Программные системы и вычислительные методы
Правильная ссылка на статью:
Конников Е.А., Крыжко Д.А. Двухступенчатая семантическая кластеризация эмбеддингов как альтернатива LDA для инфометрического анализа отраслевых новостей // Программные системы и вычислительные методы. 2025. № 3. DOI: 10.7256/2454-0714.2025.3.75348 EDN: HSKPLS URL: https://nbpublish.com/library_read_article.php?id=75348

Двухступенчатая семантическая кластеризация эмбеддингов как альтернатива LDA для инфометрического анализа отраслевых новостей

Конников Евгений Александрович

кандидат экономических наук

доцент; институт промышленного менеджмента, экономики и торговли; Санкт-Петербургский политехнический университет Петра Великого

194021, Россия, г. Санкт-Петербург, Выборгский р-н, ул. Новороссийская, д. 50

konnikov.evgeniy@gmail.com

Крыжко Дарья Александровна

кандидат экономических наук

доцент; институт промышленного менеджмента, экономики и торговли; Санкт-Петербургский политехнический университет Петра Великого

194021, Россия, г. Санкт-Петербург, Выборгский р-н, ул. Новороссийская, д. 50

darya.kryz@yandex.ru

Статья из рубрики "Системный анализ, поиск, анализ и фильтрация информации"

DOI:

10.7256/2454-0714.2025.3.75348

EDN:

HSKPLS

Дата направления статьи в редакцию:

27-07-2025

Дата публикации:

03-08-2025

Аннотация: Предметом исследования является разработка и валидация

альтернативного подхода к тематическому моделированию текстов, направленного на преодоление ограничений классического латентного размещения Дирихле (LDA). Объектом исследования выступают короткие русскоязычные новостные тексты об атомной энергетике, представленные в виде корпуса «AtomicNews». Авторы подробно рассматривают такие аспекты темы, как влияние разреженности на качество тематического моделирования, проблемы интерпретируемости тем и ограничения априорной фиксации числа тем. Особое внимание уделяется геометрической интерпретации семантики текстов, в частности, преобразованию лексических единиц в пространство предобученных эмбеддингов и последующей кластеризации с целью формирования документных тематических профилей. Исследование фокусируется на сравнительном анализе нового метода и LDA по метрикам когерентности, перплексии и тематического разнообразия. Предлагаемый подход направлен на создание интерпретируемой, вычислительно лёгкой и устойчивой к шуму модели, пригодной для онлайнового мониторинга новостных потоков. Методология исследования основана на двухступенчатом семантическом сглаживании – эмбеддинг-репрезентации лемм с помощью Sentence-BERT и агломеративной косинусной кластеризации с последующим применением K-means к тематическим профилям документов. Научная новизна исследования заключается в разработке и эмпирическом обосновании схемы тематического моделирования, заменяющей вероятностную генерацию слов на геометрическое сглаживание эмбеддингов. Предложенный подход отказывается от предпосылок «мешка слов» и фиксированного числа тем, формируя тематические координаты документов через плотностные кластеры в семантическом пространстве. Это позволяет повысить интерпретируемость тем, снизить чувствительность к разреженности текстов и избежать коллапса распределения тем в коротких сообщениях. Эксперименты на корпусе «AtomicNews» показали статистически значимое улучшение по сравнению с классической LDA: снижение перплексии на 5 %, рост когерентности тем на 0.15 пункта и увеличение тематического разнообразия. Метод также продемонстрировал вычислительную эффективность – вся процедура занимает секунды на CPU, что делает его пригодным для применения в условиях ограниченных ресурсов. Таким образом, переход от вероятностной декомпозиции к геометрическому анализу эмбеддингов представляет собой перспективное направление в тематическом моделировании отраслевых текстов.

