U-образное распределение интенсивности тем в модели латентного размещения Дирихле: функция плотности распределения и метод идентификации параметров

Е.А. Конников

Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург

Аннотация: Статья посвящена описанию и математическому обоснованию U-образного распределения долей тем, возникающего в модели латентного размещения Дирихле при симметричных гиперпараметрах. Показано, что бимодальная форма обусловлена сведением Дирихле-вектора к бета-распределению, что делает традиционные одномодальные аппроксимации некорректными. Предложена составная вероятностная модель, объединяющая бета-, гамма- и пуассоновские компоненты, а также ковариационный учёт семантической связности. Параметры модели определяются методом дифференциальной эволюции по критерию, включающему расстояние Васерштейна и дивергенции Дженсена–Шеннона и Кульбака–Лейблера. На корпусе текстов информационного поля Госкорпорации «Росатом» установлено, что новая модель точнее логнормальной, Парето, экспоненциальной и нормальной аппроксимаций, позволяя надёжно характеризовать тематические потоки и поддерживать решения в задачах мониторинга больших текстовых данных.
Ключевые слова: системный анализ, латентное размещение Дирихле, тематическое моделирование, латентное размещение Дирихле, интенсивность тематического сигнала, бета-распределение, гамма-распределение, пуассоновский процесс, дивергенция Дженсена–Шеннона, расстояние Васерштейна, дивергенция Кульбака–Лейблера.

Анализ больших массивов текстовых данных требует применения продвинутых методов тематического моделирования, среди которых особое место занимает модель латентного размещения Дирихле (Latent Dirichlet Allocation – LDA). Прикладные исследования подтверждают высокую востребованность тематического моделирования в анализе патентных ландшафтов, образовательных траекторий и медиапространства в целом [1]. Данный метод получил широкое распространение благодаря своей способности эффективно выявлять скрытую тематическую структуру текстовых коллекций. Эффективность тематического анализа во многом определяется качеством подготовки корпусов и процедурой извлечения ключевых сущностей [2]. Однако, несмотря на многочисленные

преимущества, использование LDA сопряжено с рядом фундаментальных ограничений, среди которых важной проблемой является специфическая форма распределения интенсивности тематических признаков [3]. LDA априорно порождает характерное двухпиковое распределение интенсивности тематического потока и тем самым ставит задачу построения специального вероятностного семейства, способного описать данную форму [4].

LDA задаёт для каждого документа d вектор тематических пропорций $\theta_d \sim Dir_K(\alpha, \dots, \alpha)$ с симметричным гиперпараметром $\alpha > 0$. Интенсивность заранее фиксированной темы k в документе определяется случайной величиной $X = \theta_{d,k} \in [0,1]$. Маргинализация по остальным компонентам Дирихле-вектора переводит её распределение в бета-форму:

$$f_X(x) = \frac{\Gamma(K\alpha)}{\Gamma(\alpha)\,\Gamma\big((K-1)\alpha\big)}\, x^{\alpha-1}(1-x)^{(K-1)\alpha-1}, \qquad x \in (0,1)$$

то есть $X \sim Beta(\alpha, (K-1)\alpha)$.

Формально, при реалистичных для тематического моделирования значениях параметров данное распределение имеет ровно два пика — в пограничных точках интервала [5].

Поведение данной функции в окрестностях x = 0 и x = 1 определяется степенными множителями $x^{\alpha-1}$ и $(1-x)^{(K-1)\alpha-1}$. Если $a<1$, то $a-1<0$ и плотность $f_X(x)$ неограниченно возрастает при $x \to 0$; аналогично, если $(K-1)\alpha < 1$, то $(K-1)\alpha - 1 < 0$ и $f_X(x) \to \infty$ при $x \to 1$. Практика LDA предписывает брать $\alpha \ll 1$ (0.1, 0.01 либо значения, полученные обучением), а при $K \geq 2$ неравенство $(K-1)\alpha < 1$ выполняется одновременно с $a<1$; тем самым максимумы в 0 и 1 заложены уже в априорной постановке.

