УДК 004.8:303.72
DOI:10.25729/ESI.2026.41.1.010

Цифровой портрет регионального развития: анализ грантовых проектов методами LDA и BERTopic

Митрофанова Татьяна Валерьевна, Христофорова Анастасия Владимировна

Чувашский государственный университет им. И.Н. Ульянова,
Россия, Чебоксары, *mitrofanova_tv@mail.ru*

Аннотация. В статье представлен подход к построению цифрового портрета социально-экономического развития региона на основе автоматизированного анализа текстовых данных. В качестве ключевого индикатора рассматриваются описания грантовых проектов, поддержанных в Республике Марий Эл за 2023–2025 годы. С помощью комбинации методов тематического моделирования – классического LDA и современного нейросетевого BERTopic – выявлены и визуализированы латентные тематические паттерны, формирующие актуальную повестку развития территории. Полученный «портрет» позволяет объективно оценить структуру гражданских инициатив, выявить доминирующие направления (социальная поддержка, развитие человеческого капитала, спорт, образование) и определить их соответствие стратегическим целям региона.

В результате исследования установлено, что смыслообразующим ядром грантовой повестки выступает целевая группа «дети», вокруг которой концентрируются кластеры образовательных, реабилитационных и инклюзивных проектов. LDA-моделирование выделило пять устойчивых тематических направлений, в том числе поддержку семей с детьми и адаптацию лиц с ограниченными возможностями здоровья. BERTopic, в свою очередь, позволил детализировать узкие нишевые практики, например, инклюзивный спорт, реабилитацию людей с нарушениями зрения и культурно-досуговые программы для детей с ОВЗ (ограниченными возможностями здоровья), что подтверждает высокую семантическую чувствительность нейросетевого подхода. Сопоставление полученных тематических кластеров с проектом Стратегии социально-экономического развития Республики Марий Эл выявило как зоны полного совпадения приоритетов (поддержка детства, инклюзия), так и стратегические дисбалансы: слабую представленность проектов в сферах креативных индустрий, туризма и работы с молодёжью. Предложенная методика демонстрирует, что методы машинного обучения открывают новые возможности для мониторинга и диагностики состояния региональных и муниципальных систем, обеспечивая управленческие команды Data-Driven-инструментом для принятия решений и коррекции грантовой политики. Таким образом, тематическое моделирование позволяет перевести массив неструктурированных текстовых данных в плоскость объективного анализа, превращая описание гражданских инициатив в надежный инструмент выявления реальных социальных приоритетов и основу для стратегического планирования.

Ключевые слова: цифровой портрет региона, тематическое моделирование, LDA, BERTopic, управление на основе данных, анализ грантов, региональное развитие, машинное обучение, обработка естественного языка (NLP), Республика Марий Эл

Цитирование: Митрофанова Т.В. Цифровой портрет регионального развития: анализ грантовых проектов методами LDA и BERTopic / Т.В. Митрофанова, А.В. Христофорова // Информационные и математические технологии в науке и управлении, 2026. – № 1(41). – С. 136-149 – DOI:10.25729/ESI.2026.41.1.010.

Введение. Современная парадигма управления региональным развитием все в большей степени опирается на данные и цифровые инструменты, позволяющие перейти от интуитивных оценок к объективному, измеримому анализу состояния территории. В этом контексте формирование комплексного «цифрового портрета» региона, отражающего ключевые тенденции и приоритеты его развития, становится критически важной задачей. Одним из наиболее релевантных, но пока недостаточно изученных источников для построения такого портрета являются текстовые данные, генерируемые в рамках гражданской и проектной активности. В частности, описания грантовых проектов, получающих государственную поддержку, содержат в себе концентрированное выражение актуальных общественных запросов и официальных приоритетов.

Таким образом, цифровой портрет регионального развития представляет собой многомерную тематическую модель социально-экономического состояния территории,

построенную на основе семантического анализа текстовых данных о гражданской и проектной активности. В представляемой работе предлагается методика построения цифрового портрета социально-экономического развития на примере Республики Марий Эл через призму анализа грантовых проектов. В качестве инструментария применяется комбинация двух методов тематического моделирования – латентного размещения Дирихле (LDA) и нейросетевого подхода BERTopic. LDA, будучи классическим и интерпретируемым методом, позволяет выявить устойчивые тематические кластеры в корпусе текстов описаний проектов. BERTopic, использующий современные языковые модели, обеспечивает более глубокое семантическое понимание контекста и позволяет выявить более тонкие и специфичные темы. Синтез этих подходов обеспечивает создание многомерной и достоверной тематической карты региона.

Цель исследования – сформировать и апробировать методику построения цифрового портрета региона на основе текстового анализа данных грантовой поддержки с использованием методов машинного обучения. Для достижения этой цели решаются следующие задачи:

- выявить латентные тематические структуры в описаниях проектов;
- визуализировать полученные результаты для наглядной интерпретации;
- определить практическую ценность подхода для органов управления региональными и муниципальными системами.

Дополнительной задачей исследования является оценка адекватности полученного цифрового портрета регионального развития. Под адекватностью понимается соответствие выявленных тематических кластеров стратегическим документам региона, приоритетам государственной политики в социальной сфере и фактической структуре распределения грантовой поддержки. Валидность цифрового портрета обеспечивается репрезентативностью данных, применением интерпретируемых моделей тематического анализа и сопоставлением результатов с официальными направлениями развития региона.