Ключевые слова:

тематическое моделирование, эмбеддинги слов, косинусная кластеризация, когерентность тем, латентное размещение Дирихле, Sentence-BERT, текстовая информация, лемматизация, тематический профиль, медиамониторинг

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)

Введение

Латентное размещение Дирихле (LDA) на протяжении последних двух десятилетий стало одним из наиболее распространённых подходов для тематического моделирования текстов [1]. Данная вероятностная модель генерирует фиксированный набор тем и представляет каждый документ в виде распределения этих тем, что доказало свою

эффективность при кластеризации текстовых коллекций и обеспечении достаточно связных наборов ключевых слов для многих приложений [2]. Тем не менее, классическая LDA накладывает ряд упрощающих априорных предположений, которые не всегда соответствуют реальному характеру данных [3]. Во-первых, модель предполагает условную независимость слов в документе и не учитывает порядок слов, из-за чего теряется лингвистический контекст. Это приводит к тому, что LDA не способна различать значения омонимичных слов и учитывать устойчивые словосочетания [4,5]. Во-вторых, требование зафиксировать число тем K заранее является серьезным ограничением: на практике оптимальное значение K неизвестно и может варьироваться для разных корпусов документов [6]. Непараметрические расширения позволяют делать K переменным случайным параметром, но они усложняют построение модели и повышение интерпретируемости результатов [6,7]. В-третьих, LDA испытывает трудности при моделировании коротких текстов, где наблюдается острая разреженность данных и низкая совстречаемость слов [8]. Как показано в исследованиях, традиционные методы вроде LDA заметно деградируют на коротких сообщениях (например, в социальных сетях) именно из-за нехватки статистики совместной встречаемости терминов [9]. Даже при успешном выделении тем LDA распределяет все темы «на равных» и не предоставляет средств автоматически оценить, какие из них наиболее осмысленны или полезны для аналитика [6]. Наконец, качество тематической модели существенно зависит от гиперпараметров (таких как параметры дирихлеевских распределений α и β). Их оптимизация затруднена, поскольку изменение этих параметров или дополнительных регуляризаторов может непредсказуемо влиять на получаемые темы [10]. Например, даже широко используемая мера перплексии не всегда адекватно отражает качество тем, перплексия монотонно уменьшается с увеличением сложности модели и не имеет явного минимума, из-за чего ее снижение не гарантирует улучшения интерпретируемости. Более того, перплексия не чувствительна к устранению стоп-слов и не позволяет напрямую сравнивать модели с разными словарями или учитывать биграммы. В то же время, такая прикладная мера как когерентность тем, напротив, показала высокую корреляцию с качеством тематической интерпретации по мнению человека-эксперта [6].

Ограничения базовой модели LDA стимулировали появление множества усовершенствованных методик тематического моделирования. Во-первых, были предложены модели, учитывающие динамику тем во времени и позволяющие отличать устаревшие темы от новых [11]. Например, динамическая тематическая модель (DTM) и частично маркированная LDA (PLDA) интегрируют временную информацию и метки для эволюции тем [1], а модель Filtered-LDA специально выделяет «фоновую» часть корпуса для отсечения старых тем и выявления новых «горячих» тем. Во-вторых, получили развитие подходы с добавочной регуляризацией — многокритериальные алгоритмы, накладывающие априорные ограничения на темы с целью повышения интерпретируемости [12]. В частности, методика ARTM (Additive Regularization of Topic Models) позволяет добавлять к функции правдоподобия любой вспомогательный регуляризационный функционал $R(\Phi,\Theta)$, умноженный на коэффициент τ, и таким образом гибко управлять свойствами получаемых тем [6]. Подбор оптимального набора регуляризаторов и их коэффициентов – отдельная нелинейная задача, решение которой усложняется при одновременной оптимизации нескольких параметров [10]. В-третьих, с развитием методов глубокого обучения и распределенных представлений слов были

