Сравнение методов векторизации названий товаров: Компромисс между точностью и вычислительной эффективностью в e-commerce

Ф.В. Краснов

Аннотация — **Настоящая статья представляет собой эмпирическое исследование и сравнительный анализ эффективности современных методов векторизации терминов в контексте задачи Information Retrieval (IR), сфокусированное на обработке коротких текстовых данных, представленных в виде названий товаров. Целью работы является определение оптимального метода, способного наиболее точно воспроизводить глобальную структуру семантических связей корпуса при сохранении высокой вычислительной эффективности. В качестве ключевого критерия оценки выбрана Норма Фробениуса ($\|\cdot\|_F$) разницы между нормализованной целевой матрицей совстречаемости терминов (C_{target}) и матрицей косинусной близости, полученной из векторных представлений (C_{hat}).**

Исследование было проведено в три последовательных этапа. Первый эксперимент включал сравнительную оценку классических методов матричной факторизации (SVD/LSA, NMF, LDA) и моделей локального окна (Word2Vec, FastText), используя базовую токенизацию по пробелам. На этом этапе алгоритм LDA продемонстрировал минимальную ошибку (191.00), указывая на его наибольшее соответствие глобальной структуре корпуса.

На втором этапе для всесторонней оценки была применена токенизация, совместимая с BERT (BPE-подобная), а к сравнению добавлена предобученная контекстная модель-трансформер BERT. Для обеспечения методологической чистоты, BERT был оценен в режиме статического усреднённого эмбеддинга (фиксированное векторное представление). Экспериментальные данные подтвердили, что алгоритм LDA сохраняет лидерство с ошибкой 156.9 показывая более высокую точность в данной задаче, чем модель BERT, которая достигла ошибки 253.17.

Третий эксперимент был посвящен многокритериальной оптимизации гиперпараметров наиболее эффективного метода LDA. С использованием библиотеки Optuna был найден Парето-Фронт решений, отражающий оптимальный компромисс между внутренней согласованностью (max Log-Likelihood) и эмпирической точностью (min Норма Фробениуса).

Полученные результаты подтверждают, что для задач IR, не требующих глубокого контекстуального понимания, методы, основанные на факторизации глобальных частотных связей (LDA), являются наиболее экономически и технически оправданными, превосходя сложные нейросетевые модели по ключевой метрике воспроизведения семантической структуры.

Ключевые слова — **Векторизация терминов, Information Retrieval, Норма Фробениуса, Бинарная матрица, Совстречаемость, SVD, NMF, LDA, Word2Vec, FastText, BERT.**

I. Введение

Развитие систем Information Retrieval (IR) и машинного обучения привело к фундаментальному сдвигу в подходах к обработке текстовых данных, особенно в коммерческом секторе. В сфере электронной коммерции (e-commerce) высокоточное семантическое понимание пользовательских запросов и названий товаров является критически важным для достижения бизнес-целей.

Векторные представления (эмбеддинги) лежат в основе множества ключевых функций, включая мачинг товаров для определения дубликатов в каталоге, генерацию поисковых подсказок (autocompletion), исправление опечаток с учетом контекста, семантическую сортировку результатов поиска и повышение релевантности рекомендательных систем. Традиционно эти задачи решались методами, основанными на частотном анализе и матричной факторизации, которые обеспечивали высокую точность и вычислительную эффективность. Классические задачи IR, такие, как поиск по ключевым словам и ранжирование, имеют прочную теоретическую базу и решаются с приемлемой точностью уже более пятидесяти лет, о чем свидетельствует долговечность таких моделей, как BM25 и TF-IDF [1].

Однако современный фокус исследований в области Natural Language Processing (NLP) смещен в сторону более сложных задач, таких как генерация текста, машинный перевод и тонкий анализ эмоциональной окраски, требующих глубокого контекстуального понимания. В результате, наиболее современные методы векторизации предлагают энкодеры на основе нейронных сетей с архитектурой Трансформера, которые формируют эмбеддинги с учетом глубокого контекста. Это создает предположение о безусловной

Статья получена 1 декабря 2025.

Ф.В.Краснов, fkrasnov2@yandex.ru, http://orcid.org/0000-0002-9881-7371.

универсальности таких моделей: исследователи и разработчики часто допускают, что более сложная модель автоматически гарантирует превосходный результат даже в задачах, не требующих глубокого контекста, таких как векторизация коротких названий товаров.

