К

Проблемы искусственного интеллекта. 2026. N 3 (42). С. 24-35
Problems of Artificial Intelligence. 2026;3(42):24-35.
Искусственный интеллект и машинное обучение
Научная статья

УДК 004.912 **DOI** 10.24412/2413-7383-2026-3-42-24-35

И. А. Коршакова
Общество с ограниченной ответственностью «Газпром информ»
Площадь Победы, дом 2, литера А, помещение 1-Н, кабинет 422,
Санкт-Петербург, 196143, Россия
korshakova40@gmail.com; ORCID: 0009-0003-3591-1509

АВТОМАТИЗАЦИЯ ВЫЯВЛЕНИЯ ДУБЛИКАТОВ РАЦИОНАЛИЗАТОРСКИХ ПРЕДЛОЖЕНИЙ ЗА СЧЕТ АНАЛИЗА ТЕКСТОВОЙ СХОЖЕСТИ

I. A. Korshakova
Gazprom Inform Limited Liability Company
2 Pobedy Square, Litera A, Premises 1-N, Office 422, Saint Petersburg, 196143, Russia
korshakova40@gmail.com; ORCID: 0009-0003-3591-1509

AUTOMATING THE DETECTION OF DUPLICATE RATIONALIZATION PROPOSALS USING TEXT SIMILARITY ANALYSIS

Рассматривается задача поиска семантически схожих и дублирующих рационализаторских предложений в исполняемом бизнес-процессе «Управление рационализаторской деятельностью». Для оценки смысловой близости исследовались модели преобразования текста в числовые представления (embedding-модели). Эффективность моделей оценивалась на основе степени сходства текстов и результатов экспертной оценки. Для повышения точности поиска дубликатов выполнена оптимизация весов объектов предложений с помощью эволюционного алгоритма. На основе метрик бенчмарка ruMTEB и экспертной оценки выявлена наиболее эффективная модель.
Ключевые слова: рационализаторские предложения, семантическая близость текстов, векторное представление текста, поиск дубликатов, эволюционный алгоритм.

This paper addresses the problem of identifying semantically similar and duplicate rationalization proposals within the “Rationalization Activity Management” executable business process. To assess semantic similarity, models that transform textual data into numerical representations were investigated. The performance of the models was evaluated using cosine similarity and expert assessment results. To improve the accuracy of duplicate detection, the weights assigned to different textual attributes of the proposals were optimized using an evolutionary algorithm. As a result, based on the ruMTEB benchmark metrics and expert assessment, the most effective model for identifying duplicate rationalization proposals was determined.
Key words: rationalization proposals, semantic text similarity, text vector representation, duplicate detection, evolutionary algorithm.

Введение

Актуальность исследования обусловлена необходимостью повышения эффективности управления рационализаторской деятельностью в дочерних обществах предприятия «ТЭК». В ходе эксплуатации информационной системы была выявлена проблема дублирования предложений. Отсутствие механизма автоматического поиска семантически схожих инициатив приводит к повторному рассмотрению уже существующих решений, увеличению трудозатрат экспертов и снижению эффективности тиражирования успешных практик [1]. Поскольку предложения могут быть сформулированы различными способами при сохранении одинакового содержания, применение традиционных методов текстового поиска оказывается недостаточно эффективным [2]. Для решения подобных задач применяются методы семантического анализа текста [3], [4], в том числе современные архитектуры представления смысла [5]. В связи с этим актуальной является задача использования современных методов обработки естественного языка для автоматического выявления семантически близких предложений [6].

Целью исследования является разработка механизма поиска семантически схожих рационализаторских предложений на основе embedding-моделей для последующего выявления дубликатов и рекомендаций повторного использования ранее внедрённых решений.

Для достижения поставленной цели в работе решаются следующие **задачи**:

1) подготовка обучающего набора данных из рационализаторских предложений и выбор embedding-моделей;
2) построение векторных представлений текстов с использованием embedding-моделей;
3) сравнение семантической близости текстов на основе меры косинусного сходства;
4) проведение экспертной оценки результатов работы моделей и выбор наилучшей;
5) оценивание экономической эффективности внедрения программного модуля на внешних тестах, реализующего разработанные алгоритмы повышения качества классификации пар предложений на дубликаты и недубликаты.

Описание предметной области

Рационализаторские предложения представляют собой инициативы работников, направленные на совершенствование процессов предприятия. В информационной системе управления рационализаторской деятельностью накапливается множество предложений, среди которых встречаются дублирующие и семантически схожие инициативы. Это приводит к повторному рассмотрению решений, увеличению трудозатрат экспертов и снижению эффективности использования ранее внедрённых практик.

