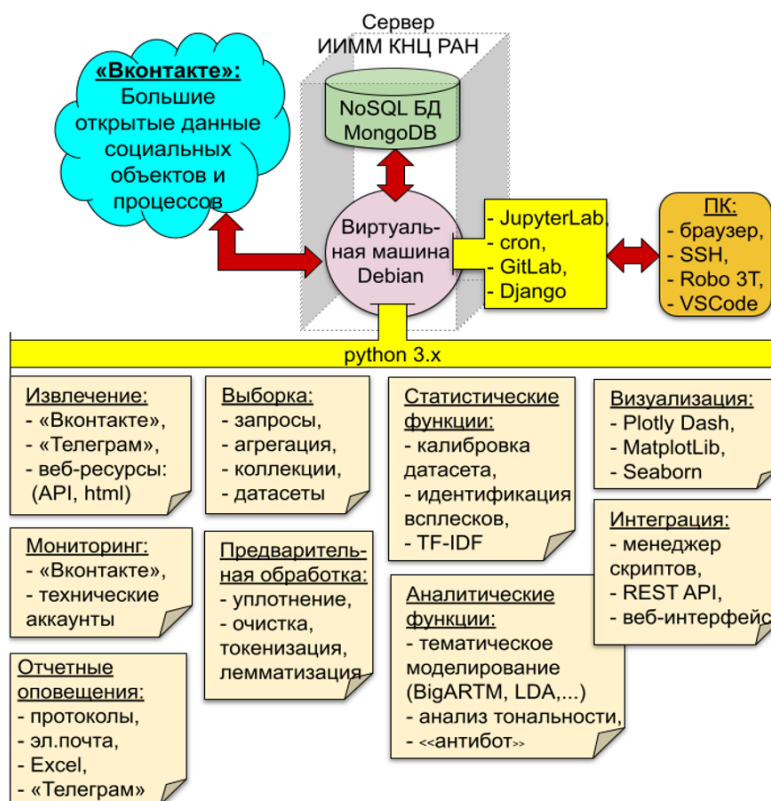
* Компания Meta Platforms Inc. признана в России экстремистской организацией. Принадлежащие ей соцсети «Фейсбук» и «Инстаграм» в России запрещены.

из различных источников (например, различных социальных медиа), выделение ключевых факторов, которые человек может упустить из виду, а также в способности к самостоятельному обучению, то есть адаптации к изменениям входных данных (например, тематической повестки, связанной с мировыми событиями). Вышеперечисленные функции прогнозирования и другие функции анализа, вероятно, могут быть реализованы посредством применения формирующего ИИ для развития рассмотренных и создания новых систем мониторинга и анализа.

Также следует отметить, что процесс пиара в существующих системах рассматривается преимущественно как односторонний — воздействие государственной власти на общество. Однако изначально из определения пиара следует «выстраивание взаимоотношений» — и это не только «подслушивание» социальных медиа на предмет «что говорит народ». Другими словами, гражданское общество тоже должно владеть информацией о состоянии власти или даже иметь возможность мониторинга и анализа информационных ресурсов, в том числе официальных ресурсов власти, что существенно повысило бы социальное самочувствие общества и, возможно, даже уровень ответственности представителей власти.

Практика построения системы сбора и аналитической обработки открытых данных социальных медиа

Несмотря на имеющийся на рынке инструментарий, необходимо развивать технологии и создавать новые инструменты интеллектуальной обработки больших данных социальных сетей. Авторы данной работы в исследовании [30] представили общую схему фреймворка для решения задачи оценки вовлеченности гражданского общества в процессы управления территориями. Именно эта схема стала основой будущих работ авторов над фреймворком, спектр решаемых задач которого расширялся согласно динамике окружающего мира. Сегодня разрабатываемый фреймворк представляется авторам как комплексное, модульное, адаптируемое решение для информационной поддержки регионального управления (рисунок).



Общая схема фреймворка для информационной поддержки регионального управления

Системная инфраструктура

Представленная разработка реализована на сервере, имеющем следующие технические характеристики:

- процессор Intel Xeon E5-2620 v4 2.1GHz (TB up to 3.0GHz) 20Mb 4 × DDR4-1600/18666, 2 ед.;
- память серверная DIMM DDR4 16384Mb PC21300 2666MHz S, 6 ед.;
- платформа Supermicro SYS-6028R-WTR (LGA2011, 8 x 3.5" SAS/SATA HDD, 16 × DDR4, 740W, 2U;
- материнская плата X10DRW-i;
- накопители SATA III WDC WD5000AAKX-00ERMA0 (WD Caviar Blue) 500 GB, WD Gold 7200rpm, 2Tb, 2 ед.;
- сетевое хранилище Asustor.

Фактически для работы используются ресурсы, предоставляемые виртуальной машиной, работающей под управлением операционной системы (ОС) Debian GNU/Linux 10 (buster), с оперативной памятью 10 ГБ и дисковым пространством 200 ГБ. Дополнительно для хранения данных в базе MongoDB используется диск 400 ГБ.

Инфраструктура разработчиков и пользователей

Рабочие места разработчиков и пользователей описываемой системы представляют собой типовые персональные компьютеры (ПК), к техническим характеристикам которых не предъявляется особых требований. Обычно используются ПК под управлением ОС Microsoft Windows поддерживаемых компанией Microsoft версий. Для непосредственной работы требуется наличие браузера и приложения-клиента ssh (например, PuTTY). Подключение к базе данных MongoDB осуществляется любым NoSQL-клиентом (например, Robo 3T). При необходимости разработки могут использоваться локальные среды разработки, такие как Microsoft Visual Studio Code (MS VSCode) или PyCharm. Однако основным средством разработки являлась выполняемая в виде сервиса на стороне сервера среда JupyterLab, работа с которой осуществляется через браузер и для безопасности работы которой необходимо использовать SSH-туннелирование. Для безопасного хранения результатов программных разработок и соблюдения соответствующих прав Института информатики и математического моделирования (ИИММ) на получаемые результаты интеллектуальной собственности используется локальный сервер GitLab.

