УДК 004.912
DOI 10.17513/snt.40617

РОБАСТНОЕ ТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ В ТЕКСТОВЫХ ПОТОКАХ НА ОСНОВЕ ГИБРИДНОГО ПОДХОДА

Родионов Д.Г. ORCID ID 0000-0002-1254-0464,
Поляков П.А. ORCID ID 0000-0003-1362-6283,
Конников Е.А. ORCID ID 0000-0002-4685-8569

Федеральное государственное автономное образовательное учреждение высшего образования «Санкт-Петербургский политехнический университет Петра Великого», Санкт-Петербург, Российская Федерация, e-mail: prohor@polyakov-box.ru

В данной работе рассматривается задача автоматического выявления и прогнозирования тематических трендов в потоке новостных текстов на примере ядерной энергетики. Цель исследования состоит в разработке и экспериментальной верификации гибридной методологии робастного тематического моделирования потоков новостных текстов по проблематике ядерной энергетики, объединяющей классическую модель Latent Dirichlet Allocation с семантически обогащённой фильтрацией корпуса по ключевым доменным терминам и ориентированной на повышение интерпретируемости и устойчивости выделяемых тематических структур для последующего прогнозирования тематических трендов. Предлагается гибридный метод, сочетающий классическое тематическое моделирование Latent Dirichlet Allocation и семантически обогащенную фильтрацию документов по ключевым доменным терминам, направленный на повышение устойчивости (робастности) и интерпретируемости результатов. Проведен эксперимент на корпусе из 1000 новостных документов об атомной энергетике. Baseline-модель (стандартная Latent Dirichlet Allocation) сравнивается с robust-моделью (с применением предлагаемого подхода). Для оценки результатов использовалась симуляция Монте-Карло – 25 повторных прогонов тематического моделирования с разными инициализациями. Ключевые метрики включают интерпретируемость тем (доля ядерных терминов среди топ-N слов темы) и стабильность распространённости тем (стандартное отклонение доли темы в корпусе между прогонами). Полученные результаты демонстрируют, что гибридный подход значительно повышает интерпретируемость с 0.05 до 0.27. Таким образом, предложенный метод позволяет более надёжно и осмысленно выявлять тематические структуры в новостных текстовых потоках, что важно для последующего прогнозирования тематических трендов. Это создает основу для построения устойчивых систем мониторинга и аналитики отраслевых информационных потоков в реальном времени.

Ключевые слова: тематическое моделирование, Latent Dirichlet Allocation, интерпретируемость тем, стабильность тем, симуляция Монте-Карло, семантическое обогащение, гибридный метод, ядерная энергетика

ROBUST THEMATIC MODELING IN TEXT FLOWS BASED ON A HYBRID APPROACH

Rodionov D.G. ORCID ID 0000-0002-1254-0464,
Polyakov P.A. ORCID ID 0000-0003-1362-6283,
Konnikov E.A. ORCID ID 0000-0002-4685-8569

Federal State Autonomous Educational Institution of Higher Education "Peter the Great St. Petersburg Polytechnic University", St. Petersburg, Russian Federation, e-mail: prohor@polyakov-box.ru

This paper considers the problem of automatic detection and prediction of thematic trends in a stream of news texts using the example of nuclear energy. A hybrid method is proposed that combines classical Latent Dirichlet Allocation thematic modeling and semantically enriched filtering of documents by key domain terms, aimed at improving the robustness and interpretability of the results. An experiment was conducted on a corpus of 1,000 news documents about nuclear energy. The baseline model (standard Latent Dirichlet Allocation) is compared with the robust model (using the proposed approach). Monte Carlo simulation was used to evaluate the results-25 repeated runs of thematic modeling with different initializations. Key metrics include topic interpretability (the proportion of nuclear terms among the top N words of a topic) and topic prevalence stability (the standard deviation of the proportion of a topic in the corpus between runs). The results demonstrate that the hybrid approach significantly increases interpretability from 0.05 to 0.27. Thus, the proposed method allows for more reliable and meaningful identification of thematic structures in news text streams, which is important for subsequent forecasting of thematic trends. This creates a basis for building robust systems for real-time monitoring and analysis of industry information flows.