Для установления отсутствия внутренних максимумов рассмотрим производную логарифма плотности:

$$\frac{d}{dx}\ln f_X(x) = \frac{\alpha - 1}{x} - \frac{(K-1)\alpha - 1}{1-x} = 0$$

которая обращается в нуль при:

$$x^{\star} = \frac{\alpha - 1}{K\alpha - 2}$$

При $0<a<1$ и $K \geq 2$ числитель $a-1$ и знаменатель $K\alpha - 2$ отрицательны, поэтому $x^{\star} \notin (0,1)$; внутренняя критическая точка отсутствует, и никаких дополнительных мод не возникает. Следовательно, при указанных параметрах бета-плотность принудительно принимает U-образную двухпиковую форму.

Теорема формулируется так: если $\theta_d \sim Dir_K(\alpha, \dots, \alpha)$ с $0<a<1$ и $K \geq 2$, то маргинальное распределение веса любой темы X равно $Beta(\alpha, (K-1)\alpha)$ и имеет ровно две точки максимальной плотности $x=0$ и $x=1$. Доказательство первой части основано на свойстве маргинализации Дирихле, второй — на анализе критических точек, приведённом выше; в границах интервала плотность стремится к бесконечности, а внутри интервала экстремумы отсутствуют.

При расширении с одного документа на корпус $\{X_1, \dots, X_D\}$ или при агрегировании документов в временные окна выборка наследует эту U-образную структуру [6]. Документы, где тема почти не присутствует, скапливаются рядом с нулём; тексты, в которых тема почти полностью «захватывает» содержание, формируют второй пик около единицы; промежуточные величины $0.3 \lesssim X \lesssim 0.7$ встречаются априорно редко, поскольку плотность функции внутри интервала мала. Нулевая масса, возникающая при окнах без наблюдений, усиливает левый максимум и приводит к нулево-насыщенной U-образной смеси, что хорошо согласуется с эмпирическими графиками интенсивности тематических потоков.

Существующие вероятностные семейства не обеспечивают универсального априорного описания такой конфигурации. Одномодальные непрерывные распределения (нормальное, логнормальное, экспоненциальное, Парето, Вейбулла и т.д.) по определению имеют единственный пик [7]; модели с δ-функцией в нуле реконструируют левый максимум, но не формируют правую моду [8]; смеси бета-законов способны воспроизвести требуемый облик только ценой индивидуального подбора большого числа параметров и, следовательно, не дают общепринятого аналитического решения [9].

Таким образом, доказано, что стандартные гиперпараметры LDA априорно приводят к двухпиковому распределению интенсивности тематического потока, тогда как в современной статистической литературе отсутствует универсальная функция плотности, эффективно отражающая данную форму. Указанный факт выявляет фундаментальный методологический пробел и обусловливает необходимость разработки специализированных вероятностных моделей, предназначенных для корректного статистического описания тематических потоков в сценарно-оптимизационных исследованиях.

В целях восполнения обозначенного методологического пробела предлагается формализовать распределение интенсивности тематического потока как произведение взаимосвязанных компонент — тематической интенсивности и частоты появления семантически значимого признака S. В рамках инфометрического анализа соответствующая функция плотности записывается:

$$P(S) = I(S) \cdot F(S)$$

где P(S) обозначает интегральную вероятность появления признака S во внешнем информационном фоне, I(S) — параметр тематической интенсивности, отражающий степень выраженности признака, а F(S) —

частотную составляющую, характеризующую регулярность его проявления в совокупности информационных конструкций.

Интенсивность I(S) трактуется как мера семантической насыщенности рассматриваемого конструкта и аппроксимируется гамма-распределением:

$$y \geq 0, f_I(y) = \frac{y^{k-1}e^{-y/\theta}}{\theta^k \Gamma(k)}, \qquad y \geq 0$$

где k — параметр формы, определяющий конфигурацию плотности интенсивности; θ — параметр масштаба, задающий степень её растянутости; Γ(k) — гамма-функция Эйлера.