Обзор литературы. Методы тематического моделирования стали незаменимыми инструментами для извлечения тематической информации из больших текстовых массивов, особенно в таких областях, как «цифровые гуманитарные науки» (Digital Humanities) и политический анализ [1, 2]. Эти подходы к неконтролируемому машинному обучению облегчают систематическую организацию, понимание и извлечение значимых шаблонов из огромных объемов неструктурированных текстовых данных, где ручной анализ часто непрактичен и неосуществим [3]. Необходимость в таких автоматизированных процедурах обусловлена экспоненциальным ростом объема цифровой информации, что требует надежных алгоритмов, способных распознавать, классифицировать и организовывать информацию, имитируя возможности человека в извлечении и обобщении сложных текстовых наборов данных [2]. Среди известных методов латентное размещение Дирихле и двунаправленные представления кодировщика на основе трансформеров BERT выделяются своими отличными подходами к выявлению скрытых тем в документах. Автор [4] проводит сравнительный анализ современных алгоритмов тематического моделирования коротких текстов, включая LDA, NMF, ETM, ProdLDA, CTM, SBMTM и BERTopic, и показывает, что традиционные вероятностные методы хуже работают на коротких сообщениях из-за ограниченного контекста, и обосновывает эффективность сетевых и нейросетевых подходов. LDA, генеративная статистическая модель, предполагает, что документы представляют собой смесь различных тем, каждая из которых характеризуется определенным распределением слов, что позволяет идентифицировать преобладающие темы с помощью шаблонов совпадения слов [5]. BERTopic, наоборот, использует контекстуальные вложения для группировки семантически схожих документов, предлагая более детальное понимание базовых тематических структур за

счет интеграции предварительно подготовленных контекстуальных представлений [6, 7]. Это различие позволяет BERTopic улавливать более сложные семантические взаимосвязи по сравнению с предположением LDA о наборе слов, что может быть особенно полезно в сложных и развивающихся областях, таких, как региональное развитие [1].

В общем виде цифровой профиль региона понимается, как совокупность его типологических особенностей в цифровой сфере, определяющих количественные и качественные характеристики цифровых ресурсов и цифровых потребностей [8]. В статье А.А. Былинской [9] рассматриваются методологические подходы к формированию цифрового профиля российских регионов на основе оценки цифрового спроса и предложения. Особое внимание уделено проблеме цифрового неравенства: различия между регионами объясняются преимущественно уровнем человеческого капитала и стимулирующими политиками, а не экономическими ресурсами. В статье [10] предложен метод выделения ключевых факторов среди экономических, социальных и цифровых процессов российских регионов, позволяющий выявить основные тенденции их развития. Автор акцентирует внимание на необходимости комплексного подхода к анализу региональных характеристик, учитывающего взаимодействие различных факторов.

Как показали в исследовании авторы [11], применение методов корреляционно-регрессионного анализа позволяет выявить ключевые факторы, определяющие динамику валового регионального продукта (ВРП), и построить высококачественные прогнозные модели для субъектов Центрального федерального округа.

В работе [12] демонстрируется создание информационной системы для анализа социально-экономического развития, что требует решения задач сбора, нормализации и агрегации данных из множества официальных и экспертных источников, включая Федеральную службу государственной статистики, ВЦИОМ, рейтинги Ассоциации инновационных регионов России и материалы исследовательских проектов. Авторами был реализован ETL-процесс, обеспечивающий приведение данных к единому стандарту и их последующее использование для моделирования. В контексте нашего исследования, посвященного построению цифрового портрета регионального развития Республики Марий Эл, аналогичный подход применяется для формирования корпуса текстовых данных из описаний грантовых проектов.

Предложенная авторами [13] методика основана на индексном методе и аддитивных моделях и позволяет проводить сравнительный анализ регионов по уровню внедрения «сквозных» цифровых технологий. Однако подобные системы оценки опираются преимущественно на количественные статистические показатели, нормативно закрепленные в стратегиях цифровой трансформации.

В работе [14] показано, что «важным инструментом анализа является применение... отзывов предпринимателей об условиях ведения бизнеса в регионе», а также подчеркивают важность «активности местных органов власти» и «поддержки малого и среднего бизнеса».

Анализ грантов позволяет перейти от внешней оценки к внутренней диагностике, построив цифровой портрет Марий Эл не через призму статистических показателей, а через призму его собственных стратегических устремлений и проектных инициатив.

Методология. В качестве источника данных был подготовлен датасет с информацией о проектах-победителях грантовых конкурсов в Республике Марий Эл за период, начиная с первого грантового конкурса 2023 года, по второй конкурс 2025 года. Данные представлены в формате Microsoft Excel и включают сведения о проектах, поддержанных Фондом президентских грантов (таблица 1).

Таблица 1. Основные поля датасета

Поле	Описание
Название проекта	Полное название поддержанного проекта
Организация	Организация, получившая финансирование
Сумма гранта	Размер предоставленного гранта в рублях
Год	Год проведения конкурса (2023, 2024, 2025)
Тематика	Тематическое направление проекта
Краткое описание	Краткое описание содержания проекта

Объектом исследования выступает корпус текстов, содержащих краткие описания проектов-победителей региональных грантовых конкурсов в Республике Марий Эл за 2023–2025 годы. Исходные данные были собраны в виде таблицы Excel и включают информацию о названии проекта, его кратком описании, годе проведения, размере поддержки и тематике.

Для анализа текстов использовались методы обработки естественного языка, включая очистку текстов от стоп-слов и пунктуации, нормализацию (приведение к нижнему регистру), векторизацию с помощью CountVectorizer.

Для лингвистической нормализации текстов использовалась лемматизация с применением библиотеки Stanza. Выбор был обусловлен тем, что, в отличие от простых словарных подходов pymorphy2, Mystem, Stanza использует обученные нейросетевые модели, демонстрирующие более высокую точность в разрешении морфологических амбиграмм, особенно в контексте социальных и проектных текстов.

Применяется комбинированный подход к тематическому моделированию, сочетающий классический метод LDA и современный нейросетевой подход BERTopic. Такой синтез методов позволяет максимально полно выявить тематическую структуру описаний грантовых проектов. LDA был выбран благодаря своей прозрачности, хорошей интерпретируемости результатов и удобству визуализации с использованием pyLDAvis. В то же время BERTopic, основанный на трансформерных моделях, обеспечивает более глубокий семантический анализ и лучше учитывает контекстные взаимосвязи между словами.