Keywords: thematic modeling, Latent Dirichlet Allocation, interpretability of topics, stability of topics, Monte Carlo simulation, semantic enrichment, hybrid method, nuclear energy

Введение

В эпоху информационной перегрузки возрастает необходимость в автоматизированных методах анализа больших потоков текстовых данных, таких как новостные ленты. В частности, актуальна задача выявления и прогнозирования тематических трендов – определение скрытых тем в тексте и отслеживание их динамики во времени для предсказания тенденций. Тради-

ционно для обнаружения тем в большом корпусе документов применяются методы тематического моделирования. Одним из самых популярных подходов является Latent Dirichlet Allocation или LDA – байесовская вероятностная модель, представляющая документы как смесь латентных тем, а темы – как распределения слов [1]. Метод LDA и его варианты успешно применялись для анализа коллекций документов в различных предметных областях, включая новостные статьи, научные публикации и социальные медиа. В частности, тематическое моделирование применяется для анализа онлайн-сообществ [2], изучения образовательной повестки в соцсетях [3], автоматического анализа обращений в службах поддержки [4], исследования влияния новостных сюжетов на финансовые рынки, а также в прикладных задачах банковского сектора.

Однако прямое применение LDA к реальным потокам текстов сталкивается с рядом проблем [5; 6]. Во-первых, получаемые автоматически темы не всегда легко интерпретируемы человеком. Топ-слова темы могут быть слишком общими или несвязанными, особенно в предметно-специализированных областях. Известно, что стандартный LDA может выделять темы, имеющие смешанный или шумовой характер, что затрудняет их семантическое толкование. Во-вторых, устойчивость (стабильность) тематического моделирования вызывает обеспокоенность. Поскольку обучение LDA включает случайные инициализации и стохастические процедуры, результаты разных прогонов на одном корпусе могут различаться. Это означает, что обнаруженные темы и их важность могут меняться от запуска к запуску, что затрудняет надёжное отслеживание тематических трендов: исследователь может получить иную тематическую структуру при повторном анализе того же самого потока документов. В литературе отмечается, что нестабильность LDA способна приводить к противоречивым выводам и снижает доверие к модели.

Для решения указанных проблем в последнее десятилетие предлагается ряд усовершенствований тематических моделей. Повышению интерпретируемости тем посвящены работы, вводящие меры когерентности (семантической согласованности) тем и методы, позволяющие встраивать априорные знания о предметной области в модель [7]. Например, существуют направляемые тематические модели, где пользователю предоставляется возможность задавать списки ключевых слов для тем или связи между словами. В рамках LDA такие подходы включают добавление априорных ограничений, таких как must-link и cannot-link для пар слов (например, через Dirichlet Forest-приоры) или начальное «посевное» задание тематических слов (seed words). В русскоязычной литературе одним из заметных направлений развития LDA является метод аддитивной регуляризации тематических моделей (ARTM) [8; 9], предлагающий набор регуляризаторов для повышения разреженности, разнообразия и привязки к внешним знаниям [10]. Подобные методы демонстрируют, что использование онтологий, тезаурусов или экспертных словарей способно сделать темы более осмысленными для человека, выделяя действительно терминологически насыщенные темы, соответствующие понятиям предметной области [11].

Проблема стабильности тематических моделей также привлекла внимание исследователей. Одно из направлений – оптимизация параметров LDA (число тем, гиперпараметры инициализации) с целью максимизации согласованности результатов между запусками. Другой подход – выполнение нескольких запусков модели и последующий анализ кластеров, полученных тем для выявления консенсусных (стабильных) тем. Стабильность тем можно количественно оценивать через метрику сходства тем между разными прогонами или вариабельность распределения весов тем. Например, Greene et al. [12] предлагают анализ стабильности как критерий для выбора оптимального числа тем. Более устойчивой считается модель, в которой тем меньше, но они воспроизводимы при случайных возмущениях данных. Agrawal и соавт. [13] прямо указывают на нестабильность LDA как на серьёзный недостаток и показывают, что перебор параметров с эволюционным алгоритмом способен частично повысить стабильность, однако вопрос остаётся открытым.