Особенностью рассматриваемой задачи является то, что предложения с одинаковым смыслом могут существенно различаться по формулировкам, используемой терминологии и структуре текста. Вследствие этого методы поиска, основанные на лексическом совпадении слов и выражений, не обеспечивают требуемого качества выявления семантически близких инициатив. Для решения данной задачи целесообразно применение современных методов обработки естественного языка, позволяющих учитывать смысловое содержание текста и определять степень его семантической близости [7–9].

Алгоритм работы модуля

Предлагаемый механизм автоматического поиска семантически схожих рационализаторских предложений представляет собой последовательный алгоритм обработки текстовых данных, предназначенный для выявления предложений с близким

смысловым содержанием независимо от особенностей их формулировки. Основными этапами работы механизма являются подготовка текстовых данных, формирование их векторных представлений, вычисление семантической близости и классификация сравниваемых предложений (в соответствии с рисунком 1, алгоритм 1).

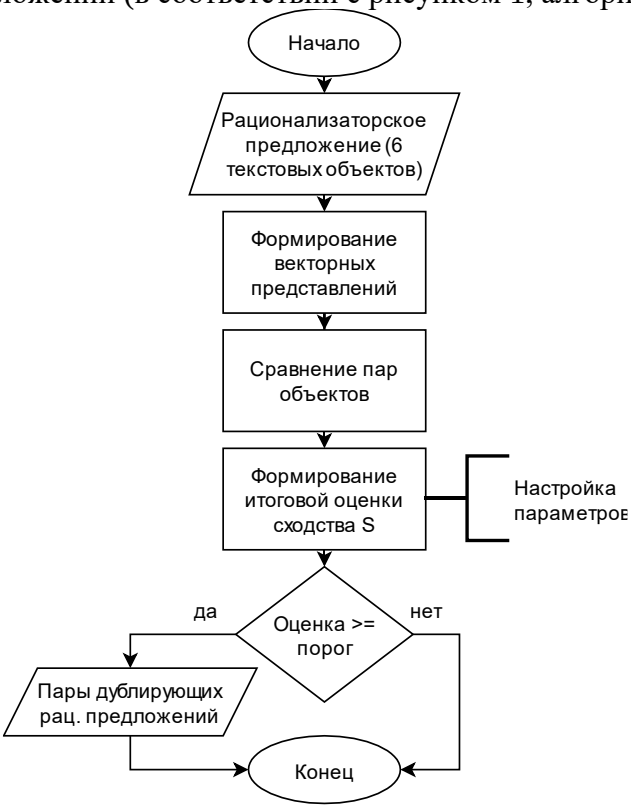


Рисунок 1 – Блок-схема алгоритма выявления дублирующих рационализаторских предложений

Алгоритм 1 – Выявление дублирующих рационализаторских предложений

1. Вход: рационализаторское предложение P, представленное шестью текстовыми объектами; пороговое значение сходства θ.
2. Для каждого текстового объекта предложения P с помощью embedding-модели сформировать векторное представление, отражающее его смысловое содержание.
3. Для каждой пары соответствующих текстовых объектов предложений P и P' вычислить меру семантической близости (косинусное сходство векторных представлений).
4. Сформировать итоговую оценку сходства $S(P, P')$ как взвешенную сумму значений близости отдельных текстовых объектов с учётом их значимости.
5. Если $S(P, P') \geq \theta$, пометить пару (P, P') как потенциальный дубликат и уведомить эксперта о возможном совпадении с ранее рассмотренным предложением или уже внедрённым решением; иначе обработать P в штатном порядке.
6. Выход: пары дублирующих предложений.

Предложенный механизм позволяет перейти от поиска предложений по лексическому совпадению к сравнению их **семантического содержания**, что обеспечивает выявление потенциальных дубликатов даже при существенных различиях в формулировках, терминологии и структуре текста.

Выбор embedding-моделей

Для выбора embedding-моделей использовался русскоязычный тестовый набор данных (бенчмарк) Russian Massive Text Embedding Benchmark (ruMTEB), предназначенный для комплексной оценки качества моделей представления текста в задачах семантического поиска, кластеризации и ранжирования [10].

На основе результатов бенчмарка ruMTEB и требований задачи были выбраны три embedding-модели, показавшие высокую эффективность в задачах семантического поиска и анализа текстов на русском языке. Результаты сравнительного анализа по метрикам бенчмарка представлены в соответствии с таблицей 1.