Блок извлечения данных

В текущей версии фреймворка реализованы программные механизмы извлечения открытых данных из социальной сети «ВКонтакте» [31] и мессенджера «Телеграм». Для этого используются библиотеки на языке программирования python 3.x, которые предоставляют доступ к API [32] этих ресурсов. Технической особенностью работы по извлечению данных являются введенные разработчиками социальной сети и мессенджера ограничения на количество запросов в единицу времени и их суммарное количество в течении суток. В условиях работы с большими данными эти ограничения оказывают свое влияние на скорость и объемы извлекаемых данных. Однако этот процесс организован без какого бы то ни было нарушения лицензионных требований и пользовательских соглашений рассматриваемых ресурсов. Помимо работы с API проводятся тестовые эксперименты по добавлению в инструментарий фреймворка механизмов парсинга исходного HTML-кода страниц веб-сайтов исследуемых ресурсов. Внедрение таких подходов позволит снять большинство технических ограничений на скорость и объемы извлекаемых из открытых источников больших данных.

Блок мониторинга

Реализация блока функций для мониторинга [33] стала логичным развитием функциональных возможностей блока извлечения данных из открытых источников. Здесь главной особенностью является обеспечение регулярной, бесперебойной автоматической работы алгоритмов извлечения данных в течение продолжительного времени. Запуск программ-скриптов по расписанию организован средствами стандартной для ОС UNIX системной службы cron. В целом такой режим работы увеличил интенсивность запросов к открытым ресурсам социальной сети «ВКонтакте». Для корректной работы

с имеющимися ограничениями данной платформы была разработана подсистема управления техническими аккаунтами. В результате гибкой масштабируемой конфигурации эта подсистема позволяет одновременно и вести мониторинг, и проводить другие отдельные эксперименты по обращению к открытым ресурсам «ВКонтакте» без нарушения правил работы с данной платформой.

Также для обеспечения работы мониторинга в непрерывном режиме необходим большой объем хранилища для записи поступающих данных. В качестве хранилища используется NoSQL (база данных MongoDB) и дисковое пространство сервера ИИММ. Объем данных постоянно растет, но применяемые технологии и средства позволяют масштабировать объем используемого хранилища.

Блоки выборки и предварительной обработки данных

Особенностью собранных в процессе мониторинга данных являются их избыточность и дублирование. Это связано с тем, что среди очередной порции извлекаемых данных лишь только часть успевает измениться за прошедший интервал времени. С целью получения из собранного массива данных необходимой информации в рассматриваемом фреймворке предусмотрен соответствующий функциональный блок. Выборка данных производится с помощью специфического языка запросов к базе данных MongoDB, который отличается от традиционного SQL в силу особенностей хранения элементов не в виде записей, а в виде отдельных документов в формате json. С учетом характеристик данных мониторинга, самой востребованной операцией является агрегация. С помощью неё большие объемы дублирующихся данных перерабатываются в коллекции данных и форматированные наборы (датасеты), пригодные для дальнейшей обработки.

После уплотнения данные готовятся для дальнейшей статистической и аналитической обработки и визуализации. Для этого производится их очистка от стоп-слов и других мусорных элементов, документы преобразуются в массивы лексем (токенов), которые нормализуются с помощью процедур лемматизации или стеммирования.

Блок статистических функций

Имеющиеся во фреймворке функции обработки данных представлены в виде двух блоков. Программные инструменты первого блока позволяют проводить первичную статистическую обработку [34] сформированных на предыдущих этапах датасетов. С помощью оценки полученных характеристик исследуемого набора данных и его специфических особенностей можно получить начальное представление об исследуемых объектах и процессах. Например, в данных, представленных в виде числовых рядов, можно идентифицировать и локализовать нетипичные резкие изменения (всплески), а для текстовых данных можно сформировать их векторные представления в виде меры TF-IDF. Для всего набора данных обычно формируется гистограмма распределения.

Блок аналитических функций

Функции данного блока предназначены для выявления закономерностей, более сложных, чем подсчет общеизвестных статистических характеристик. Наиболее проработанными программными модулями являются модули, относящиеся к тематическому моделированию.

Тематическое моделирование. Одним из ключевых преимуществ оперативного и стратегического управления регионом является получение тематической повестки сообществ социальных медиа, которая в информационном пространстве является важным компонентом общества. Для автоматизированного анализа текстов постов и комментариев используются методы тематического моделирования [35]. Недавние исследования авторов данной работы преимущественно были сосредоточены на тематическом моделировании и улучшении качества получаемых тематических моделей. Так, в работе [36] было использовано несколько методов вероятностного тематического моделирования на подготовленном датасете, состоящем из постов сообществ онлайн-социальной сети «ВКонтакте». На основе результатов экспериментов показано, что подход ARTM является одним из самых быстрых по времени обучения модели. В качестве другого преимущества BigARTM отмечена мультимодальность, потенциально значимая для моделирования сообществ социальных медиа, поскольку в них присутствует сопутствующая информация, причем не всегда являющаяся обычным текстом.

Отдельное внимание было уделено автоматической оценке качества тем [37], получаемых с помощью различных методов тематического моделирования. Такой оценкой качества была выбрана метрика когерентности UMass, поскольку исследователи отмечали ее наибольшую корреляцию с человеческими оценками семантической интерпретируемости тем [38, 39].

В роли дополнительного критерия качества использовалась сумма вероятностей десяти верхних токенов темы (так называемая чистота темы) [40]. Однако в предыдущих экспериментах значения метрики когерентности UMass тематических моделей ARTM получились не выше, чем у других методов. Метод LDA существенно превзошел ARTM и другие методы по значению когерентности UMass. Результаты объясняются использованием ARTM без регуляризаторов, поскольку в таком режиме подход ARTM является обычным PLSA [41].