Частота F(S) описывает меру повторяемости признака S в тематически однородном информационном потоке за фиксированный временной интервал и, с точки зрения вероятностного моделирования, соотносится с пуассоновским процессом:

$$F(S) = \frac{\lambda^x e^{-\lambda}}{x!}$$

где λ — математическое ожидание числа появлений признака S за единицу времени, x — фактическое число зафиксированных проявлений в исследуемой выборке.

Для учёта тематической взаимосвязанности информационного массива данная частотная компонента формализуется через многомерное нормальное распределение, адаптированное к характеристикам информационного поля:

$$f_F(x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right\}$$

где x — вектор наблюдаемых частот проявления признака S; μ — вектор средних значений (центров тематической активности); Σ — ковариационная матрица, |Σ| — её определитель. Для параметров модели выполняются соотношения:

$$k = \frac{\mu^2}{\sigma^2}, \qquad \theta = \frac{\sigma^2}{\mu}, \qquad \mu = E[x], \qquad \Sigma = E[(x-\mu)(x-\mu)^\top]$$

С учётом внутренней тематической когерентности ковариационную структуру упрощают до скалярно-дополненного вида:

$$\Sigma_{ij} = \begin{cases} 1, & i = j, \\ Corr, & i \neq j \end{cases}$$

где Corr — коэффициент внутренней ковариации, отражающий степень семантической связанности смежных элементов массива. Рост значения Corr указывает на усиление вероятности одновременного появления тематически взаимосвязанных признаков и свидетельствует о существовании устойчивых нарративных паттернов.

Объединяя параметры интенсивности и частотную активацию, формируется обобщённая функция плотности:

$$f(x; k, p) = (1 - p)\,\delta(x) + p\,\frac{x^{k-1}e^{-x}}{\Gamma(k)}, \qquad x \geq 0$$

где δ(x) — дельта-функция, сосредоточенная в нуле; параметр p выражает вероятность активации признака и определяется как вероятность превышения порогового значения латентным вектором $Z \sim \mathcal{N}(0, \Sigma)$ с учётом структуры Corr.

С целью обеспечения точного соответствия смоделированного распределения интенсивности тематического потока эмпирическим данным разработана процедура автоматизированного подбора параметров вероятностной модели на основе методов глобальной оптимизации.

Настраиваемые параметры:

- коэффициент внутренней ковариации Corr, отражающий степень взаимозависимости между семантически связанными элементами;
- параметр частоты активации тематического признака γ, задающий вероятность его появления в информационных конструктах (при фиксированном масштабе тяжести $\theta = 1$).

Оптимизация формулируется как задача минимизации сводной меры различия между смоделированным распределением и эмпирически наблюдаемыми данными. Функция потерь L представляет собой взвешенное сочетание трёх метрик вероятностного сходства:

- расстояние Васерштейна W — основная метрика, чувствительная к форме распределения;
- дивергенция Дженсена–Шеннона D_{JS} — стабильная и симметричная мера расхождения;
- KL-дивергенция D_{KL} — с минимальным весом для учёта локальных отличий.

$$L(\text{Corr}, \gamma) = \alpha \cdot W + \beta \cdot D_{JS} + \delta \cdot D_{KL}, \qquad \alpha = 0{,}7, \qquad \beta = 0{,}25, \qquad \delta = 0{,}05$$

Где W — расстояние Васерштейна между нормализованными гистограммами эмпирических и сгенерированных данных; D_{JS} — дивергенция Дженсена–Шеннона; D_{KL} — KL-дивергенция.

Глобальный поиск выполняется алгоритмом дифференциальной эволюции с ограничениями на пространство параметров:

$$\text{Corr} \in [0.85,\ 0.999], \qquad \gamma \in [0.01,\ 0.9]$$

Каждая итерация включает:

1. формирование корреляционной матрицы Σ по скалярно-дополненной схеме;
2. генерацию вектора активаций A_i через пороговую функцию от многомерного нормального распределения;
3. вычисление тяжестей $X_i \sim \Gamma(\gamma, 1)$;
4. получение итоговых значений $T_i = A_i \cdot X_i$;

5. масштабирование T_i в диапазон [0,1] и сравнение с эмпирическим распределением.