Таким образом, существует необходимость в методах, которые одновременно повышают интерпретируемость и стабильность тематического моделирования на специальных доменах. В идеале, инкорпорируя знание о предметной области (что улучшает содержательность тем), можно также отфильтровать часть «шума» – документов или слов, не относящихся к интересующей тематике, тем самым повысив устойчивость модели. В работе реализован именно такой подход. Перед обучением LDA производится фильтрация корпуса на основе списка ключевых ядерных терминов (семантическое обогащение корпуса), после чего модель обучается на очищенном подкорпусе. Авторы ожидают, что полученные темы будут более сфокусированы на ядерной энергетике и менее подвержены влиянию нере-

левантных данных, что повысит повторяемость результатов при повторных запусках.

Цель исследования состоит в разработке и экспериментальной верификации гибридной методологии робастного тематического моделирования потоков новостных текстов по проблематике ядерной энергетики, объединяющей классическую модель LDA с семантически обогащённой фильтрацией корпуса по ключевым доменным терминам и ориентированной на повышение интерпретируемости и устойчивости выделяемых тематических структур для последующего прогнозирования тематических трендов.

Материалы и методы исследования

Исследование выполнено на базе Санкт-Петербургского политехнического университета Петра Великого в рамках проекта «Разработка методологии формирования инструментальной базы анализа и моделирования пространственного социально-экономического развития систем в условиях цифровизации с опорой на внутренние резервы» (FSEG-2023-0008). Эмпирической базой работы служит корпус из 1000 русскоязычных новостных материалов по атомной энергетике, опубликованных в открытых интернет-источниках в период с 2010 по 2025 г. Исследование носит вычислительный характер и опирается на методы компьютерной лингвистики и статистического анализа текстов: предобработку (удаление стоп-слов, лемматизацию, TF-IDF-взвешивание), тематическое моделирование методом Latent Dirichlet Allocation (LDA) и многократные прогоны модели (симуляция Монте-Карло) для оценки устойчивости результатов.

Источниками документов послужили открытые интернет-ресурсы новостей энергетики. Временной охват – приблизительно 2010–2025 гг., что позволяет охватить длительный период развития тем. Для каждого документа доступны заголовок и основной текст. В рамках апробации они были объединены для совместной обработки, поскольку заголовок часто содержит важные ключевые слова темы. Предварительная обработка текста включала удаление стоп-слов, приведение слов к нижнему регистру и лемматизацию для унификации терминологии. Для представления документов использована схема взвешивания TF-IDF, которая уменьшает вклад частых слов и усиливает значимость редких терминов, что полезно при тематическом моделировании специализированных текстов.

Ключевое отличие предлагаемой robust-модели от стандартной заключается в этапе семантической фильтрации корпуса. Сформирован список ядерных терминов – слов и устойчивых выражений, однозначно относящихся к тематике атомной энергетики. В него вошли, в частности, названия и типы ядерных реакторов, материалы и топливо, названия организаций и проектов, а также общетехнические термины, характерные для данной отрасли. Данный словарь был составлен экспертным путем на основе анализа предметной области и частотности встречаемости слов в корпусе. Из исходных 1000 текстов исключались те, в которых не найдено ни одного слова из списка ядерных терминов. Предполагалось, что такие документы либо нерелевантны тематике, либо содержат лишь косвенные упоминания без детального тематического содержания. После фильтрации остался 921 документ (около 8% корпуса были отброшены как шумовые или нерелевантные). На этом отфильтрованном подкорпусе далее выполнялось тематическое моделирование аналогично baseline-подходу.

Кроме фильтрации, был реализован шаг семантического обогащения текста. Для некоторых терминов из словаря были добавлены их синонимы и связанные понятия, чтобы учесть возможное разнообразие изложения. Например, помимо слова «реактор» учитывались словосочетания «атомный реактор», «ядерный реактор», для «радиации» – близкие понятия «радиоактивность», «излучение». Это позволило не упустить документы, где использована альтернативная лексика описания ядерных тем. Обогащение было реализовано на этапе предварительной обработки текста. После лемматизации каждое слово проверялось на принадлежность к расширенному словарю. Если слово было синонимично ключевому термину, ему присваивался тот же признак. Таким образом, семантически эквивалентные термины были нормализованы к одному виду с точки зрения модели.