Для проверки адекватности цифрового портрета была проведена комплексная оценка, включающая сопоставление выделенных тематических кластеров с официальными направлениями Стратегии социально-экономического развития Республики Марий Эл [15], анализ согласованности результатов, полученных с использованием моделей LDA и BERTopic, а также количественную оценку долей тематических кластеров. Дополнительно была выполнена экспертная интерпретация выделенных тем, основанная на анализе практики реализации социально-ориентированных проектов в регионе, что позволило подтвердить содержательную релевантность полученных результатов.

Результаты. В результате тематического моделирования было выявлено пять основных тем, каждая из которых характеризуется набором ключевых слов, отражающих содержание соответствующей группы проектов (Таблица 2).

Таблица 2. Выявленные темы в результате LDA

№	Тема	Ключевые слова	Интерпретация
1.	Индивидуальные занятия и развитие детей	ребенок, занятие, провести, развитие, специалист, коррекция, индивидуальный, бесплатный, курс	Тема отражает направления деятельности, связанные с организацией индивидуальных и коррекционных занятий, направленных на развитие детей. В центре внимания – работа специалистов, проведение бесплатных курсов и развитие персонализированных программ обучения. Подобные практики характерны для образовательных и социально-реабилита-

			ционных учреждений, ориентированных на поддержку детей с особыми образова-тельными потребностями.
2.	Образовательные и развивающие мероприятия для детей	ребенок, развитие, занятие, навык, человек, марий, мероприятие, участие, организация	Тематика объединяет широкий спектр мероприятий, направленных на формирование и развитие когнитивных, коммуникативных и социальных навыков у детей. Особое место занимают развивающие занятия и образовательные инициативы, реализуемые как на уровне отдельных организаций, так и в рамках региональных программ. Акцент делается на участии различных социальных субъектов и активном вовлечении детей в образовательный процесс.
3.	Мероприятия и поддержка людей с инвалидностью	инвалид, мероприятие, зрение, республика, марий, проведение, физический, общество	Тематика характеризуется упором на проведение мероприятий, ориентированных на людей с инвалидностью, включая лиц с нарушениями зрения и другими ограничениями по здоровью. Особое внимание уделяется деятельности общественных организаций и республи-канских структур, которые обеспечивают проведение специализированных меропри-ятий и развитие инклюзивной среды. Это свидетельствует о формировании устойчивой системы поддержки данной категории населения.
4.	Социальная поддержка семей с детьми	ребенок, семья, помощь, социальный, республика, поддержка, родитель, реализация	Данная тема охватывает различные формы социальной помощи и поддержки, предоставляемые семьям с детьми. В неё входят инициативы по оказанию психологической, материальной и организационной помощи, а также меры государственной и региональной политики. Тематика подчёркивает важность участия семьи в программах поддержки и реализацию комплексного подхода к обеспечению благополучия детей.
5.	Игровые и творческие программы для детей с ОВЗ	игра, овз, команда, участник, количество, творческий, мероприятие, город	Тематика отражает организацию игровых, творческих и командных мероприятий, адаптированных для детей с ограниченными возможностями здоровья. Подобные практики способствуют развитию социальных навыков, раскрытию творческого потенциала и интеграции детей в коллектив. Проведение таких мероприятий, как правило, осуществляется на уровне муниципальных и общественных инициатив, что свидетельствует о расширении инклюзивных форм работы.

Визуализация результатов тематического моделирования с использованием библиотеки pyLDAvis (рисунок 1) позволила более детально интерпретировать структуру полученных тем и их взаимосвязи. На карте межтематических расстояний (Intertopic Distance Map) каждая тема представлена в виде круга, площадь которого пропорциональна доле соответствующей темы в корпусе текстов.

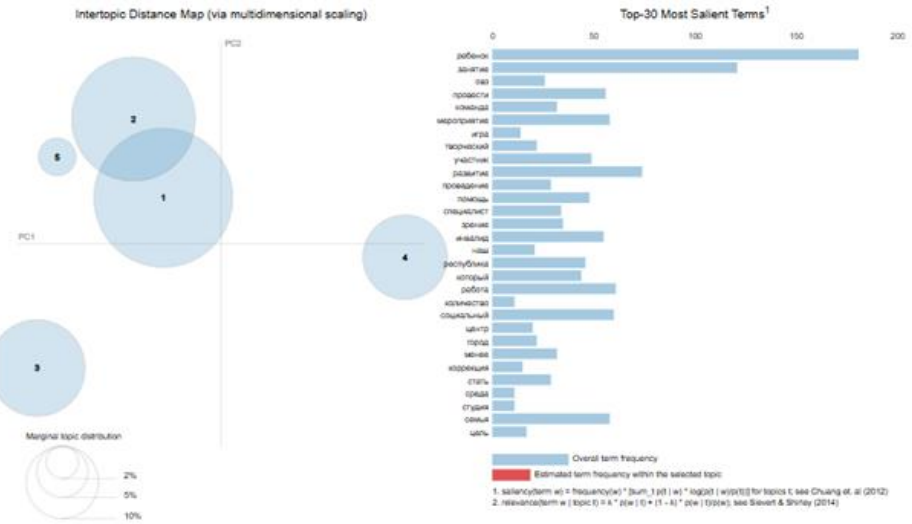

Рис. 1. Визуализация результатов тематического моделирования LDA

Были выявлены 5 основных тем, различающихся по объёму и степени взаимосвязи. Темы 1, 2 демонстрируют частичное пересечение, что свидетельствует о наличии общих лексических маркеров, отражающих родственные направления в описаниях проектов. Тема 5 занимает промежуточное положение и содержит как универсальные, так и уникальные термины. Темы 3 и 4 отчётливо выделяются, занимая обособленную позицию на карте, что указывает на их семантическую уникальность и тематическую специфичность.