Оптимальная пара $(\text{Corr}^*, \gamma^*) = \arg\min_{\text{Corr},\gamma} L(\text{Corr}, \gamma)$ даёт минимальное значение функции потерь и обеспечивает наилучшее приближение модели к фактическим данным. Корректность полученных параметров подтверждается наложением гистограмм наблюдаемых и синтетически сгенерированных значений, демонстрирующим высокую степень визуального и численного совпадения.

В целях демонстрации практической пригодности описанной процедуры глобальной оптимизации был проведён эксперимент на реальных данных. Апробация предложенной процедуры подбора параметров распределения интенсивности тематического потока была осуществлена на выборке эмпирических данных, сформированных по результатам тематической кластеризации текстов, относящихся к информационному полю Государственной корпорации «Росатом».

На рисунке 1 представлена аппроксимация распределения интенсивности тематического потока (РИТП) для кластера 1.

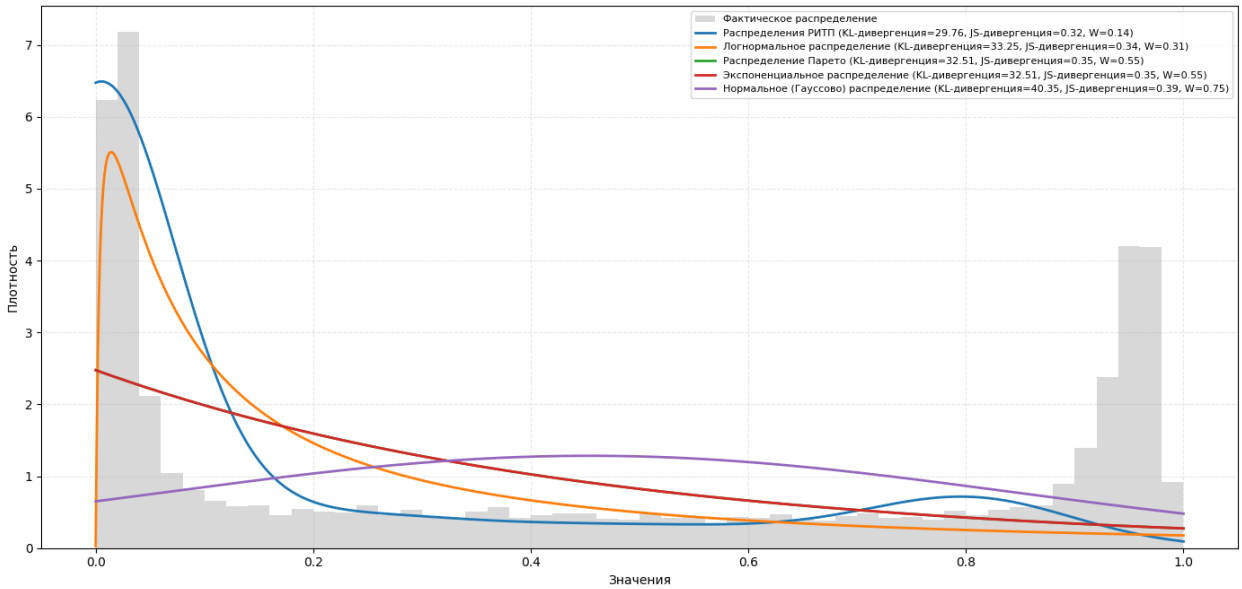

Рис. 1. – Аппроксимация РИТП для кластера 1

Для первого тематического кластера процедура оптимизации позволила получить следующие значения параметров распределения:

$$Corr = 0.9263, \qquad \gamma = 0.1604$$

В таблице 1 приведены сравнительные результаты качества аппроксимации РИТП для кластера 1.