В рамках развития работ по тематическому моделированию сообществ социальных сетей [42] авторы провели эксперименты с настройкой базовых регуляризаторов, доступных в библиотеке BigARTM, и гиперпараметрами метода LDA. В качестве опорного (бейзлайн) для сравнения использовался метод LDA, поскольку он показал лучшие результаты в предыдущих экспериментах [36], а также метод ARTM без регуляризации, то есть фактически PLSA. На основании результатов экспериментов продемонстрировано, что подход ARTM с регуляризацией сопоставим с другими методами тематического моделирования по значениям метрики когерентности UMass и сумме вероятностей десяти верхних токенов темы. Это окончательно подтвердило применимость библиотеки BigARTM в будущих исследованиях для тематического моделирования сообществ социальных сетей. Однако было обнаружено, что метрика когерентности UMass не всегда подходит для автоматизированной оценки качества тематических моделей, полученных с помощью метода ARTM при использовании регуляризаторов. Поэтому было предложено несколько дополнительных метрик, которые могут быть полезными при оценивании качества тематической модели.

На основе проведенных авторами исследований, связанных с тематическим моделированием, был сделан вывод о невозможности на данный момент однозначной идентификации лучшего метода моделирования для коротких текстов социальных медиа. Поэтому ещё в процессе исследования качества тематических моделей применительно к специфическим текстам социальных медиа был расширен перечень возможных к применению методов построения тематических моделей за счет интеграции программных реализаций этих методов в разрабатываемый авторами фреймворк. В текущей версии фреймворка программно интегрированы различные модификации следующих методов тематического моделирования: LDA [43], PLSA [41], GSDMM [44], подход ARTM [40].

Тональность сообщений и «Антибот». Модули, относящиеся к определению тональности сообщений и «Антибот» [45] находятся на стадии разработки: протестировано несколько существующих решений-библиотек, ведутся работы по повышению качества результатов. Оценка тональности позволяет дополнить статистические данные и тематические характеристики сообщений вероятностным значением, определяющим их эмоциональную окраску в виде меток: «негатив», «позитив», «нейтраль». В свою очередь, принципиальное назначение функции «Антибот» состоит в отнесении обрабатываемых текстов сообщений с некоторой вероятностью к тем, которые написаны реальными людьми, и тем, авторство которых принадлежит техническим аккаунтам. Применение данных функций позволяет проводить более качественный анализ.

Следует отметить большой потенциал развития аналитического блока за счет применения в нём в будущем технологий формирующего ИИ.

Блок отчетных оповещений

Программные модули, входящие в состав блока, позволяют формировать регулярный статистический отчет по данным мониторинга. Доступны стандартные форматы .csv и .xlsx. Предусмотрены настройки оповещений, реализованных на базе электронной почты, а также прорабатывается возможность оповещения с помощью мессенджера «Телеграм». Пользовательский интерфейс для гибкой настройки формы представления отчета находится на стадии проектирования.

Блок интеграции и визуализации

Все многообразие функциональных инструментов разрабатываемого фреймворка представляется в виде отдельных скриптов на языке программирования python 3.x. Совместное использование этих функций при реализации отдельных проектов предлагается реализовать через технологию REST API [46]. Для автоматизированного формирования интерфейсов доступа к каждому скрипту разработан и проходит тестовую отладку менеджер скриптов. Предполагается, что доступ к функциям и модулям будет осуществляться и через программный интерфейс, и через пользовательские веб-интерфейсы.

Для визуализации результатов моделирования и анализа используются традиционные библиотеки и фреймворки на базе python, такие как Plotly Dash, Matplotlib, Seaborn.

Заключение

Повсеместная цифровизация, датафикация и социальные медиа, в частности, как основной источник больших открытых данных, позволяют проводить их сбор, агрегацию и впоследствии анализировать и представлять результаты для принятия различных управленческих решений, в том числе для поддержки регионального развития.

В данной работе представлено концептуальное описание информационно-аналитической системы поддержки управления региональным развитием на основе открытых больших данных социальных медиа. Эти идеи поддержаны практическим опытом программной реализации ряда компонентов специализированного фреймворка и создания соответствующей системной и пользовательской инфраструктур. Все представленные разработки являются развитием известного практического опыта, но сконцентрированы на решении задач поддержки регионального развития на примере Мурманской области. Важные для региона социальные объекты и процессы находят свое отражение в виртуальном пространстве социальных сетей. Автоматизированная обработка больших открытых данных, выделение и моделирование в них этих сущностей позволяют повысить качество управленческих решений.

В настоящее время большие открытые данные представляют собой огромный интерес для исследователей. Появляется все больше различных подходов, технологий и инструментальных средств для их обработки и анализа. Авторы данной работы планируют развивать наработки в соответствии с новыми результатами своих исследований. Особый интерес представляет применение в разработках компонентов формирующего ИИ.

Список сокращений

PR (пиар — Public Relations) — публичные отношения, связи с общественностью, отношения с общественностью, общественные связи, общественное взаимодействие.

HR (Human resources — эйчар, «человеческие ресурсы») — персонал компании, сотрудники. Также сегодня «эйчаром» называются специалисты в области управления персоналом (менеджеры по персоналу, рекрутеры, специалисты по оплате труда, бизнес-тренеры).

СМИ (средства массовой информации) — совокупность органов публичной передачи информации с помощью технических средств.

SMM (Social Media Marketing) — использование социальных сетей в качестве каналов для продвижения бренда, увеличения целевой аудитории, привлечения трафика на сайт и повышения продаж.

ИИ (англ. Artificial intelligence, AI) — искусственный интеллект. Существует множество определений ИИ, поскольку понятие имеет философский оттенок. Наиболее популярные определения: 1) направление в информатике и информационных технологиях, задачей которого является воссоздание с помощью вычислительных систем и иных искусственных устройств разумных рассуждений и действий; 2) способность системы правильно интерпретировать внешние данные, извлекать знания из таких данных и использовать полученные знания для достижения конкретных целей и задач при помощи гибкой адаптации. Некоторые исследователи призывают расшифровывать ИИ более корректно — имитация интеллекта.