На правой панели визуализированы 30 наиболее значимых (salient) терминов, определённых с учётом их частоты и различимости между темами. «ребёнок», «занятие», «развитие», «мероприятие», «инвалид», «помощь», «семья», «игра», «ОВЗ», «специалист». Данный набор лексем отражает направленность проектов преимущественно на развитие детей, поддержку семей, реализацию инклюзивных практик и организацию образовательных и социальных мероприятий.

Для углубленного анализа тематической структуры был дополнительно применён метод BERTopic, основанный на трансформерных языковых моделях (рисунок 2).

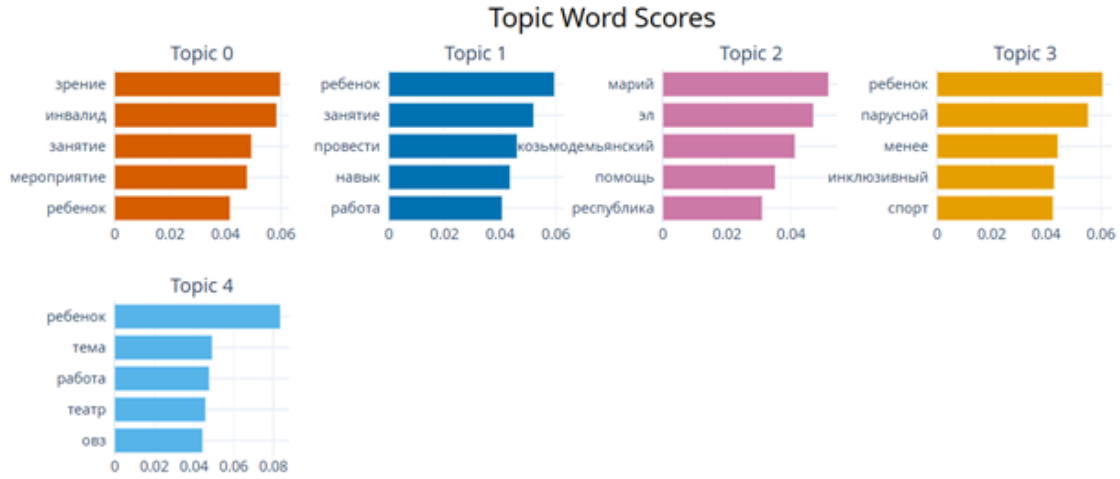

Рис. 2. Визуализация результатов тематического моделирования BERTopic

Темы (таблица 3), выделенные в результате моделирования BERTopic, обладают интерпретируемостью и логической связностью. Лексема «ребёнок» выполняет роль смыслообразующего ядра всего семантического пространства. Её повторяемость не является

статистическим шумом, а, напротив, чётко указывает на основную целевую группу, для которой выстраиваются все описанные социальные, реабилитационные и образовательные практики. Это позволяет классифицировать весь корпус текстов, как дискурс «социально-педагогической и реабилитационной поддержки детей, в особенности с ограниченными возможностями здоровья (ОВЗ)».

Таблица 3. Выделенные темы и соответствующие ключевые слова

№	Ключевые слова	Интерпретация для темы
1.	зрение, инвалид, занятие, мероприятие, ребенок	Социальная и реабилитационная поддержка лиц с нарушениями зрения
2.	ребёнок, тема, работа, театр, овз	Инклюзивная культурно-досуговая деятельность для детей с ОВЗ
3.	ребёнок, занятие, провести, навык, работа	Формирование компетенций у детей через организованную деятельность
4.	марий, эл, козьмодемьянский, помощь, республика	Региональный аспект социальной поддержки
5.	ребёнок, парусной, менее, инклюзивный, спорт	Инклюзивные практики в детском спорте

Таким образом, тематическое моделирование позволило структурировать содержательное пространство грантовых проектов и выделить доминирующие направления, что даёт основание говорить о приоритетах в региональной поддержке в социальной сфере.

Для повышения надёжности результатов тематического моделирования текстов описаний социальных проектов были применены два метода – BERTopic и LDA. Оба алгоритма позволили выявить релевантные тематические кластеры, однако в силу различий в подходах к построению тем наблюдаются как совпадения, так и расхождения.

Анализ текстового корпуса с помощью LDA и BERTopic выявил как общие, так и специфические тематические кластеры. Оба метода фиксируют центральные направления деятельности: развитие и обучение детей, инклюзивные практики для детей с ОВЗ, социальную поддержку семей и людей с инвалидностью. Повторяются ключевые слова, отражающие эти темы: «ребёнок», «занятие», «мероприятие», «инвалид/ОВЗ», «развитие», «поддержка».

При этом LDA выделяет более широкие, обобщённые направления, например, «Образовательные мероприятия» или «Социальная поддержка семей», концентрируясь на функциональных аспектах деятельности. BERTopic, напротив, позволяет выделять более узкие и контекстно насыщенные кластеры: культурно-досуговая деятельность, спортивные практики, конкретные виды инвалидности или региональные инициативы.

Таким образом, LDA показывает стратегическую «карту» тем корпуса, а BERTopic добавляет детализацию, позволяя увидеть конкретные формы реализации программ. Совместное использование обеих моделей обеспечивает многослойное понимание содержания текстов и полезно для аналитики и планирования социальных и образовательных инициатив.

Распределение тематик грантов по официальным направлениям Фонда президентских грантов подтверждает выявленную в результате тематического моделирования доминанту проектов, связанных с поддержкой детства, семьи, социальной защиты и здоровья (таблица 4). Согласно данным корпуса (34 проекта за 2023–2025 гг.), наибольший удельный вес приходился на социальную поддержку населения (38,2%), охрану здоровья и пропаганду здорового образа жизни (20,6%), а также поддержку семей и детей (11,8%). Эти значения

эмпирически согласуются с тематическими кластерами, выделенными моделью BERTopic/LDA, что является свидетельством адекватности цифрового портрета региона.