Для обеих стратегий – baseline и robust – использовалась одинаковая настройка LDA, что позволяет корректно сравнивать результаты. Модель LDA реализована с помощью библиотеки scikit-learn. Число тем фиксировано равным 8. Выбор именно 8 тем обоснован предварительным анализом. При меньшем числе часть разнородных аспектов ядерной тематики сливалась в одни темы, теряя детализацию, а при большем – некоторые темы становились избыточно дробными и менее устойчивыми. Таким образом, 8 – компромиссное число, подтвержденное также рекомендациями по тематическому моделированию. В параметрах LDA был использован вариационный байесовский вывод (learning_method='batch'), концентрация Дирихле по документ-темам α и по словам-темам β взяты по умолчанию

число итераций обучения – до сходимости. Для каждого документа модель порождает распределение по 8 темам, а для каждой темы – распределение вероятностей по словам из словаря, взвешенного TF-IDF.

Чтобы оценить стабильность тематического моделирования, проведен многократный повтор эксперимента. Для каждой из двух моделей было выполнено 25 итераций с различными случайными инициализациями. В scikit-learn LDA используется случайность при инициализации тем, поэтому разные запуски на одном корпусе приводят к несколько различающимся решениям локального максимума правдоподобия. Каждый прогон состоял из обучения LDA на соответствующем корпусе и сохранения полученных результатов: топ-слов для каждой из 8 тем, распределения долей тем по документам. Впоследствии эти результаты объединялись для вычисления метрик. Следует отметить, что 25 повторов – это компромисс между статистической достоверностью оценок и вычислительной затратностью. Согласно литературе, даже 10–20 повторных запусков могут достаточно надежно выявить степень нестабильности модели.

Основными целевыми показателями качества темы были выбраны:

• интерпретируемость темы – доля доменных (ядерных) терминов среди топ-N слов темы. В расчетах N=15 – топ-15 наиболее вероятных слов каждой темы рассматривались как характерные слова темы;

• стабильность распространённости темы – стандартное отклонение доли данной темы в корпусе между разными запусками. Под долей темы в корпусе понимается суммарная вероятность этой темы по всем документам;

• сводный индекс стабильности модели представляет собой агрегированный показатель, интегрирующий стабильность всех тем. В качестве такого индекса использовалась величина $S = 1 / (1 + \text{mean\_std})$.

Помимо указанных метрик, анализировались также качественные характеристики полученных тем [14]. Для интерпретации приводились топ-слова тем, им давались условные названия, сравнивались похожие темы между моделями baseline и robust. Однако основной упор сделан на количественных метриках, перечисленных выше, так как именно они позволяют строго сравнить два подхода. Полученные тематические распределения могут использоваться для последующего анализа временной динамики тематик в энергетической повестке и прогнозирования тематических трендов [15].

Результаты исследования и их обсуждение

После проведения 25 прогонов тематического моделирования для каждой из моделей были получены усреднённые показатели интерпретируемости и стабильности, а также распределения этих метрик. Сравнение результатов baseline- и robust-подходов представлено в таблице.

Из таблицы видно, что robust-модель существенно превосходит baseline по интерпретируемости тем. Доля профильных (ядерных) терминов среди ключевых слов тем увеличилась с 0.048 до 0.270. Иными словами, в стандартной LDA лишь около 5% слов в списках топ-15 относятся непосредственно к ядерной энергетике, тогда как при применении фильтрации этот показатель вырос до 27%. Разница более чем в 5 раз указывает на значительно более высокий уровень тематической «чистоты» в robust-модели. Для лучшего понимания распределения этого показателя на рисунке 1 ниже приведен boxplot интерпретируемости по темам.

Для каждой из 8 тем отображается медиана и размах доли ядерных терминов в топ-словах по 25 запускам. Можно заметить, что у baseline-модели большинство наблюдений сосредоточены около нуля, что говорит о слабой выраженности профильных терминов. Напротив, в robust-модели наблюдения смещены вверх – медианные значения ~0.25, а по некоторым темам достигают 0.4–0.5, причём разброс внутри одной темы относительно невелик. Это подтверждает, что семантически обогащенная модель стабильно генерирует темы с высоким содержанием доменной лексики независимо от случайной инициализации.