Umass (Юниверсити оф Массачусетс) — общепринятое обозначение, используется в данной работе для идентификации метрики когерентности, созданной именно в Университете Массачусетса, по причине существования альтернативной метрики когерентности Uci, созданной на базе Калифорнийского университета в Ирвайне (англ. University of California, Irvine).

TF-IDF (TF — term frequency, IDF — inverse document frequency) — статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов (корпуса). Вес некоторого слова пропорционален частоте употребления этого слова в документе и обратно пропорционален частоте употребления слова во всех документах коллекции.

PLSA (Probabilistic Latent Semantic Analysis — вероятностный латентный семантический анализ) — вероятностная тематическая модель представления текста на естественном языке. Модель называется латентной, так как предполагает введение скрытого (латентного) параметра — темы. Модель предложена Томасом Хофманном в 1999 г.

LDA (Latent Dirichlet allocation — латентное размещение Дирихле) — порождающая модель, метод тематического моделирования, впервые был представлен в качестве графовой модели для обнаружения тематик Дэвидом Блеем, Эндрю Ёном и Майклом Джорданом в 2003 г. Похож на PLSA, отличие заключается в том, что в LDA предполагается, что распределение тематик имеет в качестве априорного распределение Дирихле.

ARTM (Additive Regularization of Topic Models — аддитивная регуляризация тематических моделей) — математический подход тематического моделирования, основанный на максимизации взвешенной суммы логарифма правдоподобия и дополнительных критериев — регуляризаторов.

BigARTM — библиотека с открытым кодом для тематического моделирования больших коллекций текстовых документов и массивов транзакционных данных. Эффективная потоковая параллельная реализация вероятностного тематического моделирования на основе аддитивной регуляризации.

GSDMM (Gibbs Sampling Dirichlet Multinomial Mixture) — метод тематического моделирования, основан на предположении, что один документ посвящен только одной теме. Авторы этого метода (Цзяньхуа Инь и Цзяньюн Ван) утверждают, что он отлично (лучше, чем метод Vanilla LDA) работает с короткими текстами, такими как твиты и обзоры фильмов.

API (Application Programming Interface) — описание способов взаимодействия одной компьютерной программы с другими.

RPC (Remote Procedure Call — удалённый вызов процедур, иногда — вызов удалённых процедур) — класс технологий, позволяющих программам вызывать функции или процедуры в другом адресном пространстве (на удалённых узлах либо в независимой сторонней системе на том же узле).

REST (Representational State Transfer — дословно «передача репрезентативного состояния» или «передача “самоописываемого” состояния») — архитектурный стиль взаимодействия компонентов распределённого приложения в сети. Другими словами, REST — это набор правил для программиста: как организовать написание кода серверного приложения, чтобы упростить обмен данными между всеми системами и масштабирование приложения. REST является альтернативой RPC. В Интернете вызов удалённой процедуры может представлять собой обычный HTTP-запрос (обычно GET или POST; такой запрос называют REST-запросом), а необходимые данные передаются в качестве параметров запроса. Для веб-служб, построенных с учётом REST (то есть не нарушающих накладываемых им ограничений), применяют термин “RESTful”.

HTML (от англ. HyperText Markup Language — язык гипертекстовой разметки) — стандартизированный язык гипертекстовой разметки документов для просмотра веб-страниц в браузере.

HTTP (HyperText Transfer Protocol — протокол передачи гипертекста) — протокол прикладного уровня передачи данных, изначально — в виде гипертекстовых документов в формате HTML, в настоящее время используется для передачи произвольных данных.

JSON (JavaScript Object Notation) — текстовый формат обмена данными, основанный на JavaScript.

NoSQL — обозначение класса разнородных систем управления базами данных (СУБД), появившихся в конце 2000-х — начале 2010-х гг. и существенно отличающихся от традиционных реляционных СУБД с доступом к данным средствами языка SQL. Другими словами, это семейство СУБД, которые хранят данные в формате, отличном от реляционных таблиц.

MongoDB — документоориентированная СУБД, не требующая описания схемы таблиц. Считается одним из классических примеров NoSQL-систем, использует JSON-подобные документы и схему базы данных. Написана на языке программирования C++.

ГБ (гигабайт) — единица измерения количества информации.

WD (Western Digital Corporation, иногда WDC, по-русски — «Вестерн Диджитал») — компания, производитель компьютерной электроники.

HDD (hard (magnetic) disk drive, иногда — HMDD) — жёсткий диск, запоминающее устройство (накопитель) произвольного доступа, основанное на принципе магнитной записи.

DDR (Double Data Rate) — удвоенная скорость передачи данных. DDR SDRAM, DDR2 SDRAM, DDR3 SDRAM, DDR4 SDRAM и DDR5 SDRAM — типы компьютерной оперативной памяти.

ОС — операционная система, комплекс программ, предназначенных для управления ресурсами компьютера и организации взаимодействия с пользователем.

UNIX — семейство переносимых, многозадачных и многопользовательских операционных систем, которые основаны на идеях оригинального проекта AT & T Unix, разработанного в 1970-х гг. в исследовательском центре Bell Labs Кеном Томпсоном, Деннисом Ритчи и другими.

PyCharm — это кроссплатформенная интегрированная среда разработки для языка программирования Python.

MS VSCode (Microsoft Visual Studio Code) — редактор исходного кода, разработанный Microsoft для Windows, Linux и macOS. Позиционируется как «лёгкий» редактор кода для кроссплатформенной разработки веб- и облачных приложений.

JupyterLab — это интерактивная среда разработки для работы с блокнотами (Jupyter Notebook), сценариями Python, текстовыми редакторами и терминалами в рабочей области с вкладками. IDE (Integrated Development Environment — интегрированная или единая среда разработки) JupyterLab, включенная в IBM Watson Studio, содержит все строительные блоки для разработки интерактивных аналитических исследовательских вычислительных процедур с использованием Python.