Таблица 4. Распределения проектов по тематикам грантов

Тематика	Кол-во проектов	Доля (%)
Социальное обслуживание, социальная поддержка и защита граждан	13	38,2%
Охрана здоровья, пропаганда ЗОЖ	7	20,6%
Поддержка семьи, материнства, отцовства и детства	4	11,8%
Поддержка молодежных инициатив	2	5,9%
Охрана окружающей среды и защита животных	2	5,9%
Защита прав и свобод граждан	2	5,9%
Наука, образование, просвещение	2	5,9%
Сохранение исторической памяти	2	5,9%

Финансовое распределение также подтверждает выявленные тематические доминанты (таблица 5). Наибольший средний объём финансирования характерен для проектов в сфере образования и просвещения и сохранения исторической памяти, тогда как наибольший суммарный объём поддержки наблюдается в области социальной защиты населения. Это согласуется с тематическими акцентами, полученными при помощи LDA и BERTopic, что служит дополнительным подтверждением валидности сформированного цифрового портрета региона.

Таблица 5. Объем финансирования грантов по направлениям

Тематика	Кол-во проектов	Средний размер гранта (руб.)	Общий объём финансирования (руб.)
Поддержка проектов в области науки, образования и просвещения	2	4 511 424	9 022 848
Сохранение исторической памяти	2	3 131 201	6 262 402
Поддержка семьи, материнства, отцовства и детства	4	2 426 548	9 706 191
Социальное обслуживание, социальная поддержка и защита граждан	13	2 161 744	28 102 678
Поддержка молодежных проектов	2	1 479 025	2 958 050
Охрана здоровья и пропаганда ЗОЖ	7	1 895 493	13 268 453
Защита прав и свобод граждан	2	999 845	1 999 690
Охрана окружающей среды и защита животных	2	718 688	1 437 376

В свою очередь, относительно более низкий средний объём грантов в сферах экологии, правозащитной деятельности и молодёжных инициатив отражает сравнительно меньший масштаб заявляемых программ. Таким образом, структурные и финансовые параметры грантовой поддержки демонстрируют высокую согласованность с тематическими результатами моделирования, что подтверждает достоверность и практическую значимость сформированного цифрового портрета региона.

Обсуждение. Выявленная тематическая структура (таблицы 2 и 3) однозначно указывает на то, что доминирующим фокусом гражданских инициатив и, как следствие, государственной поддержки в Республике Марий Эл является сфера детства. Полученный цифровой портрет рисует регион, активно инвестирующий в развитие человеческого капитала через поддержку образования, социальной защиты и инклюзии. Такая структура грантовой повестки в целом соответствует общероссийским стратегическим приоритетам в области демографии, семьи и детства. Полученные результаты согласуются с социально-демографическим профилем Республики Марий Эл, для которой характерна значительная доля детей и семей с детьми (по данным Маристат доля детей в общей численности населения региона составила 21,3% на начало 2025 года). Доминирование тем, связанных с развитием детей, образованием и инклюзией, соответствовало целям национальных проектов до 2024 года «Демография» и «Образование», а на 2025 год – национальным проектам «Семья» и «Молодежь и дети», а также направлениям, закреплённым в Стратегии социально-экономического развития Республики Марий Эл до 2030 года [15]. Согласованность результатов двух независимых методов тематического моделирования (LDA и BERTopic) подтверждает устойчивость и достоверность выявленных тематических кластеров.

Синтез LDA и BERTopic оказался не просто техническим приемом, а методологически обоснованным решением. LDA, как ожидалось, предоставил устойчивую и хорошо интерпретируемую «карту» крупных тематических направлений («Социальная поддержка семей», «Образовательные мероприятия»), что ценно для стратегического управления и выявления макроприоритетов. В свою очередь, BERTopic позволил погрузиться вглубь этих направлений, выявив специфические ниши и практики: например, реабилитацию людей с нарушениями зрения, инклюзивный театр или парусный спорт для детей с ОВЗ. Это позволяет органам власти не только видеть общие тренды, но и идентифицировать конкретные, точечные инициативы, заслуживающие тиражирования или дополнительной поддержки.

Предложенная методика позволяет в режиме, близком к реальному времени, отслеживать смещение акцентов в гражданской активности и оценивать соответствие распределения грантовых средств заявленным стратегическим целям региона. Выявленный дисбаланс в пользу поддержки детей (при возможном недостатке проектов для молодежи, старшего поколения или развития инфраструктуры) может стать основанием для корректировки конкурсной документации региональных грантов и информационной работы с НКО (некоммерческими организациями).

Полученный в ходе исследования «цифровой портрет» грантовой повестки позволяет не только зафиксировать актуальные направления, но и выявить ключевые дисбалансы, требующие стратегического внимания. В частности, наше предположение о недостаточном фокусе грантовых проектов на молодежной повестке находит полное подтверждение в положениях проекта Стратегии социально-экономического развития Марий Эл до 2036 года [16]. В данном документе молодежная политика названа одним из трех ключевых социальных приоритетов. Стратегия прямо признает наличие ключевых вызовов, таких, как миграционный отток молодежи в возрасте 25-34 лет, высокий уровень молодежной безработицы и нехватка досуговых молодежных пространств. Таким образом, доминирование в тематическом моделировании LDA и BERTopic тем, связанных с «детьми» и «социальной поддержкой семей», при минимальном отражении проблем «молодежи», объективно отражает сложившийся дисбаланс, который на уровне стратегического планирования был признан критически важным.

Кроме того, выявленный узкий тематический спектр проектов, преимущественно сосредоточенных в социальной сфере, может быть объяснен не только приоритетами финансирования, но и структурными ограничениями некоммерческого сектора региона.

Проекта Стратегии социально-экономического развития Марий Эл до 2036 года указывает на слабость развития третьего сектора в республике Марий Эл, отмечая низкое место региона в общероссийских рейтингах по развитию НКО. НКО региона, согласно данным Стратегии, отмечают недостаток постоянного финансирования и поддержки со стороны региональных властей. Это позволяет предположить, что полученный нами «цифровой портрет» фиксирует узкую нишу, где некоммерческий сектор смог наработать экспертизу для привлечения финансирования (преимущественно через федеральные гранты), но не отражает всего спектра социальных и экономических задач региона.