Указанное допущение сопряжено с прямыми экономическими рисками для сектора e-commerce. Использование сложных предобученных Трансформеров (например, BERT) для простых задач приводит к неэффективному использованию вычислительных ресурсов из-за необходимости поддержания крупномасштабной модели в режиме инференса. Более того, это удлиняет время разработки и усложняет обслуживание системы из-за высокой ресурсоемкости обучения, тонкой настройки и отладки глубоких нейронных сетей. Например, при мачинге товаров достаточно определить, что "Apple iPhone 13 Pro 256GB" семантически близок к "iPhone 13 Pro Apple 256 GB", чего часто можно добиться с помощью эффективной матричной факторизации.

Таким образом, возникает острая необходимость в проведении строгого сравнительного анализа, который бы оценил, насколько выигрыш в точности, предоставляемый глубокими контекстными моделями, оправдывает многократное увеличение их вычислительных требований по сравнению с проверенными и быстрыми методами, такими как SVD, NMF, Word2Vec и FastText. Данное исследование направлено на устранение этого пробела, предоставляя эмпирически обоснованные рекомендации для выбора оптимальной стратегии векторизации в e-commerce на основе метрики, отражающей глобальную структуру совстречаемости терминов.

На основе анализа существующих подходов, представленного в следующем разделе, мы перейдем к детальному описанию выбранных методов, что позволит глубже понять их применимость в контексте e-commerce.

II. Обзор моделей векторизации терминов

Анализ существующих подходов к векторизации терминов демонстрирует эволюцию от статистических методов, основанных на частотах, к нейросетевым архитектурам, способным учитывать сложный лингвистический контекст.

Классическая модель Латентно-семантического анализа (LSA), основанная на сингулярном разложении матрицы Документ-Терм (SVD) [2], заложила основы для снижения размерности и выявления скрытых семантических факторов.

Неотрицательная матричная факторизация (NMF) [3] была предложена как альтернатива, обеспечивающая более интерпретируемые компоненты за счет наложения ограничения неотрицательности.

Метод Латентного размещения Дирихле (LDA) [4] является статистической моделью, которая аппроксимирует распределение терминов по темам и, по сути, также формирует векторное представление терминов в пространстве тем.

Последующий прорыв в области эмбеддингов был связан с появлением моделей, основанных на локальных окнах. Модель Word2Vec (W2V), представленная Миколовым и соавторами [5], впервые продемонстрировала возможность обучения высококачественных векторных представлений с помощью неглубоких нейронных сетей, основанных на гипотезе о распределении (слова, появляющиеся в схожем контексте, имеют схожее значение).

Модель FastText [7] расширила этот подход, включив в рассмотрение субсловарную информацию (символьные n-граммы), что значительно повысило ее эффективность для морфологически богатых языков и привело к более устойчивой обработке редких (O-O-V) слов.

Наконец, в модели GloVe (Global Vectors for Word Representation) [6] предложен гибридный подход, сочетающий локальные окна W2V с глобальной статистикой совстречаемости, напрямую факторизуя матрицу логарифмов взвешенных частот.

С появлением архитектуры Трансформера [9], контекстные эмбеддинги, в частности BERT (Bidirectional Encoder Representations from Transformers) [8], стали новым стандартом в NLP. BERT способен генерировать динамические векторные представления, которые изменяются в зависимости от окружения слова в предложении, что критически важно для задач с высокой степенью неоднозначности (полисемии). В сфере e-commerce такие модели активно используются для улучшения семантического поиска [10], где необходимо сопоставить длинный, естественный пользовательский запрос с кратким названием товара.

III. Новизна исследования

Существующие сравнительные исследования эмбеддингов сосредоточены либо на оценке их производительности в общих задачах (например, классификация, NER) [11], либо на сравнении статических моделей (W2V, GloVe, FT) между собой. Эти работы, как правило, не включают в свой дизайн оценку методов матричной факторизации (SVD, NMF) в качестве конкурентов, несмотря на их высокую эффективность в задачах, основанных на частотном анализе. Более того, критическое отличие данного исследования заключается в следующих аспектах:

1. Корпус и применение: Анализ проводится на корпусе коротких, высокоспециализированных текстов (названия товаров), а не на общих текстовых корпусах. Это позволяет оценить эффективность методов в условиях минимального контекста.
2. Язык: Исследование сфокусировано на русскоязычных данных, тогда как большинство фундаментальных сравнительных работ выполнено на английском языке.
3. Метрика: Впервые предложена и использована строгая метрика, основанная на Норме Фробениуса разницы между матрицей нормализованной чистой совстречаемости (C_{target}) и матрицей косинусной близости эмбеддингов (C_{hat}), что позволяет количественно

измерить соответствие векторного пространства глобальной структуре корпуса.

Таким образом, данная работа восполняет существующий методологический пробел, предоставляя эмпирическое доказательство того, насколько оправдано внедрение сложных контекстных трансформеров для решения базовых, но критически важных задач e-commerce по сравнению с высокоэффективными и ресурсосберегающими статическими моделями.