Debian GNU/Linux — один из самых популярных и важных дистрибутивов GNU/Linux, в первичной форме оказавший значительное влияние на развитие этого типа ОС в целом.

SSH (Secure Shell — безопасная оболочка) — сетевой протокол прикладного уровня, позволяющий производить удалённое управление ОС и туннелирование TCP-соединений.

TCP (Transmission Control Protocol — протокол управления передачей) — один из основных протоколов передачи данных Интернета.

PuTTY — свободно распространяемый клиент для различных протоколов удалённого доступа, включая SSH, Telnet, rlogin.

Django — свободный фреймворк для веб-приложений на языке Python, использующий шаблон проектирования MVC.

MVC (Model-View-Controller) — схема разделения данных приложения и управляющей логики на три отдельных компонента (модель, представление и контроллер) таким образом, что модификация каждого компонента может осуществляться независимо. Модель предоставляет данные и реагирует на команды контроллера, изменяя своё состояние.

Robo 3T (ранее Robomongo) — кроссплатформенный инструмент управления MongoDB, ориентированный на оболочку. В отличие от большинства других пользовательских интерфейсов MongoDB, Robo 3T встраивает реальную оболочку mongo в интерфейс с вкладками с доступом к командной строке оболочки, а также взаимодействует с графическим интерфейсом.

cron — классический демон, использующийся для периодического выполнения заданий в определённое время. Регулярные действия описываются инструкциями, помещёнными в файлы crontab и специальные каталоги. Название cron образовано от греческого слова χρόνος — время.

Список источников

1. Вечкинзова Е. А. К вопросу об определении понятия «региональное развитие» // Вестник Саратовского государственного социально-экономического университета. 2010. № 5. С. 71–73. URL: <https://cyberleninka.ru/article/n/k-voprosu-ob-opredelenii-ponyatiya-regionalnoe-razvitiye> (дата обращения: 10.10.2022).

2. Bærenholdt, Jørgen Ole. Regional Development and Noneconomic Factors // *International Encyclopedia of Human Geography* / editors Rob Kitchin; Nigel Thrift. Vol. 9. Amsterdam: Pergamon Press, 2009. P. 181–186.
3. Сайт Федеральной службы государственной статистики. URL: <https://rosstat.gov.ru/>.
4. Алиева А. Х. Ключевые аспекты использования социальных сетей для бизнес-модели B2B // *Журнал прикладных исследований*. 2021. № 6. С. 719–724. URL: <https://cyberleninka.ru/article/n/klyuchevye-aspekty-ispolzovaniya-sotsialnyh-setey-dlya-biznes-modeli-b2b> (дата обращения: 18.10.2022).
5. Рослякова М. В. Социальные сети в профессиональной деятельности государственных служащих: российская практика и зарубежный опыт // *Социодинамика*. 2019. № 9. С. 82–99. DOI: 10.25136/2409-7144.2019.9.30298. URL: https://nbpublish.com/library_read_article.php?id=30298 (дата обращения: 18.10.2022).
6. Симачевский Р. Муниципальный SMM: как социальные сети помогают местному самоуправлению. URL: <https://vc.ru/marketing/198267-municipalnyy-smm-kak-socialnye-seti-pomogayut-mestnomu-samoupravleniyu> (дата обращения: 19.10.2022).
7. Alqudsi, Taghreed. Creative use of Social Media in the Revolutions of Tunisia, Egypt & Libya // *The International Journal of Interdisciplinary Social Sciences: Annual Review*. 2012. Vol. 6. P. 147–158. DOI: 10.18848/1833-1882/CGP/v06i06/52097
8. Губанов Д. А., Чхартишвили А. Г. Концептуальный подход к анализу онлайн-социальных сетей // *УБС*. 2013. № 45. С. 222–236. URL: <https://cyberleninka.ru/article/n/kontseptualnyy-podhod-k-analizu-onlaynovyh-sotsialnyh-setey> (дата обращения: 19.10.2022).
9. Безбогова М. С., Ионцева М. В. Социально-психологические аспекты взаимодействия пользователей в виртуальных социальных сетях // *Мир науки. Педагогика и психология*. 2016. № 5. URL: <https://cyberleninka.ru/article/n/sotsialno-psihologicheskie-aspekty-vzaimodeystviya-polzovateley-v-virtualnyh-sotsialnyh-setyah> (дата обращения: 24.10.2022).
10. Сайт «Лидеры Севера». URL: <https://лидеры-севера.рф/>.
11. Best Social Media Monitoring Software. URL: <https://www.g2.com/categories/social-media-monitoring>.
12. Рязанова Н. Ю., Сперцяк К. М. Сравнительный анализ методов определения эмоциональной окраски сообщений в социальных сетях с применением обучения с учителем // *Новые информационные технологии в автоматизированных системах*. 2018. № 21. С. 417–423. URL: <https://cyberleninka.ru/article/n/sravnitelnyy-analiz-metodov-opredeleniya-emotsionalnoy-okraski-soobscheniy-v-sotsialnyh-setyah-s-primeneniem-obucheniya-s-uchitelem> (дата обращения: 21.10.2022).
13. Богданов А. Л., Дуля И. С. Сентимент-анализ коротких русскоязычных текстов в социальных медиа // *Вестн. Том. гос. ун-та. Экономика*. 2019. № 47. С. 220–241. URL: <https://cyberleninka.ru/article/n/sentiment-analiz-korotkih-russkoyazychnyh-tekstov-v-sotsialnyh-media> (дата обращения: 21.10.2022).
14. Крибрум. URL: <https://www.ashmanov.com/tech-i-services/kribrum/>.
15. Медиалогия. URL: <https://medialogia.com/products/pr/>.
16. [М 13] Мониторинг и аналитика сообщений СМИ и социальных сетей. URL: <https://www.m13.ru/>.
17. Система мониторинга и анализа СМИ и соцмедиа // *СКАН-Интерфакс*. URL: <https://scan-interfax.ru/media-monitoring/>.
18. Babkee — система мониторинга социальных медиа и СМИ. URL: <https://www.babkee.ru>.
19. Exlibris. URL: <https://exlibris.ru/>.
20. Мониторинг СМИ и социальных сетей в режиме реального времени. URL: <https://pressindex.ru/>.
21. Интегрум. URL: <https://integrum.ru/>.
22. IQBuzz. URL: <http://iqbuzz.pro/>.
23. SemanticForce: Единая платформа мониторинга и анализа онлайн-медиа. URL: <https://semanticforce.ai/ru>.
24. Связь с общественностью // *Википедия*. URL: https://ru.wikipedia.org/wiki/Связи_с_общественностью.
25. Ren H., Yang Q. Predicting and Evaluating the Popularity of Online News // *Stanford University*. 2015. URL: https://cs229.stanford.edu/proj2015/328_report (дата обращения: 21.10.2022).