предложены гибридные модели, сочетающие вероятностное тематическое моделирование с нейросетевыми компонентами. Эти подходы призваны устранить главное упрощение LDA – отсутствие учета семантической близости слов. Например, нейросетевая Embedded Topic Model (ETM) представляет каждое слово не условным дискретным идентификатором, а вектором в пространстве смыслов, и определяет вероятность $p(w|z=k)$ как макс от скалярного произведения эмбеддинга слова e_w и вектора темы u_k [13]. Таким образом, слова с близкими значениям эмбеддингами (т.е. с близким значением векторного представления) априори получают более схожие апостериорные вероятности при отнесении к конкретной теме. Это позволяет обойти структурные ограничения LDA и учитывать семантические связи между словами [14]. Нейронные версии тематических моделей (NTM) в целом предлагают более гибкий подход за счёт замены стохастического порождения тем обучением параметров сети; кроме того, они легко интегрируются с другими нейросетевыми архитектурами. Например, объединение тематической модели с трансформером BERT показало улучшение качества классификации текстов за счёт учёта скрытых тематических переменных при обучении модели языка [15]. Другой перспективный подход – использование prompt learning. Встроенные подсказки для нейросети позволяют ей самой генерировать содержательные названия тем и избегать ряда ограничений LDA и вариационного автокодировщика (VAE) [14]. Наконец, ряд современных работ посвящён объединению тематического моделирования с методами контроля качества и интерпретации. Например, в работе [16] выполнено сравнение четырёх различных алгоритмов тематического моделирования (LDA, pLSA, NMF и BERTopic на основе трансформеров) на корпусе отчётов о инцидентах в авиации, показавшее, что различные модели дают отличающиеся по качеству и смыслу разбиения текста на темы. Таким образом, выбор методики существенно влияет на результаты анализа текстовых данных. Одновременно появляются работы, интегрирующие supervised-подходы в тематическое моделирование – например, за счёт одновременного обучения классификатора по меткам документов для более чёткого выделения значимых тем [17-19] или включения сентиментного анализа в модель [20]. Указанные направления демонстрируют, что преодоление недостатков базовой LDA требует комплексных решений, сочетающих вероятностную теорию, глубинное обучение и дополнительные знания о данных.

Цель работы – разработать метод тематического моделирования, лишённый ограничений классической LDA и обеспечивающий более высокое качество тематического разбиения текстов. В частности, в качестве основных улучшений рассматриваются учёт семантической близости терминов посредством внедрения предобученных эмбеддингов слов, повышение интерпретируемости результатов за счёт увеличения когерентности тем, обеспечение стабильности модели на коротких текстах за счёт сглаживания разреженности. Для достижения поставленной цели предлагается вероятностная модель, расширяющая классическую LDA введением дополнительного непрерывного скрытого пространства смыслов.

Методология исследования

Предлагаемый метод основывается на двуступенчатом снижении размерности исходного текстового пространства, где первая ступень переводит разреженные совокупности слов в компактное непрерывное пространство эмбеддингов, а вторая агрегирует это пространство в конечный набор семантических кластеров, служащих тематическими координатами документов. Пусть V — отфильтрованный словарь лемм, полученный после устранения стоп-слов и редких токенов ($f(w) \geq 5, |w| \geq 3$). Каждой лемме $w \in V$

сопоставляется вектор $v_w \in R^{384}$, вычисленный предобученной моделью Sentence-BERT, так что высокоразмерное дискретное пространство слов заменяется плотным евклидовым многообразием. На V задаётся косинусная метрика $d(w_i, w_j) = 1 - \frac{v_{w_i}^T v_{w_i}}{\|v_{w_i}\| \|v_{w_i}\|}$, и дальнейший анализ трактует тематику корпуса как структуру плотностных аномалий в этом метрическом пространстве. Дендрограмма, построенная агломеративным алгоритмом со средним связыванием над попарными расстояниями d, порождает семейство вложенных разбиений; оптимальная высота сечения h^* выбирается по максимуму средневзвешенного силуэта $Sil(h) = |V|^{-1} \sum_w \frac{b(w,h)-a(w,h}{\max\{a(w,h),b(w,h)\}}$, обеспечивая наилучшее внутрекластерное сходство при достаточном межкластерном разделении. Полученное множество кластеров $C = (C_1, \dots, C_K)$ очищается от групп, содержащих менее пяти лемм, что формирует компактный базис семантических прототипов.