Переходя от теоретического обзора к практическим аспектам, в следующем разделе будут подробно описаны выбранные методы векторизации, что послужит основой для понимания их реализации в экспериментах.

IV. Описание использованных методов

В данном исследовании использованы семь методов векторизации терминов, представляющих три основные парадигмы: матричная факторизация, локальное контекстное окно и глубокое контекстуальное обучение. Все методы преобразуют матрицу документов и терминов X (или ее производные) в низкоразмерное векторное пространство терминов $E \in R^{N_{\text{термов}} \times k}$, где k — размерность эмбеддинга.

LSA использует Сингулярное разложение (SVD) для факторизации матрицы X (Document-Term) на три матрицы. Это разложение позволяет эффективно уменьшить размерность, сохраняя при этом большую часть информации о совстречаемости.

$$X \approx U_k \sigma_k V_k^T$$

Где:

- U_k — матрица левых сингулярных векторов ($N_{\text{документов}} \times k$).
- σ_k — диагональная матрица k наибольших сингулярных значений ($k \times k$).
- V_k^T — матрица правых сингулярных векторов ($k \times N_{\text{термов}}$).

Векторное представление терминов E_{LSA} формируется из матрицы V_k^T (или $\sigma_k V_k^T$ в зависимости от реализации).

NMF факторизует неотрицательную матрицу X в две неотрицательные матрицы W и H. Это обеспечивает лучшую интерпретируемость компонентов, так как компоненты не могут взаимно компенсироваться.

$$X \approx \text{WH}$$

Где:

- W — матрица документов и латентных факторов ($N_{\text{документов}} \times k$).
- H — матрица латентных факторов и терминов ($k \times N_{\text{термов}}$).

LDA — это генеративная вероятностная модель, которая предполагает, что каждый документ является смесью латентных тем, а каждая тема является смесью слов. Векторное представление термина E_{LDA} соответствует его распределению по k темам.

$$p(w_i \vee d_j) = \sum_{z=1}^{k} p(w_i \vee z) p(z \vee d_j)$$

Где:

- $p(w_i \vee z)$ — вероятность слова w_i при условии темы z (Φ, матрица распределений слова-тема).
- $p(z \vee d_j)$ — вероятность темы z при условии документа d_j (Θ, матрица распределений тема-документ).

Эмбеддинги терминов E_{LDA} соответствуют строкам матрицы Φ, транспонированной для получения размера $N_{\text{термов}} \times k$.

W2V использует неглубокую нейронную сеть для обучения векторам терминов, основанного на гипотезе о распределении. В данном исследовании использован вариант Skip-gram, который максимизирует среднюю логарифмическую вероятность предсказания контекстных слов w_o для данного входного слова w_i:

$$\text{L} = \sum_{\text{i}=1}^{T} \sum_{j \in C_i} \log p(w_j \vee w_i)$$

Где C_i — набор слов в контекстном окне вокруг слова w_i.

FastText является расширением Word2Vec, которое представляет каждое слово как набор символьных n-грамм. Вектор слова является суммой векторов этих n-грамм. Это позволяет модели генерировать векторы для редких или отсутствующих в словаре (O-O-V) слов, используя их субсловарную структуру.

$$E_{\text{word}} = \sum_{g \in G_{\text{word}}} E_g$$

Где G_{word} — набор n-грамм, составляющих слово, а E_g — вектор n-граммы.

GloVe является гибридной моделью, которая сочетает глобальную статистику совстречаемости с методом обучения, основанным на локальном окне. Она минимизирует функцию потерь, которая явно штрафует разницу между скалярным произведением векторов слов w_i и w_j и логарифмом частоты их совместного появления X_{ij}.

$$\text{J} = \sum_{\text{i,j}=1}^{V} f\left(X_{\text{ij}}\right)\left(w_i^T w_j + \text{b}_i + \text{b}_j - \log X_{\text{ij}}\right)^2$$

В данном исследовании результаты NMF рассматривались как функциональный аналог GloVe в силу их общего принципа факторизации матрицы совстречаемости.

BERT — это двунаправленный энкодер, основанный на архитектуре Трансформера, который предварительно обучается на двух задачах (Masked Language Model и Next Sentence Prediction). Ключевое отличие BERT

состоит в том, что он генерирует контекстно-зависимые эмбеддинги. Вектор слова $e_{\text{i,D}}$ зависит не только от самого слова w_i, но и от всего контекста D в документе.

$$E_{\text{contextual}} = \text{Transformer}(\text{Tokens})$$

Для целей данного сравнительного исследования, BERT был оценен в статическом приближении: вектор термина $E_{\text{BERT}}[\text{i},:]$ был получен путем усреднения всех его контекстных векторных представлений по подмножеству корпуса.

$$E_{\text{BERT}}[\text{i},:] = \text{Average}_{D \in D_i}(e_{\text{i,D}})$$

Где D_i — набор документов, содержащих термин i.