26. Femandes K., Vinagre P., Cortez P. A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News. 2015 // Pereira F., Machado P., Costa E., Cardoso A. (eds) Progress in Artificial Intelligence. EPIA 2015. Lecture Notes in Computer Science, vol. 9273. Springer, Cham.
27. Губанов Д. А., Новиков Д. А., Чхартишвили А. Г. Модели репутации и информационного управления в социальных сетях // Управление большими системами. 2009. № 26.1. С. 209–234.
28. Губанов Д. А., Новиков Д. А., Чхартишвили А. Г. Социальные сети: модели информационного влияния, управления и противоборства. М.: Физматлит, 2010. 244 с.
29. Губанов Д. А. Обзор онлайн-систем репутации / доверия // Интернет-конференция по проблемам управления / ИПУ РАН. М., 2009. URL: http://www.mtas.ru/bitrix/components/bitrix/forum.interface/show_file.php?fid=1671 (дата обращения: 21.10.2022).
30. Datyev I. O., Fedorov A. M., Shchur A. L. Framework for civic engagement analysis based on open social media data // Silhavy, R. (ed.): CSOC 2020. AISC, vol. 1225. Springer, Cham, 2020. P. 586–597. URL: https://doi.org/10.1007/978-3-030-51971-1_48 (дата обращения: 21.10.2022).
31. Свидетельство о государственной регистрации программы для ЭВМ № 2019661630. ИАС «Региональные профили социальной сети» / Рябов Д. В., Федоров А. М., Датьев И. О.; Роспатент. 04.09.2019.
32. Использование API | Быстрый старт | VK для разработчиков. URL: <https://dev.vk.com/api/getting-started>.
33. Свидетельство о государственной регистрации программы для ЭВМ № 2020619469. Информационная Система Мониторинга Сообществ ВКонтакте («ИС МСВ») / Федоров А. М., Датьев И. О., Щур А. Л.; Роспатент. 17.08.2020.
34. Свидетельство о государственной регистрации программы для ЭВМ № 2020619468. Информационно-Аналитическая Система Социальная-Сеть-Статистика (ИАС CCC) / Федоров А. М., Датьев И. О., Щур А. Л.; Роспатент. 17.08.2020.
35. Свидетельство о государственной регистрации программы для ЭВМ № 2021668445. Подсистема организации и проведения экспериментов тематического моделирования по заданному множеству параметров / Федоров А. М., Датьев И. О.; Роспатент. 16.11.2021.
36. Fedorov, A. M., Datyev I. O., Shchur A. L. Social Media Communities Topic Modeling // Silhavy R., Silhavy P., Prokopova Z. (eds.): Data Science and Intelligent Systems. CoMeSySo 2021. Lecture Notes in Networks and Systems, vol. 231. Springer, Cham, 2021. P. 605–614. URL: https://doi.org/10.1007/978-3-030-90321-3_50 (дата обращения: 21.10.2022).
37. Свидетельство о государственной регистрации программы для ЭВМ № 2021668652. Подсистема выбора оптимальной (по заданному критерию) модели тематического моделирования / Федоров А. М., Датьев И. О.; Роспатент. 18.11.2021.
38. Optimizing semantic coherence in topic models / D. Mimno, H. Wallach, Ed. Talley, M. Leenders, A. McCallum // Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Edinburgh, Scotland, UK. — Association of Computational Linguistics, 2011. P. 262–272.
39. Automatic evaluation of topic coherence / D. Newman, J. H. Lau, K. Grieser, T. Baldwin // Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT 2010). — Association for Computational Linguistics, Stroudsburg, 2010. P. 100–108.
40. Vorontsov K., Potapenko A. Additive regularization of topic models // Mach Learn. 2015. 101. P. 303–323. URL: <https://doi.org/10.1007/s10994-014-5476-6> (дата обращения: 21.10.2022).
41. Hofmann, T. Probabilistic latent semantic indexing // Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '99). — Association for Computing Machinery, New York, NY, USA, 1999. P. 50–57. URL: <https://doi.org/10.1145/312624.312649> (дата обращения: 21.10.2022).
42. Датьев И. О., Федоров А. М. Аддитивная регуляризация при тематическом моделировании текстов сообществ онлайн-социальных сетей // Онтология проектирования. 2022. Т. 12, № 2 (44). С. 186–199. DOI: 10.18287/2223-9537-2022-12-2-186-199
43. Blei D. M., Ng A. Y., Jordan M. I. Latent Dirichlet allocation // J. Mach. Learn. Res. 2003. № 3. P. 993–1022.

44. Jianhua Yin, Jianyong Wang. A dirichlet multinomial mixture model-based approach for short text clustering // Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '14). — Association for Computing Machinery, New York, NY, USA, 2014. P. 233–242. URL: <https://doi.org/10.1145/2623330.2623715> (дата обращения: 21.10.2022).
45. Главная // ГосВонь: сайт. URL: <https://gosvon.net/>.
46. Введение в REST API — RESTful веб-сервисы. URL: <https://habr.com/ru/post/483202/>.