Сравнение моделей baseline и robust по основным метрикам

Метрика	Baseline- модель	Robust- модель
Количество документов в корпусе	1000	921
Доля доменных слов в топ-15 слов темы (интерпретируемость)	0.048	0.270
Сводный индекс стабильности распространённости тем $S = 1 / (1 + \text{avg STD})$	0.815	0.806

Примечание: составлено авторами на основе полученных данных в ходе исследования.

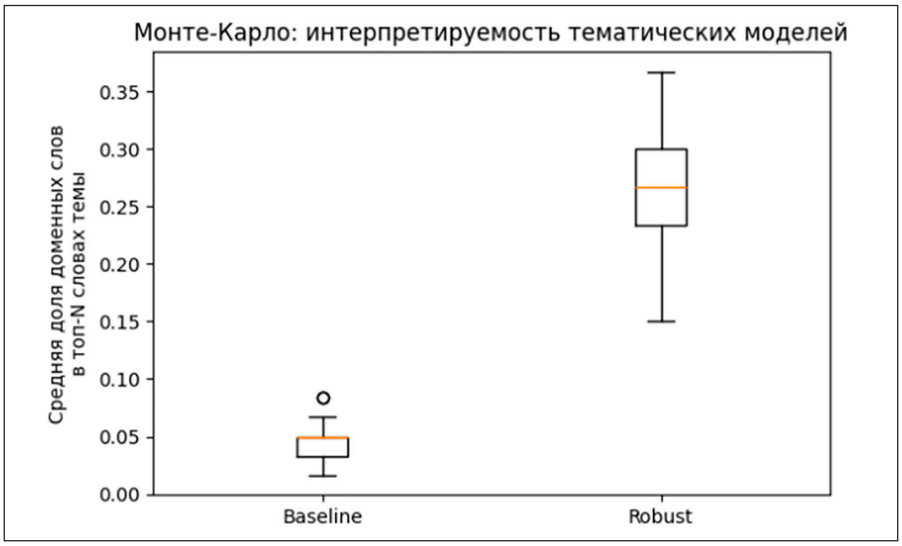

Рис. 1. Интерпретируемость тематических моделей
Источник: составлено авторами по результатам данного исследования

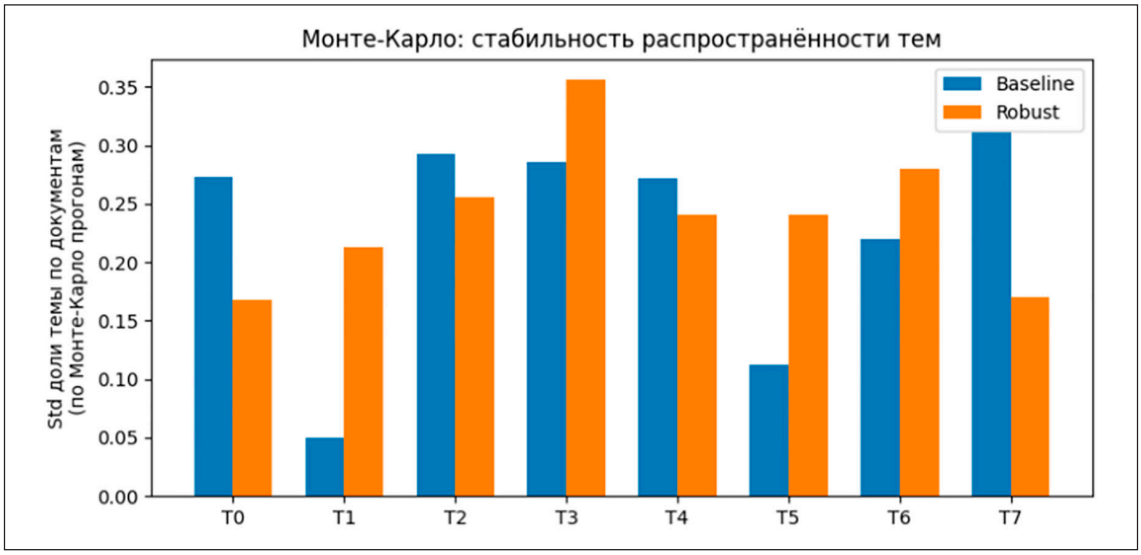

Рис. 2. Стабильность распространенности тем
Источник: составлено авторами по результатам данного исследования

Что касается стабильности тем, то усреднённый сводный индекс S равен 0.815 для baseline-модели и 0.806 для robust-варианта. При выбранном определении индекса (чем больше S, тем меньше средний разброс долей тем) baseline оказывается немного более стабильной в среднем.