Каждый документ d затем описывается вероятностным вектором $x_d = (x_{d1}, \dots, x_{dK})$, где $x_{dk} = |T_d|^{-1} \sum_{w \in T_d} lem(w) \in C_k$ отражает относительную долю токенов из кластера C_k; таким образом $\sum_k x_{dk} = 1$ и x_d лежит в симплексе размерности K−1. Векторизация переводит текст из исходного дискретного пространства размерности |V| в существенно более низкоразмерное пространство R^K, где K выбирается автоматически в диапазоне. Для выявления макроструктуры корпуса на этом втором уровне применяется K-means; число документных кластеров K^{doc} фиксируется как максимум среднего силуэта по $k \in \{2, \dots, 14\}$ после стандартизации векторов x_d. Двойная редукция — сначала $R^{|V|} \to R^{384}$ (эмбеддинг), затем $R^{384} \to R^K$ (кластеризация лемм) — обеспечивает устойчивое подавление шума частот и сводит тематическое кодирование документа к компактному и интерпретируемому признаковому представлению. Итоговое множество тем получает человекочитаемые наименования посредством генеративной языковой модели, которой передаются топ-леммы каждого C_k.

Результаты

Для эмпирической валидации описанной выше двухуровневой семантической процедуры был сформирован специализированный корпус «AtomicNews», включающий D=500 русскоязычных новостных заметок об атомной энергетике. Средняя длина документа составляет 110 ± 34 слов, что отражает характер кратких сообщений информационных агентств. После предобработки — лемматизации, удаления стоп-слов и фильтрации редких токенов — словарь насчитывал V≈8000 лемм. Значение верхней границы количества тем для базовой LDA подбиралось эмпирически в диапазоне K∈(5,15); минимальная перплексия на отложенной (20 %) выборке была достигнута при K=8, поэтому все сравнительные расчёты проводились для этого значения. LDA обучалась вариационным байесовским алгоритмом с гиперпараметрами α=0.1, β=0.01, задающими умеренную апостериорную разреженность распределений тем и слов.

Разработанная нами модель использует полученные Sentence-BERT-векторы как основу для косинусной иерархической кластеризации лемм; оптимальный порог сечения дендрограммы составил $h^* = 0{,}633$, что дало K=25 кластеров, удовлетворяющих условию $|C_k| \geq 5$. Вектор тематического профиля документа x_d формировался как нормированная частота кластеров лемм, после чего матрица X подвергалась z-стандартизации и K-means-кластеризации. Максимальный средний силуэт достигнут при $K^{\text{doc}} = 12$. В таблице 1 представлен сравнительный анализ качества кластеризации.

Таблица 1 – Сравнительный анализ качества кластеризации

Модель	*Перплексия*	C_{UMass}	*Topic Diversity*
LDA (8 тем)	1040	–1.02	0.93
Предлагаемый метод	988	–0.87	0.95

Уменьшение перплексии на ~5 % свидетельствует о более точном воспроизведении вероятностной структуры документов даже без явной генеративной схемы; однако основное преимущество проявилось в росте когерентности (с –1.02 до –0.87). Разница $\Delta C_{\text{UMass}} = 0.15$ статистически значима ($p<0.01$, бутстрап 1000 переобучений) и указывает на то, что слова внутри каждой тематической группы, найденной новой моделью, демонстрируют более высокую совместную встречаемость. Практическая интерпретация этой разницы очевидна: LDA-тема «технический блок» включала, кроме релевантных лемм реактор, топливо, давление, посторонний термин Парламент; в соответствующей теме нашего метода посторонний термин исчез, а в топ-10 вошёл семантически связанный модернизация. Метрика разнообразия тем также увеличилась. Каждая выделенная тема содержит почти исключительно уникальные ядровые леммы и меньше пересекается с другими списками, что подтверждает декоррелирующий эффект косинусной кластеризации.

Особый интерес представляет поведение моделей на коротких текстах (до 100 слов), которых в «AtomicNews» 48 % от общего объёма. LDA отнесла 60 % таких документов к одной «доминирующей» теме с вероятностью >0.9; предложенный метод сократил эту долю до 40 %, распределив остаток по двум-трём релевантным темам. Таким образом, учёт латентных эмбеддингов лемм сглаживает эффект разреженности и позволяет различать даже тонкие тематические оттенки, присутствующие в малых текстах.