Перспективным направлением методики является возможность для ее дальнейшего масштабирования за счет расширения корпуса данных, включив в него сведения о федеральных и региональных грантах, а также их сопоставление с ключевыми направлениями государственных программ и медиа-повестки для достижения максимальной полноты анализа.

Заключение. В представленной работе была успешно апробирована методика построения цифрового портрета социально-экономического развития региона на основе автоматизированного текстового анализа описаний грантовых проектов. В качестве ключевого инструментария использовалась комбинация двух методов тематического моделирования – классического LDA и современного нейросетевого BERTopic.

Было установлено, что тематическое пространство грантовой поддержки в регионе концентрируется вокруг нескольких доминирующих направлений: развитие и образование детей, социальная поддержка семей, инклюзия и адаптация людей с инвалидностью, а также организация культурно-досуговых и спортивных мероприятий. Смыслообразующим ядром всей грантовой повестки является целевая группа «дети», что указывает на стратегический приоритет инвестиций в человеческий капитал.

Сопоставление результатов тематического моделирования с проектом Стратегии социально-экономического развития Марий Эл до 2036 года демонстрирует четкую синергию: выявленные грантовые приоритеты (такие, как поддержка детей с ОВЗ, реабилитация лиц с нарушениями зрения, инклюзивное образование) полностью совпадают с действующими и планируемыми стратегическими задачами региональной политики. Это доказывает, что метод позволяет точно фиксировать направления, где гражданские инициативы и государственные приоритеты уже находятся в гармонии.

Однако наибольшая ценность метода заключается в его диагностическом потенциале для выявления стратегических пробелов и зон роста. «Цифровой портрет», который нами был зафиксирован, показывает подавляющее доминирование социальной сферы в грантовой активности. Между тем, проект Стратегии социально-экономического развития Марий Эл до 2036 года задает более широкий экономический и культурный вектор развития, фокусируясь на таких приоритетах, как «Регион передовых технологий» и «Регион культуры и гостеприимства». Стратегия указывает на необходимость развития креативных индустрий и туристической инфраструктуры. Наш анализ показывает, что некоммерческий сектор в настоящее время не вовлечен в реализацию этих новых экономических и культурных приоритетов. Таким образом, предложенный в статье инструмент может быть использован управленческими командами для директивной корректировки региональной грантовой политики, целенаправленно стимулируя появление и финансирование заявок НКО в стратегически важных, но пока слабо представленных в гражданских инициативах направлениях.

Практическая значимость работы заключается в демонстрации того, что методы машинного обучения, в частности, тематическое моделирование, превращают массив неструктурированных текстовых данных в мощный аналитический ресурс для органов

управления. Предложенный подход позволяет автоматизировать мониторинг и диагностику состояния социальной сферы, визуализировать латентные тематические паттерны для их наглядной интерпретации, обеспечивать обоснование управленческих решений на основе методологии Data-Driven.

Предложенный цифровой портрет может использоваться, как аналитический инструмент для мониторинга грантовых приоритетов региона, оценки сбалансированности поддержки различных социальных групп, а также выявления недофинансированных направлений. На основе полученных результатов могут формироваться рекомендации по корректировке грантовой политики и программ поддержки в рамках региональных конкурсов для социально ориентированных некоммерческих организаций. Органы власти могут целенаправленно стимулировать подачу заявок в недостаточно представленные сферы. Выявление нишевых инициатив – для управленца это не просто тема, а конкретная практика, которую можно проанализировать на предмет эффективности и тиражировать в других муниципалитетах. Разработанная модель обладает выраженным прикладным потенциалом для органов регионального управления и инфраструктуры поддержки НКО.

Таким образом, исследование подтверждает, что цифровой портрет региона, построенный на основе анализа грантовых проектов, является релевантным и эффективным инструментом для анализа приоритетов развития, обеспечивающим новые возможности для повышения эффективности регионального и муниципального управления. Применяемые авторами методы LDA и BERTopic дают возможность перейти от измерения «цифровой зрелости» через стандартизированные индикаторы к интерпретации латентных тематических профилей, формирующих цифровую повестку региона в его стратегических документах и проектных инициативах.

Список источников

1. Annals of computer science and information systems: proceedings of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS), Belgrade, Serbia, September 8-11, 2024, vol. 39. DOI:10.15439/978-83-969601-6-0.
2. Schiavon L. Addressing topic modelling via reduced latent space clustering. Statistical Methods & Applications, 2025, vol. 34, pp. 1–20
3. Ma L., Chen R., Ge W. et al. AI-powered topic modeling: comparing LDA and BERTopic in analyzing opioid-related cardiovascular risks in women. Experimental Biology and Medicine, 2025, vol. 250, p. 10389. Available at: https://pubmed.ncbi.nlm.nih.gov/40093658/ (accessed: 08/20/2025).
4. Ващенко В.А. Тематическое моделирование для коротких текстов: сравнительный анализ алгоритмов / В.А. Ващенко // Социология: 4М, 2023. – № 56. – С. 69-99.
5. Rejeb A., Rejeb K., Appolloni A. et al. Navigating the landscape of public–private partnership research: a novel review using latent Dirichlet allocation. International Journal of Public Sector Management, 2025, vol. 38, no. 2, pp. 213–237.
6. Wu X., Han Y., Yang F. et al. Analyzing CASIS policy data with AI: sentiment trends and topic modeling. Research Square [preprint], 2024, DOI:10.21203/rs.3.rs-5271894/v1.
7. Saheb T., Dehghani M. Artificial intelligence for sustainable energy: a contextual topic modeling and content analysis. Sustainable Computing: Informatics and Systems, 2022, vol. 35, article ID 100699, available at: https://www.sciencedirect.com/science/article/abs/pii/S2210537922000584 (accessed: 08/20/2025)
8. Мухачёва А.В. Цифровой профиль региона как фактор развития цифрового потенциала в социальной сфере / А.В. Мухачёва // Экономика, предпринимательство и право, 2025. – Т. 15, № 2. – С. 1219-1240.
9. Былинская, А.А. Цифровой портрет российских регионов: методологические основания исследования / А.А. Былинская // Социальные процессы современной России: материалы Международной научно-практической конференции (г. Нижний Новгород, 19–20 ноября 2020 г.). – Нижний Новгород : ООО «Научно-исследовательский социологический центр», 2020. – Т. 2. – С. 53-57.
10. Бушуева М.А. Метод обоснования факторного портрета регионов / М.А. Бушуева, Н.Н. Масюк, З.В. Брагина и др // АНИ: экономика и управление,2022. – № 4 (41). – С. 14-17.