На рисунке 2 представлены стандартные отклонения доли каждой темы для обеих моделей. Значения STD лежат в диапазоне примерно 0.05–0.35. Для тем T0, T2, T4 и T7 robust-модель демонстрирует меньший разброс, чем baseline, то есть для этих сюжетов доля темы по документам меняется между запусками менее существенно. Для тем T1, T3, T5 и T6 картина обратная: оранжевые столбцы выше синих, что говорит о более высокой чувствительности этих тем к случайной инициализации. В среднем суммарный разброс по всем восьми темам у baseline немного ниже, что отражается в интегральных индексах стабильности. S = 0.815 для baseline против S = 0.806 для robust. Таким образом, гибридный подход не делает модель «абсолютно более стабильной», а скорее перераспределяет вариативность: часть тем становится устойчивее, часть – несколько более чувствительной, при этом общее улучшение приходится прежде всего на интерпретируемость.

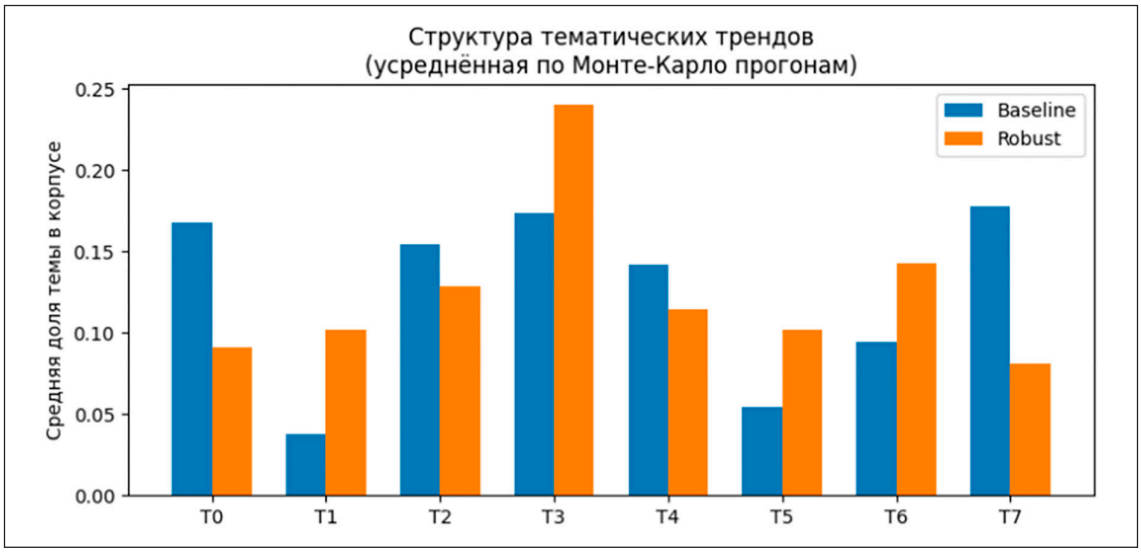

Рис. 3. Структура тематических трендов
Источник: составлено авторами по результатам данного исследования

Кроме того, был проанализирован средний тематический состав корпусов для обеих моделей. Рисунок 3 иллюстрирует усреднённую по запускам структуру тем – долю каждого из восьми топиков в корпусе для baseline- и robust-решений.

В baseline-модели наблюдаются две относительно малочисленные темы (T1 и T5, около 4–5% корпуса каждая), тогда как остальные темы распределены более-менее равномерно в диапазоне 9–18%. В robust-модели после фильтрации корпуса исчезают совсем «маленькие» темы: минимальная доля темы возрастает до ~8–9%. Однако одновременно формируется выраженный «лидер» T3 с долей около 24%, которая превосходит долю любой темы в baseline-варианте.