References

1. Vechkinzova E. A. K voprosu ob opredelenii ponyatiya “regional'noe razvitie” [To the question of the definition of the concept of “regional development”]. *Vestnik Saratovskogo gosudarstvennogo social'no-ekonomicheskogo universiteta* [Bulletin of the Saratov State Socio-Economic University], 2010, no. 5, pp. 71–73. (In Russ.). Available at: <https://cyberleninka.ru/article/n/k-voprosu-ob-opredelenii-ponyatiya-regionalnoe-razvitie> (accessed 10.10.2022).
2. Bærenholdt Jørgen Ole. *Regional Development and Noneconomic Factors. International Encyclopedia of Human Geography*. Editors Rob Kitchin, Nigel Thrift. Vol. 9. Amsterdam, Pergamon Press, 2009, pp. 181–186.
3. <https://rosstat.gov.ru/>.
4. Alieva A. H. Klyuchevye aspekty ispol'zovaniya social'nyh setej dlya biznes-modeli B2B [Key aspects of using social media for a B2B business model]. *Zhurnal prikladnyh issledovaniy* [Journal of Applied Research], 2021, no. 6, pp. 719–724. (In Russ.). Available at: <https://cyberleninka.ru/article/n/klyuchevye-aspekty-ispolzovaniya-sotsialnyh-setey-dlya-biznes-modeli-b2b> (accessed 18.10.2022).
5. Roslyakova M. V. Social'nye seti v professional'noj deyatel'nosti gosudarstvennyh sluzhashchih: rossijskaya praktika i zarubezhnyy opyt [Social networks in the professional activities of civil servants: Russian practice and foreign experience]. *Sociodinamika* [Sociodynamics], 2019, no. 9, pp. 82–99. (In Russ.). DOI: 10.25136/2409-7144.2019.9.30298. Available at: https://nbpublish.com/library_read_article.php?id=30298 (accessed 18.10.2022).
6. Simachevskij R. Municipal'nyj SMM: kak social'nye seti pomagayut mestnomu samoupravleniyu. (In Russ.). Available at: <https://vc.ru/marketing/198267-municipalnyy-smm-kak-socialnye-seti-pomagayut-mestnomu-samoupravleniyu> (accessed 19.10.2022).
7. Alqudsi Taghreed. Creative use of Social Media in the Revolutions of Tunisia, Egypt & Libya. *The International Journal of Interdisciplinary Social Sciences: Annual Review*, 2012, vol. 6, pp. 147–158. DOI: 10.18848/1833-1882/CGP/v06i06/52097
8. Gubanov D. A., Chkhartishvili A. G. Konceptual'nyj podhod k analizu onlaynovykh social'nyh setej [A conceptual approach to the analysis of online social networks]. *Upravlenie bol'shimi sistemami* [Large-Scale Systems Control], 2013, no. 45, pp. 222–236. (In Russ.). Available at: <https://cyberleninka.ru/article/n/kontseptualnyy-podhod-k-analizu-onlaynovykh-sotsialnyh-setey> (accessed 19.10.2022).
9. Bezbogova M. S., Ionceva M. V. Social'no-psihologicheskie aspekty vzaimodejstviya pol'zovatelej v virtual'nykh social'nykh setyah [Socio-psychological aspects of user interaction in virtual social networks]. *Mir nauki. Pedagogika i psihologiya* [The World of Science. Pedagogy and Psychology], 2016, no. 5. (In Russ.). Available at: <https://cyberleninka.ru/article/n/sotsialno-psihologicheskie-aspekty-vzaimodeystviya-polzovateley-v-virtualnykh-sotsialnykh-setyah> (accessed 24.10.2022).
10. <https://лидеры-севера.рф/>.
11. <https://www.g2.com/categories/social-media-monitoring>.
12. Ryazanova N. Yu., Spercyan K. M. Sravnitel'nyj analiz metodov opredeleniya emocional'noj okraski soobshchenij v social'nykh setyah s primeneniem obucheniya s uchitelem [Comparative analysis of methods for determining the emotional coloring of messages in social networks using learning with a teacher]. *Novye informacionnye tekhnologii v avtomatizirovannykh sistemah* [New Information Technologies in Automated Systems], 2018, no. 21, pp. 417–423. (In Russ.). Available at: <https://cyberleninka.ru/article/n/sravnitelnyy-analiz-metodov-opredeleniya-emotsionalnoy-okraski-soobscheniy-v-sotsialnykh-setyah-s-primeneniem-obucheniya-s-uchitelem> (accessed 21.10.2022).