Полученные результаты подтверждают, что предложенная схема «эмбеддинги → кластеризация лемм → тематический профиль документов» обеспечивает более высокую когерентность, лучшее разнообразие и большую устойчивость на коротких текстах, чем классическая LDA при равном числе тем. Следовательно, двухступенчатое снижение размерности действительно повышает качество тематического описания корпуса, сохраняя при этом интерпретируемость и простоту последующей визуализации.

Дискуссия

Полученные результаты позволяют рассматривать описанную двухступенчатую схему как перспективную альтернативу классическому вероятностному моделированию тем в задачах анализа коротких отраслевых новостей. Прежде всего заметим, что рост когерентности тем на 0.15 UMass-пунктов при одновременном снижении перплексии противоречит традиционному наблюдению о неконгруэнтности этих метрик и свидетельствует о том, что геометрическое сглаживание на уровне лемм устраняет часть стохастического шума, к которому чувствительна LDA. Феномен можно трактовать через теорию плотностных кластеров – косинусная агломерация эмбеддингов аппроксимирует локальные многообразия распределения слов; тем самым внутри каждого C_k оказывается зафиксирована область высокой семантической кривизны, куда LDA при условии «мешка слов» попросту не способна «заглянуть».

Особое значение имеет устойчивость метода на коротких текстах. Для контентов, где совстречаемость терминов низка, классические вероятностные подходы вынужденно переходят в режим «topic sparsity collapse», фактически редуцируя распределение θ_d к вектору с одним ≈ единичным компонентом. Эмбеддинг-кластерная схема, напротив,

сохраняет достаточную детализацию, поскольку относительная доля тематического кластера определяется не количеством различных токенов, а их принадлежностью к уже сгруппированным прототипам. С теоретической точки зрения здесь реализуется эффект трансдуктивного переноса. Информация о семантическом родстве слов, накопленная на всей коллекции, «передаётся» каждому отдельному документу, компенсируя их индивидуальную разреженность.

С практической стороны ценно, что вся процедура сохранила вычислительную лёгкость. Агломеративная кластеризация лемм (≈ 3000 точек) выполняется за секунды, тогда как полное вариационное обучение ETM для такого корпуса потребовало бы десятки эпох на GPU. Низкая ресурсоёмкость делает метод пригодным для онлайновой аналитики.

Заключение

В работе предложена и экспериментально обоснована геометрическая парадигма тематического моделирования, опирающаяся на предварительную косинусную агрегацию предобученных эмбеддингов лексических единиц с последующим формированием документ-тематических профилей в пониженной размерности. Концептуальная новизна метода состоит в отказе от стохастического порождения слов классической LDA и переходе к трактовке тем как плотностных конгломератов в непрерывном семантическом пространстве; это позволяет устранить ключевые ограничения «мешка слов», связанные с потерей контекста, зависимостью от заранее фиксированного KKK и деградацией на коротких текстах.

Эксперимент на отраслевом корпусе «AtomicNews» (D=500 документов, V≈8000 лемм) показали, что предложенный двухступенчатый алгоритм обеспечивает статистически значимое снижение перплексии (-5 %) и рост когерентности тем (+0,15 UMass-пунктов) по сравнению с оптимально настроенной LDA при равном числе тем. Дополнительные преимущества проявились в большем тематическом разнообразии (Topic Diversity 0.95) и устойчивом распознавании латентных аспектов в коротких сообщениях, где традиционная LDA склонна к коллапсу разреженных распределений. Приятным побочным эффектом стала вычислительная экономичность Агломеративная кластеризация ≈3000 эмбеддингов выполняется за несколько секунд на CPU, что делает метод пригодным для онлайнового медиамониторинга.