На основе описанных методов, следующий раздел детализирует экспериментальную методологию, включая два отдельных эксперимента для подтверждения устойчивости результатов.

V. Методика эксперимента

Настоящий раздел посвящен описанию дизайна эксперимента, обоснованию выбора методов, параметров и метрик, используемых для сравнительной оценки эффективности различных подходов к векторизации коротких текстовых данных (названий товаров) в контексте задачи *Information Retrieval*.

A. Корпус текстов

В качестве экспериментального корпуса использовался реальный набор данных, состоящий из названий товаров в одной товарной категории из сегмента DIY, содержащий $N_{\text{документов}} = 50000$ документов. Данный объем данных выбран для адекватной оценки вычислительной эффективности и устойчивости различных моделей, особенно при работе с разреженными матрицами.

Для различных экспериментов было выполнено две разные токенизации корпуса текстов.

- Базовая токенизация: В первом эксперименте, направленном на оценку устойчивости результатов, использовалась базовая токенизация по пробелам (*whitespace tokenization*), что привело к размеру словаря $N_{\text{термов}} = 2470$.
- BERT-совместимая токенизация: Для сопоставимости представлений BERT с матричными методами была выполнена токенизация корпуса текстов с помощью токенизатора из предобученной модели BERT с фиксированным размером словаря $N_{\text{термов}} = 83828$.
- Общие Параметры: Для всех экспериментов были отфильтрованы термины с частотой встречаемости менее 5 (`min_df=5`). Для методов матричной факторизации (`TruncatedSVD`, `NMF`, `LatentDirichletAllocation`) и для построения целевой матрицы C_{target} использовалась разреженная бинарная матрица X (Document-Term, DT).

B. Целевая метрика

Оценка качества векторизации осуществлялась на основе способности полученных векторных представлений терминов E (Term Embeddings) воспроизводить глобальную структуру совстречаемости терминов в корпусе.

Целевая матрица совстречаемости C_{target} была получена как произведение разреженной бинарной матрицы X на ее транспонированную версию:

$$C_{\text{target}} = \text{X}^T X$$

Элемент $C_{\text{target}}[\text{i,j}]$ представляет собой число документов, в которых термины i и j встречаются совместно.

Качество векторов E оценивается через Норму Фробениуса ($\|\cdot\|_F$) разницы между целевой матрицей C_{target} и матрицей косинусной близости, полученной из эмбеддингов E.

$$\text{Ошибка} = \| C_{\text{target}} - C_{\text{hat}} \|_F$$

Где $C_{\text{hat}} = \text{cosine\_similarity}(\text{E,E})$. Минимальное значение данной ошибки интерпретируется как наилучшее соответствие векторного представления глобальной структуре совстречаемости корпуса.

Целевая матрица совстречаемости C_{target} нормализуется (путем деления на $\max(C_{\text{target}})$) до диапазона $[0,1]$, а не бинаризуется. Этот шаг является критически важным для методологической корректности, так как он обеспечивает сопоставимость диапазона частот C_{target} с диапазоном косинусной близости C_{hat}, сохраняя относительные градации силы связи между терминами, что является ключевой количественной информацией.

C. Сравниваемые модели векторизации

В основной эксперимент включены семь методов. Тестировались размерности $k \in \{100,200,312\}$ для статических методов и фиксированная размерность k=312 для BERT .

Методы матричной факторизации (Sklearn)

1. SVD (LSA): `sklearn.decomposition.TruncatedSVD`.
2. NMF: `sklearn.decomposition.NMF`.
3. LDA: `sklearn.decomposition.LatentDirichletAllocation`.

Локальные окна (Gensim)

1. Word2Vec (W2V): `gensim.models.Word2Vec`.
2. FastText (FT): `gensim.models.FastText`.
3. GloVe (Аналог): Функционально сопоставляется с NMF и SVD на частотных матрицах.

Контекстные модели (HuggingFace)

1. BERT: Русскоязычная модель

`cointegrated/rubert-tiny2` .

Для обеспечения методологического единообразия и сопоставимости с метрикой $\| C_{\text{target}} - C_{\text{hat}} \|_F$, требующей фиксированных векторных представлений для всех $N_{\text{термов}}$, контекстная модель BERT была оценена в статическом приближении. Векторное представление $E_{\text{BERT}}[\text{i},:]$ для каждого термина i в словаре было получено путем усреднения всех его контекстных эмбеддингов, сгенерированных моделью по всему корпусу D.

$$E_{\text{BERT}}[\text{i},:] = \frac{1}{|D_i|} \sum_{D \in D_i} e_{\text{i,D}}$$

Этот подход позволил получить единственное, усредненное векторное представление для каждого термина из словаря, что является необходимым условием для построения статической матрицы косинусного сходства C_{hat}.