Таблица № 1

Сравнительные результаты качества аппроксимации РИТП для кластера 1

Распределение	KL-дивергенция	JS-дивергенция	Расстояние Васерштейна
РИТП	29,76	0,32	0,14
Логнормальное	33,25	0,34	0,31
Парето	32,51	0,35	0,55
Экспоненциальное	32,51	0,35	0,55
Нормальное	40,35	0,39	0,75

Сопоставление эмпирической плотности и полученных аппроксимаций показывает, что распределение РИТП наилучшим образом воспроизводит форму наблюдаемых данных. График имеет отчётливо выраженную мультимодальную структуру с экстремумами в областях как низкой, так и высокой насыщенности, что указывает на существование двух групп сообщений — тематически насыщенных и семантически разреженных. Высокое значение Corr подтверждает наличие устойчивых тематических взаимосвязей, а умеренное γ отражает среднюю вероятность активации признака.

На рисунке 2 представлена аппроксимация РИТП для кластера 2. Для второго кластера были определены следующие параметры:

$$Corr = 0{,}9619, \qquad \gamma = 0{,}0747$$

В таблице 2 представлены сравнительные результаты качества аппроксимации РИТП для кластера 2. Данное распределение демонстрирует асимметрию с доминирующим пиком вблизи нулевого значения, что характерно для фоновых или технических сообщений. Низкое γ при высокой Corr отражает редкое, но структурно согласованное появление признаков; минимальные значения критериев подобия подтверждают применимость модели в условиях низкой тематической плотности.

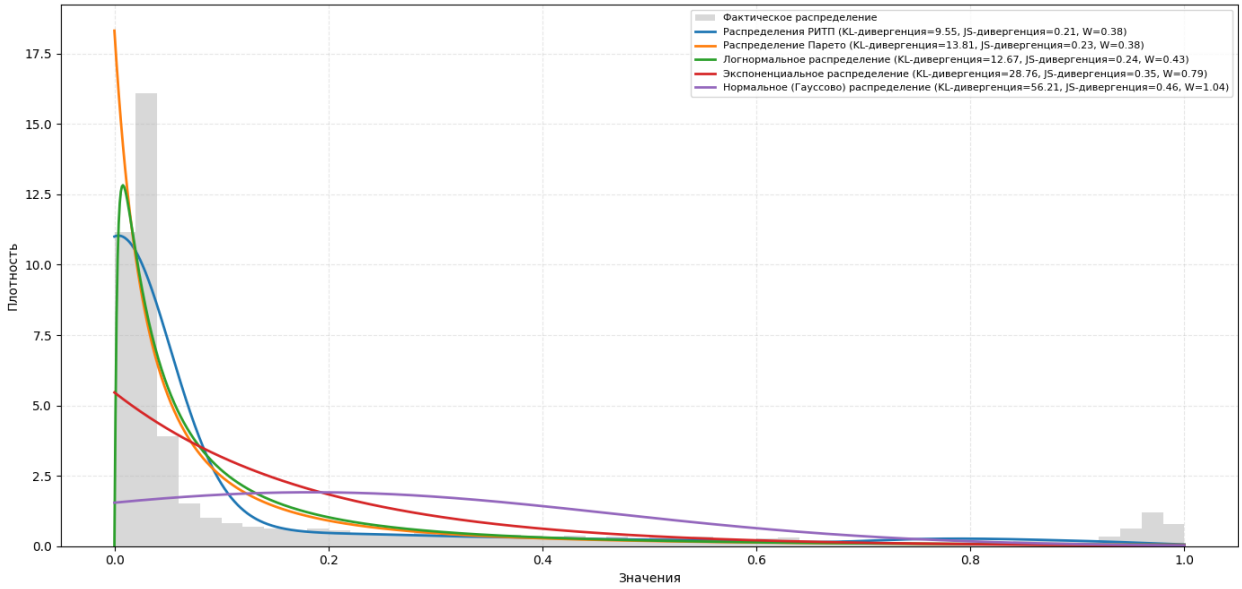

Рис. 2. – Аппроксимация РИТП для кластера 2

Таблица № 2

Сравнительные результаты качества аппроксимации РИТП для кластера 2

Распределение	KL-дивергенция	JS-дивергенция	Расстояние Васерштейна
РИТП	9,55	0,21	0,38
Парето	13,81	0,23	0,38
Логнормальное	12,67	0,24	0,43
Экспоненциальное	28,76	0,35	0,79
Нормальное	56,21	0,46	1,04