11. Гусарова О.М. Цифровые модели социально-экономического развития региональных субъектов / О.М. Гусарова, В.Д. Кузьменкова, П.И. Комаров // Фундаментальные исследования, 2018. – № 8. – С. 42-47.
12. Романчуков С.В. Информационная система для анализа и моделирования социального и экономического развития региона / С.В. Романчуков, И.А. Лызин, О.В. Марухина // Информационные и математические технологии в науке и управлении, 2020. – № 3 (19). – С. 96-104.
13. Бождай А. С. Методика численной оценки уровня цифровой трансформации приоритетных направлений социально-экономических процессов регионов / А.С. Бождай, В.В. Свиридова // Модели, системы, сети в экономике, технике, природе и обществе, 2023. – № 2. – С. 172-184.
14. Строев В. В. Анализ цифровой зрелости регионов Российской Федерации / В.В. Строев, С.В. Сидоренко // Вестник университета, 2024. – № 5. – С. 5-14.
15. Об утверждении Стратегии социально-экономического развития Республики Марий Эл на период до 2030 года: утверждена постановлением Правительства Республики Марий Эл от 17 января 2018 г. № 12. – URL: https://docs.cntd.ru/document/446647066(дата обращения: 20.08.2025)
16. Проект Стратегии социально-экономического развития Республики Марий Эл на период до 2036 года. – URL: https://mari-el.gov.ru/upload/medialibrary/ac1/2j9wij7rqpujht6dgqymgnjqcr3z07bj.pdf(дата обращения: 20.08.2025)

Митрофанова Татьяна Валерьевна. *Кандидат физико-математических наук, доцент кафедры математического и аппаратного обеспечения информационных систем, Чувашский государственный университет им. И.Н. Ульянова. AuthorID: 630193, SPIN: 2567-1670, ORCID: 0000-0002-5750-7991, mitrofanova_tv@mail.ru. 428015, Чувашская Республика, г. Чебоксары, Московский пр-т, д. 15.*

Христофорова Анастасия Владимировна. *Кандидат физико-математических наук, доцент кафедры математического и аппаратного обеспечения информационных систем, Чувашский государственный университет им. И.Н. Ульянова. AuthorID: 591162, SPIN: 4797-3333, ORCID: 0000-0003-3534-8747, dlya.nastenki@mail.ru. 428015, Чувашская Республика, г. Чебоксары, Московский пр-т, д. 15.*

UDC 004.8:303.72
DOI:10.25729/ESI.2026.41.1.010

Digital profile of regional development: analysis of grant projects using LDA and BERTopic

Tatiana V. Mitrofanova, Anastasia V. Khristoforova

Chuvash State University named after I.N. Ulyanov,
Russia, Cheboksary, *mitrofanova_tv@mail.ru*

Abstract. This article proposes an approach for creating a digital profile of a region's socio-economic development through automated text data analysis. Grant project descriptions from the Republic of Mari El (2023–2025) were used as a key indicator. By combining classical LDA with the modern neural BERTopic model, latent thematic patterns shaping the region's current development agenda were identified and visualized. The resulting digital profile allows for an objective evaluation of civic initiatives, highlights dominant areas such as social support, human capital development, sports, and education, and assesses their alignment with regional strategic goals.

The study found that the target group "children" forms the core of the grant agenda, centering on clusters of educational, rehabilitation, and inclusive projects. LDA modeling identified five consistent thematic areas, including support for families with children and the adaptation of individuals with disabilities. BERTopic, in turn, enabled the detailed elaboration of narrow niche practices, such as inclusive sports, rehabilitation for people with visual impairments, and cultural and leisure programs for children with disabilities, confirming the high semantic sensitivity of the neural network approach. Comparison of the resulting thematic clusters with the draft Strategy for Socioeconomic Development of the Republic of Mari El revealed both areas of complete priority overlap (childhood support, inclusion) and strategic imbalances: the underrepresentation of projects in the creative industries, tourism, and youth work. The proposed methodology demonstrates that machine learning methods open up new opportunities for monitoring and diagnosing the state of regional and municipal systems, providing

management teams with a data-driven tool for decision-making and adjusting grant policy. Thus, thematic modeling enables the translation of unstructured text data into an objective analysis, transforming descriptions of civic initiatives into a reliable tool for identifying real social priorities and the basis for strategic planning.