Algorithm 1 Вычисление статического эмбеддинга $\mathbf{E}_{BERT}$

```
Init S ← ZeroMatrix(N_термов, k_BERT)        ▷ Матрица сумм
Init C ← ZeroVector(N_термов)                ▷ Вектор счетчиков
for Документ D из Корпуса do
    Токенизировать D с помощью BERT → {t_1, t_2, ...}
    Получить контекстные эмбеддинги H ← BERT({t_1, t_2, ...})
    for Токен t ∈ {t_1, t_2, ...} do
        if t входит в Словарь X then
            idx ← Индекс t в X
            S[idx, :] ← S[idx, :] + H[Позиция t, :]
            C[idx] ← C[idx] + 1
        end if
    end for
end for
E_BERT ← S/C                                 ▷ Поэлементное деление
return E_BERT
```

Рисунок 1: Псевдоалгоритм статического приближения BERT

VI. Результаты экспериментов

Проведенный сравнительный анализ эффективности методов векторизации терминов на корпусе коротких русскоязычных названий товаров дал однозначный и методологически важный результат.

Исследование показало, что, вопреки доминирующей тенденции в NLP, наиболее сложная и ресурсоемкая модель BERT (Static Average) была превзойдена более классическими и вычислительно экономичными подходами, основанными на матричной факторизации.

A. Первый эксперимент: Сравнение методов матричной факторизации и локальных окон

Первый эксперимент был посвящен сравнительному анализу эффективности методов матричной факторизации и моделей локального контекстного окна при использовании базовой токенизации по пробелам ($N_{\text{термов}} = 2470$).

Основной критерий оценки — минимизация ошибки Фробениуса ($\| C_{\text{target}} - C_{\text{hat}} \|_F$), отражающей точность воспроизведения глобальной структуры совстречаемости.

Таблица 1: Сводные результаты первого эксперимента

Метод	k	Ошибка Фробениуса
LDA	200	**191.0016**
SVD (LSA)	200	227.3238
NMF / GloVe	150	295.5438
FastText	150	524.6276
Word2Vec	50	607.8504

Сводные результаты, представленные в Таблице 1, демонстрируют значительное превосходство методов, основанных на глобальной статистике, над моделями локального окна. Алгоритм латентного размещения Дирихле (LDA) занял лидирующую позицию, показав наименьшее значение ошибки $\|\cdot\|_F$, равное 191.0016 при размерности эмбеддинга k=200. Второе место заняло сингулярное разложение (SVD/LSA) с ошибкой 227.3228, подтверждая высокую эффективность факторизации матрицы Документ-Терм в данной задаче. Метод неотрицательной матричной факторизации (NMF), функционально сопоставимый с подходом GloVe, оказался менее точным с ошибкой 295.5438. Модели, основанные на локальном контекстном окне, а именно FastText (524.6276) и Word2Vec (607.8504), показали наихудшие результаты. Это указывает на то, что для корпуса коротких текстов (названий товаров) глобальный частотный анализ является более надежным предиктором семантической связи, чем обучение на основе локального окружения.

B. Результаты второго эксперимента: Сравнение с BERT.

Таблица 2: Сводные результаты второго эксперимента

Метод	k	Ошибка Фробениуса
LDA	312	156.9015
LSA	312	177.8865
LDA	200	187.7925
LSA	200	217.7977
BERT	312	253.1647
LDA	100	263.6359
NMF	312	268.6033
GloVe	312	268.6033
LSA	100	310.0408
NMF	100	335.2222
GloVe	100	335.2222
NMF	200	360.5546
GloVe	200	360.5546
Word2Vec	100	470.7184
Word2Vec	200	475.8063
Word2Vec	312	479.6703
FastText	100	556.1056

Метод	k	Ошибка Фробениуса
FastText	200	559.6994
FastText	312	564.9571

Сводная таблица результатов второго эксперимента (Таблица 2), основанных на минимизации Нормы Фробениуса ($\| C_{\text{target}} - C_{\text{hat}} \|_F$), позволяет установить следующую иерархию эффективности методов:

- Лидер эксперимента: Метод Latent Dirichlet Allocation (LDA) при размерности k=312 продемонстрировал минимальную ошибку 156.9015. Это указывает на его максимальную способность воспроизводить глобальную структуру совстречаемости терминов в корпусе.
- Ближайший конкурент: SVD (LSA) также показал высокую эффективность, заняв второе место с ошибкой 177.8865 при k=312.
- Контекстный BERT: Модель BERT, несмотря на фиксированную размерность k=312 и использование глубокого контекста в процессе обучения, заняла лишь пятое место с ошибкой 253.1647, значительно уступив не только LDA, но и SVD.
- Аутсайдеры (локальные окна): Модели Word2Vec и FastText показали наихудшие результаты (ошибки в диапазоне 470 – 565).