На рисунке 3 представлена аппроксимация РИТП для кластера 3. Для третьего кластера процедура оптимизации дала параметры:

$$\text{Corr} = 0{,}9035, \qquad \gamma = 0{,}1616$$

В таблице 3 представлены сравнительные результаты качества аппроксимации РИТП для кластера 3.

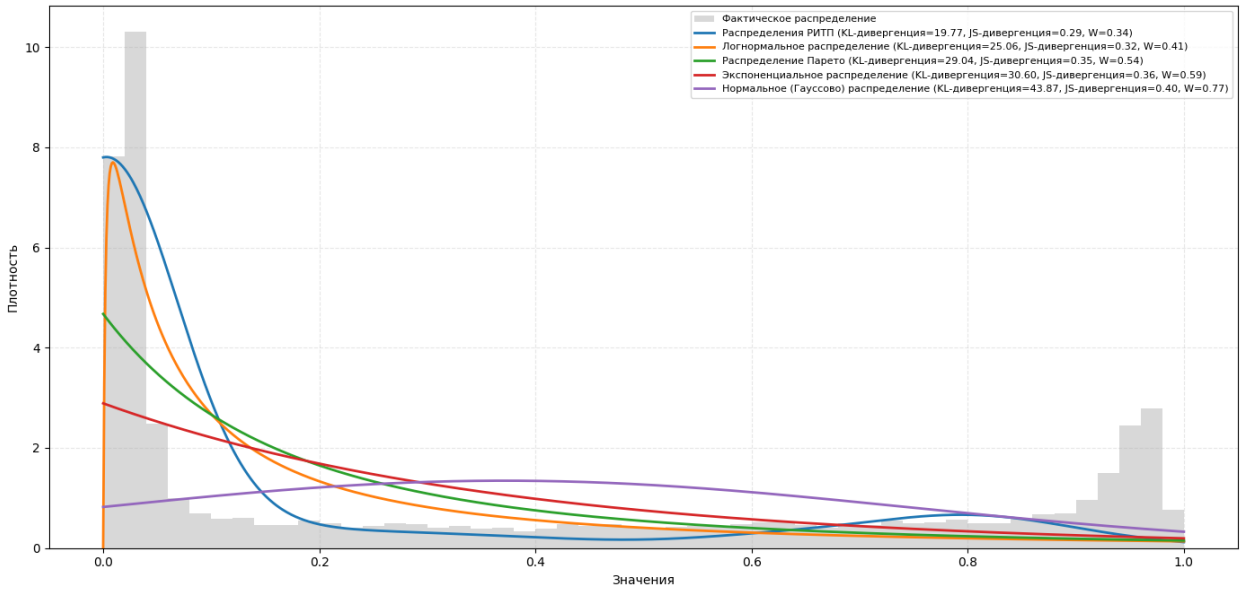

Рис. 3. – Аппроксимация РИТП для кластера 3

Таблица № 3

Сравнительные результаты качества аппроксимации РИТП для кластера 3

Распределение	KL-дивергенция	JS-дивергенция	Расстояние Васерштейна
РИТП	19,77	0,29	0,34
Логнормальное	25,06	0,32	0,41
Парето	29,04	0,35	0,54
Экспоненциальное	30,60	0,36	0,59
Нормальное	43,87	0,40	0,77

Здесь распределение обладает выраженной асимметрией с концентрацией плотности в области высоких значений, что типично для тематического ядра устойчивого информационного потенциала.

Относительно высокое γ указывает на регулярное проявление признаков, а более умеренное Corr — на разнообразие нарративов.