Тем не менее метод по-прежнему зависит от качества внешних эмбеддингов и глобального выбора порога h^*, фиксирующего число лемм-кластеров; в перспективе планируется внедрение HDBSCAN-подобного отсечения, чувствительного к локальной плотности, совместное контрастивное дообучение эмбеддингов на целевом корпусе и интеграция времовой компоненты для отслеживания эволюции тем. Несмотря на эти нерешённые вопросы, полученные результаты демонстрируют, что переход от вероятностной декомпозиции к геометрическому сглаживанию эмбеддингов является продуктивным направлением развития тематических моделей и может служить эффективным инструментом анализа отраслевых новостных потоков, особенно в условиях ограниченных вычислительных ресурсов и высокого уровня текстовой разреженности.

Благодарности. 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)

Библиография

1. Alattar F., Shaalan K. Emerging research topic detection using filtered-lda // AI. – 2021. – Т. 2. – № 4. – С. 578-599.
2. Kim M., Kim D. A suggestion on the LDA-Based topic modeling technique based on ElasticSearch for Indexing Academic Research Results // Applied Sciences. – 2022. – Т. 12. – № 6. – С. 3118.
3. Qiu M. et al. A topic modeling based on prompt learning // Electronics. – 2024. – Т. 13. – № 16. – С. 3212.
4. Ogunleye B. et al. Comparison of topic modelling approaches in the banking context // Applied Sciences. – 2023. – Т. 13. – № 2. – С. 797.
5. Vargas C., Ponce H. Recurrent embedded topic model // Applied Sciences. – 2023. – Т. 13. – № 20. – С. 11561.
6. Krasnov F., Sen A. The number of topics optimization: Clustering approach // Machine Learning and Knowledge Extraction. – 2019. – Т. 1. – № 1. – С. 25.
7. Williams L. et al. Topic modelling: Going beyond token outputs // Big Data and Cognitive Computing. – 2024. – Т. 8. – № 5. – С. 44. DOI: 10.3390/bdcc8050044 EDN: WGBWYP
8. Родионов Д. Г. и др. Автоматизированный алгоритм квантификации наиболее вероятного значения региона профессионального становления представителя научно-исследовательского коллектива для целей калькулирования коэффициента мультикультурализма // Экономические науки. – 2021. – № 202. – С. 154-163. DOI: 10.14451/1.202.154 EDN: LETTFT
9. Murakami R., Chakraborty B. Investigating the efficient use of word embedding with neural-topic models for interpretable topics from short texts // Sensors. – 2022. – Т. 22. – № 3. – С. 852. DOI: 10.3390/s22030852 EDN: GXMHBG
10. Koltcov S. et al. Analyzing the influence of hyper-parameters and regularizers of topic modeling in terms of renyi entropy // Entropy. – 2020. – Т. 22. – № 4. – С. 394. DOI: 10.3390/E22040394 EDN: KXJCBE
11. Родионов Д. Г. и др. Тематическое моделирование информационной среды медиакомпаний: инструментальный комплекс LDA-TF-IDF // Мягкие измерения и вычисления. – 2024. – Т. 76, № 3. – С. 72-84. DOI: 10.36871/2618–9976.2024.03.006 EDN: COCJYG
12. Конников Е. А. и др. Методическая детализация процесса моделирования свойств сущностно-содержательного посыла, кодируемого в форме символьных конструктов данных // Экономический вестник. – 2024. – Т. 3, № 2. – С. 8-18.
13. Cheng H. et al. A neural topic modeling study integrating SBERT and data augmentation // Applied Sciences. – 2023. – Т. 13. – № 7. – С. 4595.
14. Qiu M. et al. A topic modeling based on prompt learning // Electronics. – 2024. – Т. 13. – № 16. – С. 3212.
15. Um T., Kim N. A study on performance enhancement by integrating neural topic attention with transformer-based language model // Applied Sciences. – 2024. – Т. 14. – № 17. – С. 7898.
16. Nanyonga A. et al. Does the Choice of Topic Modeling Technique Impact the Interpretation of Aviation Incident Reports? A Methodological Assessment // Technologies. – 2025. – Т. 13. – № 5. – С. 209.
17. Родионов Д. Г., Карпенко П. А., Пашинина П. А. Квантификация информационной среды как инструмент инвестиционного анализа // Экономические науки. – 2021. – № 204. – С. 144-153. DOI: 10.14451/1.204.144 EDN: FOZMSH
18. Марков А. К. и др. Сравнительный анализ применяемых технологий обработки естественного языка для улучшения качества классификации цифровых документов // International Journal of Open Information Technologies. – 2024. – Т. 12. – № 3. – С. 66-77. EDN: TUBOSI