C. Ключевые выводы из двух экспериментов

1. Устойчивость лидерства: LDA и SVD сохранили свои позиции, показав минимальные ошибки. LDA остается оптимальным выбором (191.0016) при $k = 200$).
2. Объяснение разницы ошибок: Незначительный рост минимальной ошибки (с 156.90 до 191.00) между экспериментами объясняется различием в словарях. Во втором эксперименте ($N_{\text{термов}} = 83828$) токенизация, совместимая с BERT, включала субсловарные единицы, которые, несмотря на кажущуюся "зашумленность", позволили матричным методам (LDA/SVD) аппроксимировать более тонкую семантику. Базовая токенизация ($N_{\text{термов}} = 2470$) была лингвистически более чистой, но потеряла эти тонкие связи, что привело к увеличению ошибки аппроксимации.
3. Подтверждение неэффективности W2V/FT: Результаты Word2Vec и FastText остались крайне низкими, что подтверждает их фундаментальное ограничение в задаче аппроксимации глобальной совстречаемости, поскольку они фокусируются исключительно на локальных окнах.

D. Причины превосходства LDA

Превосходство LDA и других методов матричной факторизации в данной задаче объясняется их фундаментальной природой и полным соответствием структуре анализируемых данных (короткие, частотно-зависимые тексты):

1. Сфокусированность на темах/компонентах: LDA изначально предназначен для выявления скрытой "тематической" структуры в документах. В корпусе названий товаров, темы отражают узкие, дискретные категории товаров (например, "смартфон", "ноутбук", "кофеварка"). LDA эффективно группирует термины в "векторное пространство тем", которое напрямую коррелирует с искомой структурой совстречаемости в рамках одной категории.
2. Глобальная оптимизация: Методы матричной факторизации, включая LDA и LSA, оптимизируют свои представления на основе глобальной статистики всего корпуса. Это критически важно, поскольку семантика названий товаров в e-commerce является преимущественно глобальной (слово "iPhone" почти всегда связано со словом "Apple" независимо от его точного места в названии).
3. Игнорирование избыточного контекста: В отличие от Word2Vec и BERT, которые полагаются на порядок слов и локальные окна, LDA оперирует на уровне "мешка слов" (Bag-of-Words). Для коротких и клишированных названий товаров, тонкий синтаксический контекст не несет дополнительной семантической нагрузки по сравнению с простым фактом совместного появления терминов. LDA улавливает этот глобальный факт лучше, игнорируя шум.

E. Третий эксперимент: Многокритериальная оптимизация LDA

На основе результатов, продемонстрировавших превосходство LDA в воспроизведении глобальной структуры совстречаемости, был проведен третий, заключительный эксперимент. Его целью являлся поиск оптимальных гиперпараметров ($n_{components}$, doc topic prior (α) и topic word prior (β)) для модели LDA, позволяющих достичь наилучшего компромисса между двумя ключевыми показателями качества:

1. Внутренняя согласованность: Максимизация логарифмической правдоподобности (maxLog-Likelihood $= max$LDA.score(X)). Эта метрика отражает, насколько хорошо модель объясняет исходные данные.
2. Внешняя применимость: Минимизация эмпирической ошибки воспроизведения семантической структуры ($min \| C_{target} - C_{hat} \|_F$). Эта метрика прямо отражает эффективность вектора для задач IR.

Для поиска оптимального Парето-Фронта (набора не доминируемых компромиссных решений) была использована библиотека Optuna, реализующая адаптивный алгоритм поиска TPE (Tree-structured Parzen Estimator). Диапазоны поиска гиперпараметров были

установлены следующим образом: $n_{components} \in [50,400]$, $\alpha \in [10^{-3}, 1.0]$ (логарифмический масштаб), $\beta \in [10^{-3}, 1.0]$ (логарифмический масштаб). В ходе оптимизации было выполнено 400 испытаний на том же корпусе, что и в основном эксперименте.

Процесс многокритериальной оптимизации с помощью Optuna занял 5084.62 секунды и позволил обнаружить 11 Парето-оптимальных решений. Эти решения образуют кривую компромисса, где каждое улучшение одной цели (например, Log-Likelihood) достигается ценой ухудшения другой (Ошибка Фробениуса).