Keywords: digital regional profile, topic modeling, LDA, BERTopic, data-driven management, grant analysis, regional development, machine learning, natural language processing (NLP), Republic of Mari El

References

1. Annals of computer science and information systems: proceedings of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS), Belgrade, Serbia, September 8-11, 2024, vol. 39. DOI:10.15439/978-83-969601-6-0.
2. Schiavon L. Addressing topic modelling via reduced latent space clustering. Statistical Methods & Applications, 2025, vol. 34, pp. 1–20
3. Ma L., Chen R., Ge W. et al. AI-powered topic modeling: comparing LDA and BERTopic in analyzing opioid-related cardiovascular risks in women. Experimental Biology and Medicine, 2025, vol. 250, p. 10389. Available at: https://pubmed.ncbi.nlm.nih.gov/40093658/ (accessed: 08/20/2025).
4. Vashchenko V.A. Tematicheskoye modelirovaniye dlya korotkikh tekstov: sravnitel'nyy analiz algoritmov [Topic modeling for short texts: a comparative analysis of algorithms]. Sotsiologiya: 4M [Sociology: Methodology, Methods, Mathematical Modeling], 2023, no. 56, pp. 69-99.
5. Rejeb A., Rejeb K., Appolloni A. et al. Navigating the landscape of public–private partnership research: a novel review using latent Dirichlet allocation. International Journal of Public Sector Management, 2025, vol. 38, no. 2, pp. 213–237.
6. Wu X., Han Y., Yang F. et al. Analyzing CASIS policy data with AI: sentiment trends and topic modeling. Research Square [preprint], 2024, DOI:10.21203/rs.3.rs-5271894/v1.
7. Saheb T., Dehghani M. Artificial intelligence for sustainable energy: a contextual topic modeling and content analysis. Sustainable Computing: Informatics and Systems, 2022, vol. 35, article ID 100699, available at: https://www.sciencedirect.com/science/article/abs/pii/S2210537922000584 (accessed: 08/20/2025)
8. Mukhacheva A.V. Tsifrovoy profil' regiona kak faktor razvitiya tsifrovogo potentsiala v sotsial'noy sfere [Digital profile of the region as a factor in the development of digital potential in the social sphere]. Ekonomika, predprinimatel'stvo i pravo [Journal of economics, entrepreneurship and law], 2025, vol. 15, no. 2, pp. 1219-1240.
9. Bylinskaya A.A. Tsifrovoy portret rossiyskikh regionov: metodologicheskiye osnovaniya issledovaniya [Digital portrait of Russian regions: methodological foundations of the study]. Sotsial'nyye protsessy sovremennoy Rossii: materialy Mezhdunarodnoy nauchno-prakticheskoy konferentsii [Social processes of modern Russia: proceedings of the International scientific and practical conference]. Nizhny Novgorod, NISC Publ., 2020, vol. 2, pp. 53-57.
10. Bushueva M.A., Masyuk N.N., Bragina Z.V. et al. Metod obosnovaniya faktornogo portreta regionov [Method of substantiating the factor portrait of regions]. ANI: ekonomika i upravlenie [ANI: economics and management], 2022, no. 4 (41), pp. 14-17.
11. Gusarova O.M., Kuzmenkova V.D., Komarov P.I. Tsifrovyye modeli sotsial'no-ekonomicheskogo razvitiya regional'nykh sub"yektov [Digital models of socio-economic development of regional subjects]. Fundamental'nyye issledovaniya [Fundamental research], 2018, no. 8, pp. 42-47.
12. Romanchukov S.V., Lyzin I.A., Marukhina O.V. Informatsionnaya sistema dlya analiza i modelirovaniya sotsial'nogo i ekonomicheskogo razvitiya regiona [Information system for analysis and modeling of social and economic development of the region]. Informatsionnyye i matematicheskiye tekhnologii v nauke i upravlenii [Information and mathematical technologies in science and management], 2020, no. 3 (19), pp. 96-104.
13. Bozhday A.S., Sviridova V.V. Metodika chislennoy otsenki urovnya tsifrovoy transformatsii prioritetnykh napravleniy sotsial'no-ekonomicheskikh protsessov regionov [Methodology for numerical assessment of the digital transformation level of priority areas of socio-economic processes in regions]. Modeli, sistemy, seti v ekonomike, tekhnike, prirode i obshchestve [Models, systems, networks in economics, technology, nature and society], 2023, no. 2, pp. 172-184.
14. Stroev V.V., Sidorenko S.V. Analiz tsifrovoy zrelosti regionov Rossiyskoy Federatsii [Analysis of digital maturity of the regions of the Russian Federation]. Vestnik universiteta [University bulletin], 2024, no. 5, pp. 5-14.
15. Ob utverzhdenii Strategii sotsial'no-ekonomicheskogo razvitiya Respubliki Mariy El na period do 2030 goda [On approval of the strategy for socio-economic development of the Republic of Mari El for the period up to 2030]. Approved by Decree of the Government of the Republic of Mari El No. 12 of January 17, 2018. Available at: https://docs.cntd.ru/document/446647066 (accessed: 08.20.2025).
16. Proyekt Strategii sotsial'no-ekonomicheskogo razvitiya Respubliki Mariy El na period do 2036 goda [Draft Strategy for Socio-Economic Development of the Republic of Mari El for the period up to 2036]. Available at: https://mari-el.gov.ru/upload/medialibrary/ac1/2j9wij7rqpujht6dgqymgnjqcr3z07bj.pdf (accessed: 08.20.2025).

Mitrofanova Tatiana Valeryevna. *Candidate of Physics and Mathematics Sciences, Associate Professor Department of Mathematical and Hardware Support of Information Systems, Chuvash State University named after I.N. Ulyanov. AuthorID: 630193, SPIN: 2567-1670, ORCID: 0000-0002-5750-7991, mitrofanova_tv@mail.ru. 428015, Russia, Chuvash Republic, Cheboksary, Moskovsky Prospect, 15.*

Khristoforova Anastasia Vladimirovna*. Candidate of Physics and Mathematics Sciences, Associate Professor Department of Mathematical and Hardware Support of Information Systems, Chuvash State University named after I.N. Ulyanov. AuthorID: 591162, SPIN: 4797-3333, ORCID: 0000-0003-3534-8747, dlya.nastenki@mail.ru. 428015, Russia, Chuvash Republic, Cheboksary, Moskovsky Prospect, 15.*

Статья поступила в редакцию 08.10.2025; одобрена после рецензирования 05.11.2025; принята к публикации 06.02.2026.

The article was submitted 10/08/2025; approved after reviewing 11/05/2025; accepted for publication 02/06/2026.