13. Bogdanov A. L., Dulya I. S. Sentiment-analiz korotkih russkoyazychnyh tekstov v social'nyh media [Sentiment analysis of short Russian-language texts in social media]. *Vestn. Tom. gos. un-ta. Ekonomika* [Bulletin of Tomsk State University. Economy], 2019, no. 47, pp. 220–241. (In Russ.). Available at: <https://cyberleninka.ru/article/n/sentiment-analiz-korotkih-russkoyazychnyh-tekstov-v-sotsialnyh-media> (accessed 21.10.2022).
14. <https://www.ashmanov.com/tech-i-services/kribrum/>.
15. <https://medialogia.com/products/pr/>.
16. <https://www.m13.ru/>.
17. <https://scan-interfax.ru/media-monitoring/>.
18. <https://www.babkee.ru>.
19. <https://exlibris.ru/>.
20. <https://pressindex.ru/>.
21. <https://integrum.ru/>.
22. <http://iqbuzz.pro/>.
23. <https://semanticforce.ai/ru>.
24. https://ru.wikipedia.org/wiki/Связи_с_общественностью.
25. Ren H., Yang Q. *Predicting and Evaluating the Popularity of Online News*. Stanford University, 2015. Available at: https://cs229.stanford.edu › proj2015 › 328_report (accessed 21.10.2022).
26. Fernandes K., Vinagre P., Cortez P. A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News. 2015. In: Pereira F., Machado P., Costa E., Cardoso A. (eds). *Progress in Artificial Intelligence. EPIA 2015. Lecture Notes in Computer Science*, vol. 9273. Springer, Cham.
27. Gubanov D. A., Novikov D. A., Chkhartishvili A. G. Modeli reputacii i informacionnogo upravleniya v social'nyh setyah [Models of reputation and information management in social networks]. *Upravlenie bol'shimi sistemami* [Large System Management], 2009, no. 26.1, pp. 209–234. (In Russ.).
28. Gubanov D. A., Novikov D. A., Chkhartishvili A. G. *Social'nye seti: modeli informacionnogo vliyaniya, upravleniya i protivoborstva* [Social networks: models of information influence, control and confrontation]. Moscow, Fizmatlit, 2010, 244 p. (In Russ.).
29. Gubanov D. A. Obzor onlajnovykh sistem reputacii / doveriya [Overview of Online Reputation / Trust Systems]. *Internet-konferenciya po problemam upravleniya, IPU RAN* [Internet conference on control problems, Institute of Control Sciences V. A. Trapeznikov Academy of Sciences]. Moscow, 2009. (In Russ.). Available at: http://www.mtas.ru/bitrix/components/bitrix/forum.interface/show_file.php?fid=1671 (accessed 21.10.2022).
30. Datyev I. O., Fedorov A. M., Shchur A. L. Framework for civic engagement analysis based on open social media data. In: Silhavy R. (ed.): CSOC 2020. AISC, vol. 1225, Springer, Cham, 2020, pp. 586–597. Available at: https://doi.org/10.1007/978-3-030-51971-1_48 (accessed 21.10.2022).
31. Ryabov D. V., Fedorov A. M., Datyev I. O. IAS “Regional profiles of the social network”. The Certificate on official registration of the computer program. No. 2019661630, 2019. (In Russ.).
32. <https://dev.vk.com/api/getting-started>.
33. Fedorov A. M., Datyev I. O., Shchur A. L. VKontakte Community Monitoring Information System (“MSV IS”). The Certificate on official registration of the computer program. No. 2020619469, 2020. (In Russ.).
34. Fedorov A. M., Datyev I. O., Shchur A. L. Information-Analytical System Social-Network-Statistics (IAS SSS). The Certificate on official registration of the computer program. No. 2020619468, 2020. (In Russ.).
35. Fedorov A. M., Datyev I. O. Subsystem for organizing and conducting experiments of topic modeling for a given set of parameters. The Certificate on official registration of the computer program. No. 2021668445, 2021. (In Russ.).
36. Fedorov A. M., Datyev I. O., Shchur A. L. Social Media Communities Topic Modeling. In: Silhavy R., Silhavy P., Prokopova Z. (eds.): *Data Science and Intelligent Systems. CoMeSySo 2021. Lecture Notes in Networks and Systems*, vol. 231. Springer, Cham, 2021, pp. 605–614. Available at: https://doi.org/10.1007/978-3-030-90321-3_50 (accessed 21.10.2022).

37. Fedorov A. M., Datyev I. O. Subsystem for selecting the optimal (according to a given criterion) model of topic modeling. The Certificate on official registration of the computer program. No. 202166865, 2021. (In Russ.).
38. Mimno D., Wallach H., Talley Ed., Leenders M., McCallum A. Optimizing semantic coherence in topic models. In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland, UK. — Association of Computational Linguistics, 2011, pp. 262–272.
39. Newman D., Lau J. H., Grieser K., Baldwin T. Automatic evaluation of topic coherence. In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT 2010)*. — Association for Computational Linguistics, Stroudsburg, 2010, pp. 100–108.
40. Vorontsov, K., Potapenko, A. Additive regularization of topic models. *Mach Learn*, 2015, no. 101, pp. 303–323. Available at: <https://doi.org/10.1007/s10994-014-5476-6> (accessed 21.10.2022).
41. Hofmann T. Probabilistic latent semantic indexing. In: *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '99)*. — Association for Computing Machinery, New York, NY, USA, 1999, pp. 50–57. Available at: <https://doi.org/10.1145/312624.312649> (accessed 21.10.2022).
42. Datyev I. O., Fedorov A. M. Additivnaya regulyariatsiya pri tematicheskoy modelirovaniy tekстов soobshchestv onlaynovykh social'nykh setey [Additive Regularization in Topic Modeling of Texts of Communities of Online Social Networks]. *Ontologiya proektirovaniya* [Ontology of Designing], 2022, no. 2 (44), pp. 186–199. (In Russ.). DOI: 10.18287/2223-9537-2022-12-2-186-199
43. Blei D. M., Ng A. Y., Jordan M. I. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 2003, no. 3, pp. 993–1022.
44. Jianhua Yin, Jianyong Wang. A dirichlet multinomial mixture model-based approach for short text clustering. In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '14)*. Association for Computing Machinery, New York, NY, USA, 2014, pp. 233–242. Available at: <https://doi.org/10.1145/2623330.2623715> (accessed 21.10.2022).
45. <https://gosvon.net/>.
46. <https://habr.com/ru/post/483202/>.

Информация об авторах

Андрей Михайлович Федоров — кандидат технических наук, ведущий научный сотрудник;

Игорь Олегович Датьев — кандидат технических наук, старший научный сотрудник;

Максим Геннадьевич Шишаев — доктор технических наук, профессор, главный научный сотрудник;

Сергей Сергеевич Федотов — программист, аспирант;

Иван Геннадьевич Вишняков — системный администратор, магистрант.

Information about the authors

Andrey M. Fedorov — Candidate of Science (Tech.), Leading Researcher;

Igor O. Datyev — Candidate of Science (Tech.), Senior Researcher;

Maksim G. Shishaev — Doctor of Science (Tech.), Chief Research Fellow;

Sergey S. Fedotov — Programmer, graduate student;

Ivan G. Vishnyakov — System Administrator, undergraduate.

Статья поступила в редакцию 15.10.2022; одобрена после рецензирования 29.10.2022; принята к публикации 08.11.2022.

The article was submitted 15.10.2022; approved after reviewing 29.10.2022; accepted for publication 08.11.2022.