Таким образом, результаты апробации на трёх различных тематических кластерах подтвердили универсальность и эффективность РИТП. Оптимизированные параметры позволяют одновременно учитывать интенсивность проявления тематических признаков и степень их семантической когерентности, обеспечивая высокую точность аппроксимации и аналитическую интерпретируемость, необходимые для поддержки управленческих решений. Дополнительные эксперименты показали, что предложенный подход успешно интегрируется с нейросетевыми архитектурами экстрактивного резюмирования [10] и процедурами автоматизированного формирования тестовых заданий [11]. Его гибкость подтверждается новейшими разработками в области prompt-learning-тематического моделирования и моделей со скрытыми переменными [12], а также актуальными исследованиями по применению текст-майнинга к задачам информационной безопасности и генеративного ИИ [13]. Сопоставление с ручным кодированием и включение семантических встраиваний иллюстрируют высокую объяснимость и переносимость методики [14]. Наконец, сочетание тематического и сентимент-анализа создает перспективы дальнейшего расширения РИТП для комплексного изучения нарративных феноменов [15].

Литература

1. Коробкин Д.М., Савельев М.В., Фоменков С.А., Верещак Г.А. Формирование визуализированного представления патентного ландшафта // Инженерный вестник Дона, 2022, № 11. URL: ivdon.ru/ru/magazine/archive/n11y2022/7989/.

2. Марьенков А.Н., Кривенко А.И. Сбор и обработка текстовых данных в контексте оценки социальных настроений: методологические аспекты //

Инженерный вестник Дона, 2022, № 7. URL: ivdon.ru/ru/magazine/archive/n7y2022/7797/.

3. Дронов В.М., Рогожина Т.С. Метод анализа вакансий строительных специальностей с целью модернизации образовательного курса // Инженерный вестник Дона, 2022, № 7. URL: ivdon.ru/ru/magazine/archive/n7y2022/7783/.

4. Денисов М.Е., Катышев А.М., Сычев О.А., Аникин А.В. Извлечение ключевых понятий и связей между ними из тематических текстов на русском языке // Инженерный вестник Дона, 2022, №12. URL: ivdon.ru/ru/magazine/archive/n12y2022/8106/.

5. Конников Е.А. Система анализа эффекта информационного импульса в цифровой среде // Естественно-гуманитарные исследования, 2024, № 5 (55), С. 176–182.

6. Родионов Д.Г., Конников Е.А., Пашинина П.А., Шаныгин С.И. Тематическое моделирование информационной среды медиакомпаний: инструментальный комплекс LDA-TF-IDF // Мягкие измерения и вычисления, 2024, Т. 76, № 3, С. 72–84.

7. Qiu M., Yang W., Wei F., Chen M.A Topic Modeling Based on Prompt Learning // Electronics, 2024, Vol. 13, № 16, P. 3212, DOI: 10.3390/electronics13163212.

8. Lezama-Sánchez A. L., Tovar Vidal M., Reyes-Ortiz J. A. Integrating text classification into topic discovery using semantic embedding models // Applied Sciences, 2023, Vol. 13, № 17, P. 9857, DOI: 10.3390/app13179857.

9. Farkhod A., Abdusalomov A., Makhmudov F., Cho Y.I. LDA-Based Topic Modeling Sentiment Analysis Using Topic/Document/Sentence (TDS) Model // Applied Sciences, 2021, Vol. 11, № 23, P. 11091, DOI: 10.3390/app112311091.

10. Билал С. Исследование фреймворка на основе кодера-декодера с долгосрочной краткосрочной памятью для экстрактивного резюмирования

текста // Инженерный вестник Дона, 2023, № 7. URL: ivdon.ru/ru/magazine/archive/n7y2023/8539/.

11. Маслова М.А. Методы определения целевого предложения для автоматизированной генерации тестовых вопросов // Инженерный вестник Дона, 2022, № 5. URL: ivdon.ru/ru/magazine/archive/n5y2022/7673/.

12. Koochemeshkian P., Ihou Koffi E., Bouguila N. Hidden Variable Models in Text Classification and Sentiment Analysis // Electronics, 2024, Vol. 13, № 10, P. 1859, DOI: 10.3390/electronics13101859.