Резюмируя, сообщим, что гибридный подход продемонстрировал значительное улучшение качества тематического моделирования по ключевому содержательному параметру. Интерпретируемость тем возросла примерно в 5–6 раз, что подтверждает эффективность семантического обогащения – модель концентрируется на терминологии ядерной отрасли. При этом влияние на стабильность неоднозначно. По интегральному индексу baseline остаётся немного более стабильной по долям тем, тогда как robust-подход перераспределяет вариативность между темами. Тем не менее robust-модель убирает наиболее маргинальные, слабо интерпретируемые темы и формирует набор более предметно насыщенных топиков, что делает её более удобной для практического анализа тематических трендов.

Полученные результаты демонстрируют, что объединение статистического подхода LDA с простыми эвристиками на основе экспертных знаний о домене может привести к существенному выигрышу. Рост доли тематических терминов в топ-словах тем с ~0.05 до ~0.27 – весьма внушительный показатель. Это означает, что robust-модель фактически «узнала» терминологию ядерной энергетики значительно лучше, чем стандартная LDA, которая, видимо, «отвлекалась» на общие слова.

Наконец, важно подчеркнуть, что гибридный метод относительно прост и интерпретируем со стороны процесса. В отличие от «чёрного ящика» сложных нейронных моделей тематического анализа, описываемый в данной работе подход явно использует понятные шаги фильтрации. Это облегчает его применение в прикладных сценариях – от аналитики новостей до мониторинга социальных медиа – где требуется объяснимость.

Заключение

Влияние гибридного подхода на стабильность тематической структуры оказалось более сложным. Интегральный индекс стабильности долей тем между повторами обучения для baseline-варианта несколько выше, чем для robust-модели, однако детальный анализ показывает перераспределение вариативности между отдельными темами. Часть сюжетов становится устойчивее, часть - более чувствительной к случайной инициализации. Дополнительно гибридный подход удаляет около 8% нерелевантных документов, за счёт чего исчезают откровенно «мусорные» темы, перегруженные общими словами, а оставшийся набор топиков лучше соответствует содержатель-

ным аспектам корпуса. Тем самым модель формирует более осмысленный тематический срез новостного потока, несмотря на то что формальные показатели стабильности улучшаются не по всем темам.

Практическая значимость полученных результатов состоит в том, что устойчивые и предметно интерпретируемые темы могут использоваться для мониторинга информационных трендов, например для отслеживания интереса к новым реакторным технологиям, к вопросам безопасности и обращения с отходами или к международным проектам в области атомной энергетики. Более робастное тематическое разбиение снижает зависимость выводов аналитика от случайных артефактов моделирования и делает прогнозы трендов ближе к реальной динамике содержания текстов.

Перспективным направлением развития предлагаемой гибридной методологии является автоматизация формирования доменного словаря. В частности, словарь может пополняться автоматически на основе процедур извлечения терминов и устойчивых словосочетаний, статистик значимости и семантической близости, с последующей верификацией человеком. Это позволит снижать трудоёмкость адаптации метода к новым поддоменам и источникам, а также быстрее учитывать появление новых сущностей в отраслевой повестке. Дополнительно метод может быть расширен за счёт интеграции с динамическими тематическими моделями (DTM), которые позволяют оценивать эволюцию тем во времени и выявлять устойчивые и зарождающиеся тренды на временных срезах. В такой связке семантически отфильтрованный корпус может служить более «чистым» входом для DTM, а также использоваться для задания априорных ограничений, что потенциально улучшит интерпретируемость временных траекторий тем и устойчивость их отслеживания.

Список литературы

1. Blei D.M., Ng A.Y., Jordan M.I. Latent Dirichlet Allocation // Journal of Machine Learning Research. 2003. Vol. 3. P. 993–1022. URL: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf (дата обращения: 25.11.2025). DOI: 10.1162/jmlr.2003.3.4-5.993.

2. Kim S.-H., Cho H.-G. User–Topic Modeling for Online Community Analysis // Applied Sciences. 2020. Vol. 10. № 10. Art. 3388. DOI: 10.3390/app10103388.