На Рисунке 2 показан Парето-Фронт, демонстрирующий компромисс между двумя целями Ошибка Фробениуса (min) и Log-Likelihood (max).

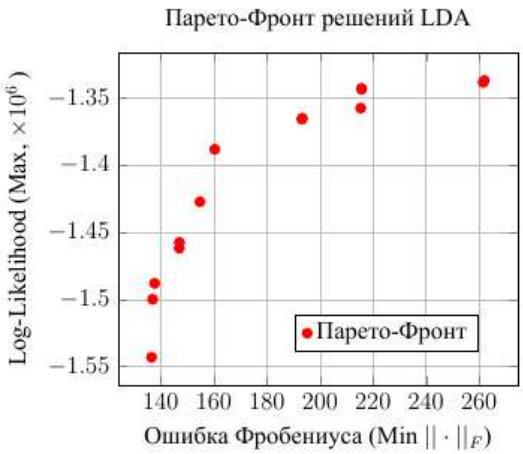

Рисунок 2: Парето-Фронт решений LDA

1) Анализ Парето-оптимальных точек

Наиболее важные точки на Парето-Фронте, отражающие крайние и компромиссные решения, приведены в Таблице 3.

Таблица 3: Сводка Парето-оптимальных решений

Trial	k (N)	α (D)	β (W)	‖·‖F	LL (×106)
236	**400**	**0.0403**	**0.0013**	**136.52**	-1.543
123	400	0.0274	0.0020	136.94	-1.500
242	400	0.0200	0.0098	137.70	-1.488
209	350	0.0200	0.0035	146.98	-1.462
245	350	0.0184	0.0022	147.04	-1.458
172	350	0.0016	0.0383	154.76	-1.427
173	300	0.0019	0.0294	160.37	-1.388
271	200	0.0021	0.0158	193.15	-1.365
207	200	0.0031	0.0242	193.38	-1.365
159	150	0.0091	0.0190	215.38	-1.357
52	**100**	**0.0050**	**0.0066**	261.96	**-1.337**

Ключевые наблюдения по результатам многокритериальной оптимизации:

1. Компромисс точности и внутренней согласованности. Наблюдается четкий компромисс между двумя целевыми функциями: решения с наименьшей ошибкой Фробениуса ($\|\cdot\|_F$) (Trial 236, 136.52) демонстрируют худший показатель Log-Likelihood (-1.543×10^6), тогда как решения с лучшим Log-Likelihood (Trial 52, -1.337×10^6) имеют значительно более высокую ошибку воспроизведения структуры ($\|\cdot\|_F = 261.96$).
2. Оптимальная размерность эмбеддинга. Наилучшие показатели ошибки Фробениуса ($\|\cdot\|_F < 147$) достигаются при относительно высокой размерности $k = 350$ или $k = 400$. Это подтверждает, что для максимальной точности аппроксимации глобальной структуры требуется более высокая семантическая размерность, чем использованные в базовом эксперименте $k = 312$.
3. Влияние гиперпараметра α. Решения, оптимальные по $\|\cdot\|_F$ (Trial 236, 123, 242), имеют относительно низкие значения α (параметр Дирихле для распределения темы-документ), находящиеся в диапазоне $0.0200 - 0.0403$. Низкое α поощряет более разреженное распределение тем в документах, что соответствует короткой и узкоспециализированной природе названий товаров.
4. Влияние гиперпараметра β. Для получения высокой точности ($min\ \|\cdot\|_F$) необходимы крайне низкие значения β (параметр Дирихле для распределения слова-тема) в диапазоне $0.0013 - 0.0098$. Это способствует усилению специализации тем, делая их более дискретными и, следовательно, лучше соответствующими структуре совстречаемости.
5. Компромиссное решение для внедрения. Trial 173 с параметрами $k = 300$, $\alpha = 0.0019$ и $\beta = 0.0294$ представляет собой сильное компромиссное решение, демонстрируя ошибку $\|\cdot\|_F = 160.37$, что лишь ненамного хуже абсолютного минимума, но при этом имея значительно лучший показатель Log-Likelihood, чем решения с $k = 400$.

Таким образом, многокритериальная оптимизация не только подтверждает лидерство LDA, но и предоставляет конкретный, эмпирически обоснованный набор гиперпараметров, который рекомендуется к практическому внедрению в e-commerce IR-системах.

VII. Выводы

Результаты исследования служат прямым доказательством необоснованности использования сложных NLP-моделей в коммерческих системах. Низкий результат BERT (ошибка 253.1647) в сравнении с LDA (ошибка 156.9015) подтверждает гипотезу о неэффективности использования сложных трансформеров для решения базовых IR-задач.