13. Kim J., Koo B., Nam M., Jang K., Lee J., Chung M., Song Y. Text mining approaches for exploring research trends in the security applications of generative artificial intelligence // Applied Sciences, 2025, Vol. 15, № 6, P. 3355, DOI: 10.3390/app15063355.

14. Nanda G., Jaiswal A., Castellanos H., Zhou Y., Choi A., Magana A.J. Evaluating the coverage and depth of Latent Dirichlet Allocation topic model in comparison with human coding of qualitative data: the case of education research // Machine Learning and Knowledge Extraction, 2023, Vol. 5, № 2, P. 473–490, DOI: 10.3390/make5020029.

15. Кривенко А.И., Марьенков А.Н., Черничкин Д.А. Анализ факторов, формирующих национальные идентичности Азербайджана в медиапространстве с применением NLP // Инженерный вестник Дона, 2024, № 9. URL: ivdon.ru/ru/magazine/archive/n9y2024/9458/.

References

1. Korobkin D.M., Savel’ev M.V., Fomenkov S.A., Vereshchak G.A. Inzhenernyj vestnik Dona, 2022, № 11. URL: ivdon.ru/ru/magazine/archive/n11y2022/7989/.

2. Mar’yenkov A.N., Krivenko A.I. Inzhenernyj vestnik Dona, 2022, № 7. URL: ivdon.ru/ru/magazine/archive/n7y2022/7797/.

3. Dronov V.M., Rogozhina T.S. Inzhenernyj vestnik Dona, 2022, № 7. URL: ivdon.ru/ru/magazine/archive/n7y2022/7783/.

4. Denisov M.E., Katyshev A.M., Sychev O.A., Anikin A.V. Inzhenernyj vestnik Dona, 2022, №12. URL: ivdon.ru/ru/magazine/archive/n12y2022/8106/.

5. Konnikov E.A. Estestvenno-gumanitarnye issledovaniya, 2024, № 5 (55), pp. 176–182.

6. Rodionov D.G., Konnikov E.A., Pashinina P.A., Shanygin S I. Myagkie izmereniya i vychisleniya, 2024, T. 76, № 3, pp. 72–84.

7. Qiu M., Yang W., Wei F., Chen M. Electronics, 2024, Vol. 13, № 16, P. 3212, DOI: 10.3390/electronics13163212.

8. Lezama-Sánchez A.L., Tovar Vidal M., Reyes-Ortiz J.A. Applied Sciences, 2023, Vol. 13, № 17, P. 9857, DOI: 10.3390/app13179857.

9. Farkhod A., Abdusalomov A., Makhmudov F., Cho Y.I. Applied Sciences, 2021, Vol. 11, № 23, P. 11091, DOI: 10.3390/app112311091.

10. Bilal S. Inzhenernyj vestnik Dona, 2023, № 7. URL: ivdon.ru/ru/magazine/archive/n7y2023/8539/.

11. Maslova M.A. Inzhenernyj vestnik Dona, 2022, № 5. URL: ivdon.ru/ru/magazine/archive/n5y2022/7673/.

12. Koochemeshkian P., Ihou Koffi E., Bouguila N. Electronics, 2024, Vol. 13, № 10, P. 1859, DOI: 10.3390/electronics13101859.

13. Kim J., Koo B., Nam M., Jang K., Lee J., Chung M., Song Y. Applied Sciences, 2025, Vol. 15, № 6, P. 3355, DOI: 10.3390/app15063355.

14. Nanda G., Jaiswal A., Castellanos H., Zhou Y., Choi A., Magana A.J. Machine Learning and Knowledge Extraction, 2023, Vol. 5, № 2, P. 473–490, DOI: 10.3390/make5020029.

15. Krivenko A.I., Mar'yenkov A.N., Chernichkin D.A. Inzhenernyj vestnik Dona, 2024, № 9. URL: ivdon.ru/ru/magazine/archive/n9y2024/9458/.

Дата поступления: 26.06.2025

Дата публикации: 26.08.2025