3. Waheeb S.A., Khan N.A., Shang X. Topic Modeling and Sentiment Analysis of Online Education in the COVID-19 Era Using Social Networks Based Datasets // Electronics. 2022. Vol. 11. № 5. Art. 715. DOI: 10.3390/electronics11050715.

4. Papadia G., Pacella M., Giliberti V. Topic Modeling for Automatic Analysis of Natural Language: A Case Study in an Italian Customer Support Center // Algorithms. 2022. Vol. 15. № 6. Art. 204. DOI: 10.3390/a15060204.

5. Chen W., Rabhi F., Liao W., Al-Qudah I. Leveraging State-of-the-Art Topic Modeling for News Impact Analysis on Financial Markets: A Comparative Study // Electronics. 2023. Vol. 12. № 12. Art. 2605. DOI: 10.3390/electronics12122605.

6. Ogunleye B., Maswera T., Hirsch L., Gaudoin J., Brunsdon T., Boateng K.A. Comparison of Topic Modelling Approaches in the Banking Context // Applied Sciences. 2023. Vol. 13. № 2. Art. 797. DOI: 10.3390/app13020797.

7. Röder M., Both A., Hinneburg A. Exploring the Space of Topic Coherence Measures: A Unified Framework and Evaluation // Proceedings of the 8th ACM International Conference on Web Search and Data Mining (WSDM 2015). 2015. P. 399–408. DOI: 10.1145/2684822.2685324.

8. Воронцов К.В., Потапенко А.А. Аддитивная регуляризация тематических моделей коллекций текстовых документов // Доклады Академии наук. 2014. Т. 456. № 3. С. 268–271. DOI: 10.1134/S1064562414020185.

9. Vorontsov K.V., Potapenko A.A. Additive Regularization of Topic Models // Machine Learning. 2015. Vol. 101. № 1–3. P. 303–323. URL: https://doi.org/10.1007/s10994-014-5476-6 (дата обращения: 25.11.2025). DOI: 10.1007/s10994-014-5476-6.

10. Veselova E., Vorontsov K. Topic Balancing with Additive Regularization of Topic Models // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. 2020. P. 59–65. URL: https://aclanthology.org/2020.acl-srw.9 (дата обращения: 25.11.2025). DOI: 10.18653/v1/2020.acl-srw.9.

11. Bulatov V., Alekseev V., Vorontsov K., Polyudova D., Veselova E., Goncharov A., Egorov E. TopicNet: Making Additive Regularisation for Topic Modelling Accessible // Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC 2020). European Language Resources Association. 2020. P. 6745–6752. URL: https://aclanthology.org/2020.lrec-1.833 (дата обращения: 25.11.2025).

12. Greene D., O'Callaghan D., Cunningham P. How Many Topics? Stability Analysis for Topic Models // Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2014. Lecture Notes in Computer Science. Vol. 8724. 2014. P. 498–513. DOI: 10.1007/978-3-662-44848-9_32.

13. Agrawal A., Fu W., Menzies T. What Is Wrong with Topic Modeling? And How to Fix It Using Search-Based Software Engineering // Information and Software Technology. 2018. Vol. 98. P. 74–88. DOI: 10.1016/j.infsof.2018.02.005.

14. Williams L., Anthi E., Arman L., Burnap P. Topic Modelling: Going Beyond Token Outputs // Big Data and Cognitive Computing. 2024. Vol. 8. № 5. Art. 44. DOI: 10.3390/bdcc8050044.

15. Wang Z., Zhou R., Wang Y. DTM-Based Analysis of Hot Topics and Evolution of China's Energy Policy // Sustainability. 2024. Vol. 16. № 19. Art. 8293. DOI: 10.3390/su16198293.

Конфликт интересов: Авторы заявляют об отсутствии конфликта интересов.

Conflict of interest: The authors declare that there is no conflict of interest.

Финансирование: Работы выполнены в рамках реализации проекта «Разработка методологии формирования инструментальной базы анализа и моделирования пространственного социально-экономического развития систем в условиях цифровизации с опорой на внутренние резервы» (FSEG-2023-0008).

Financing: The work was carried out as part of the project "Development of a methodology for forming an instrumental base for analyzing and modeling the spatial socio-economic development of systems in the context of digitalization based on internal reserves" (FSEG-2023-0008).