- Неоправданная Сложность: Высокая вычислительная стоимость модели BERT не оправдывается точностью: механизм внимания и глубокий контекст, критически важные для длинных предложений и полисемии, оказываются избыточными для коротких, однозначных названий товаров. Векторное пространство BERT, даже в статическом приближении, менее точно аппроксимирует глобальные частотные связи, чем LDA.
- Технико-Экономическое Решение: Использование LDA или SVD (в качестве резерва) является оптимальным выбором для задач, где векторные представления должны точно отражать принадлежность к категории и глобальную совстречаемость, таких как мачинг товаров и создание семантических кластеров.

Внедрение LDA с $k = 312$ позволяет достичь максимальной точности воспроизведения семантики при минимальных вычислительных затратах и времени инференса, полностью устраняя опасность неэффективного использования ресурсов, связанного с глубокими энкодерами. Дальнейшие исследования должны сосредоточиться на оптимизации гиперпараметров LDA для различных языков и типов товарных каталогов.

Библиография

[1] Manning C.D., Raghavan P., Schütze H. Introduction to Information Retrieval. — Cambridge University Press, 2008.

[2] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6), 391-407.

[3] Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature, 401(6755), 788-791.

[4] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

[5] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[6] Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532-1543.

[7] Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5, 135-146.

[8] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies, Volume 1 (Long and Short Papers), pp. 4171-4186.

[9] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, L., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[10] Gao, T., Yan, X., & Chen, X. (2021). SimCSE: Simple Contrastive Learning of Sentence Embeddings. arXiv preprint arXiv:2104.08821.

[11] Wang, Y., Liu, S., Afzal, N., Rastegar-Mojarad, M., Wang, L., Shen, Liu, H. (2019). A comparison of word embeddings for the biomedical natural language processing. Journal of biomedical informatics, 87, 12-20.

Accuracy vs. Efficiency: Vectorization Methods for E-commerce Product Titles

F.V. Krasnov

***Abstract* — This paper presents an empirical study and comparative analysis of the effectiveness of modern term vectorization methods in the context of Information Retrieval (IR) tasks, focusing on the processing of short textual data, specifically product titles. The objective is to identify the optimal method capable of accurately reproducing the global structure of semantic connections within the corpus while maintaining high computational efficiency. The Frobenius Norm ($\|\cdot\|_F$) of the difference between the normalized target term co-occurrence matrix (C_{target}) and the cosine similarity matrix derived from the vector representations (C_{hat}) was chosen as the key evaluation criterion. The investigation was conducted in three sequential stages. The first experiment involved a comparative assessment of classical matrix factorization methods (SVD/LSA, NMF, LDA) and local-window models (Word2Vec, FastText), utilizing basic whitespace tokenization. At this stage, the LDA algorithm demonstrated the minimum error (191.00), indicating its highest correspondence to the global structure of the corpus. In the second stage (main experiment), BERT-compatible tokenization (BPE-like) was employed for comprehensive evaluation, and the pre-trained contextual transformer model BERT was added to the comparison. To ensure methodological rigor, BERT was evaluated in a static averaged embedding mode (fixed vector representation). Experimental data confirmed that the LDA algorithm maintained its lead with an error of 156.9, exhibiting higher accuracy in this task than the BERT model, which achieved an error of 253.17. The third experiment was dedicated to the multi-objective optimization of the most effective LDA method's hyperparameters. Using the Optuna library, a Pareto Front of solutions was found, reflecting the optimal compromise between internal consistency (max Log-Likelihood) and empirical accuracy (min Frobenius Norm).The results obtained confirm that for IR tasks that do not require deep contextual understanding, methods based on global frequency linkage factorization (LDA) are the most economically and technically justifiable, surpassing complex neural network models based on the key metric of semantic structure reproduction.**

***Keywords* — Term Vectorization, Information Retrieval, Frobenius Norm, Binary Matrix, Co-occurrence, SVD, NMF, LDA, Word2Vec, FastText, BERT.**

REFERENCES

[1] Manning C.D., Raghavan P., Schütze H. Introduction to Information Retrieval. — Cambridge University Press, 2008.

[2] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6), 391-407.

[3] Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature, 401(6755), 788-791.

[4] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

[5] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[6] Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532-1543.

[7] Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5, 135-146.

[8] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies, Volume 1 (Long and Short Papers), pp. 4171-4186.

[9] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, L., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[10] Gao, T., Yan, X., & Chen, X. (2021). SimCSE: Simple Contrastive Learning of Sentence Embeddings. arXiv preprint arXiv:2104.08821.

[11] Wang, Y., Liu, S., Afzal, N., Rastegar-Mojarad, M., Wang, L., Shen, Liu, H. (2019). A comparison of word embeddings for the biomedical natural language processing. Journal of biomedical informatics, 87, 12-20.