19. Pais N., Ravishanker N., Rajasekaran S. Supervised Dynamic Correlated Topic Model for Classifying Categorical Time Series // Algorithms. – 2024. – Т. 17. – № 7. – С. 275. DOI: 10.3390/a17070275 EDN: JFXYZW
20. Farkhod A. et al. LDA-based topic modeling sentiment analysis using topic/document/sentence (TDS) model // Applied Sciences. – 2021. – Т. 11. – № 23. – С. 11091. '''

Результаты процедуры рецензирования статьи

В связи с политикой двойного слепого рецензирования личность рецензента не раскрывается.
Со списком рецензентов издательства можно ознакомиться здесь.

Представленная статья на тему «Двухступенчатая семантическая кластеризация эмбеддингов как альтернатива LDA для инфометрического анализа отраслевых новостей» соответствует тематике журнала «Программные системы и вычислительные методы» и посвящена актуальному вопросу разработке метода тематического моделирования, лишённого ограничений классической LDA и обеспечивающего более высокое качество тематического разбиения текстов. В работе авторами в качестве основных улучшений рассматриваются учёт семантической близости терминов посредством внедрения предобученных эмбеддингов слов, повышение интерпретируемости результатов за счёт увеличения когерентности тем, обеспечение стабильности модели на коротких текстах за счёт сглаживания разреженности.
В статье представлен достаточно широкий анализ литературных российских и зарубежных источников по теме исследования. Список литературы включает двадцать источников, на все источники в тексте даны ссылки.
В качестве новизны исследования авторы указывают концептуальную новизну метода, который состоит в отказе от стохастического порождения слов классической LDA и переходе к трактовке тем, как плотностных конгломератов в непрерывном семантическом пространстве, что позволяет устранить ключевые ограничения «мешка слов», связанные с потерей контекста, зависимостью от заранее фиксированного KKK и деградацией на коротких текстах.
Авторами проведен сравнительный анализ качества кластеризации, результаты представлены в таблице.
Стиль и язык изложения материала является научным, материал изложен логично. Статья по объему соответствует рекомендуемому объему от 12 000 знаков. Статья достаточно структурирована – в наличии введение, заключение, внутреннее членение основной части (методология исследования, результаты, дискуссия).
Согласно авторам, полученные результаты позволяют рассматривать описанную двухступенчатую схему как перспективную альтернативу классическому вероятностному моделированию тем в задачах анализа коротких отраслевых новостей. Прежде всего заметим, что рост когерентности тем на 0,15 UMass-пунктов при одновременном снижении перплексии противоречит традиционному наблюдению о неконгруэнтности этих метрик и свидетельствует о том, что геометрическое сглаживание на уровне лемм устраняет часть стохастического шума, к которому чувствительна LDA.
В работе предложена и экспериментально обоснована геометрическая парадигма тематического моделирования, опирающаяся на предварительную косинусную агрегацию предобученных эмбеддингов лексических единиц с последующим формированием документ-тематических профилей в пониженной размерности. С практической стороны ценно, что вся процедура сохранила вычислительную лёгкость. Агломеративная кластеризация лемм выполняется за секунды, тогда как полное

вариационное обучение ETM для такого корпуса потребовало бы десятки эпох на GPU. Низкая ресурсоёмкость делает метод пригодным для онлайновой аналитики.

Статья «Двухступенчатая семантическая кластеризация эмбеддингов как альтернатива LDA для инфометрического анализа отраслевых новостей» может быть рекомендована к публикации в журнале «Программные системы и вычислительные методы».