Таблица 1 – Оценка embedding-моделей

Модель	**Средняя оценка**	**Семантический поиск**	**Схожесть смысла**	**Сортировка результатов**
BAAI/bge-m3	62.8	68.7	76.8	65.1
deepvk/USERv2-base	61.1	66.9	74.3	56.4
GigaEmbeddings	69.1	75+	80+	70+

По результатам бенчмарка ruMTEB для сравнения были отобраны следующие embedding-модели:

1) **BAAI/bge-m3** [11] — китайская многоязычная модель (100+ языков, включая русский), размерность d = 1024, контекст до 8192 токенов; средняя оценка — 62.8;
2) **deepvk/USERv2-base** [12], [13] — русскоязычная модель компании ВК, d = 768, контекст до 512 токенов; средняя оценка — 61.1;
3) **GigaEmbeddings** [14] — русскоязычная модель компании Сбер, d = 1024, контекст до 4096 токенов; средняя оценка — 69.1, семантический поиск — 75+, схожесть смысла — 80+, сортировка результатов — 70+.

Ввиду наилучших показателей по всем метрикам ruMTEB модель GigaEmbeddings рассматривается как приоритетный кандидат для решения задачи автоматического выявления дубликатов рационализаторских предложений.

Для измерения семантического сходства между парами рационализаторских предложений применяется метрика косинусной близости между двумя векторами a и b (1):

$$cos(a,b) = \frac{(a \cdot b)}{((\|a\|_2 \cdot \|b\|_2)}, \quad (1)$$

где $a \cdot b = \Sigma_i a_i b_i$ — скалярное произведение векторов, а $\|a\|_2 = \sqrt{(\Sigma_i {a_i}^2)}$ и $\|b\|_2 = \sqrt{(\Sigma_i {b_i}^2)}$ — евклидовы нормы вектора.

Данная мера оценивает угол между двумя векторами в многомерном пространстве и не зависит от их абсолютных длин, что делает её особенно подходящей для сравнения текстовых векторных представлений [15], [16]. Подобный подход широко используется в задачах семантического поиска и ранжирования на основе векторных представлений [17], [18].

Значение косинусной близости находится в диапазоне [−1, 1]. Значение, близкое к 1, свидетельствует о высокой семантической схожести текстов, а значение, близкое к 0, — об отсутствии смысловой связи. Отрицательные значения указывают на противоположность смыслов. В контексте данной задачи пара рационализаторских предложений считается дублирующей при $cos(a,b) \geq \theta$, где θ — пороговое значение.

Сравнительный анализ результатов работы embedding-моделей

Для оценки качества embedding-моделей проведено исследование на реальных данных информационной системы управления рационализаторской деятельностью. В качестве выборки использовалось 221 рационализаторское предложение, каждое из которых содержит шесть текстовых объектов. Всего обработано 1326 текстовых объектов.

В качестве критерия выявления семантически схожих текстов установлен порог косинусной близости 0.9. При превышении данного значения пара объектов считалась потенциально дублирующей. Сравнение проводилось отдельно для каждого типа объекта рационализаторского предложения, что позволило оценить эффективность моделей на различных видах текстовых данных.

Для детального анализа различий в результатах работы моделей рассмотрим объект «Обоснование эффекта». Для данного объекта была построена диаграмма Венна (в соответствии с рисунком 2), отражающая пересечение найденных семантически схожих пар текстов между моделями BGE-M3, USERv2 и Giga-Embeddings.

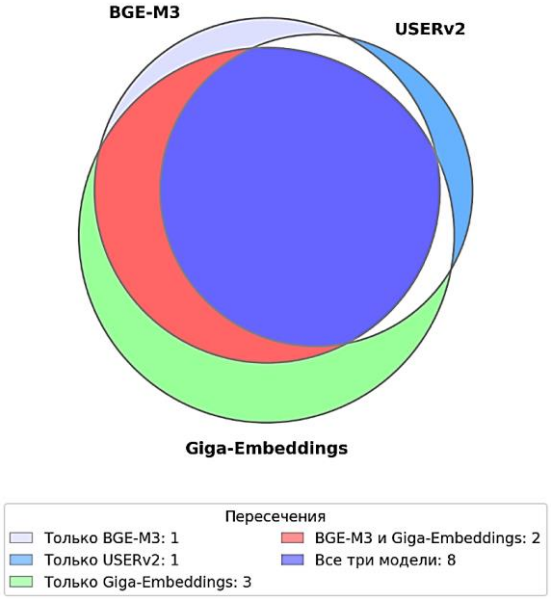


Рисунок 2 – Диаграмма Венна для объекта «Обоснование эффекта»

Согласно диаграмме Венна, восемь пар семантически схожих текстов были выявлены всеми рассматриваемыми моделями. Дополнительно одна уникальная пара была обнаружена только моделью BGE-M3, одна — только моделью USERv2 и три — только моделью Giga-Embeddings. Кроме того, две пары были совместно выявлены моделями BGE-M3 и Giga-Embeddings.

Анализ результатов показывает, что во всех исследуемых объектах наибольшее количество уникальных семантически схожих пар было выявлено моделью Giga-Embeddings. Большинство совпадений, обнаруженных другими моделями, также входят в её результаты. Наибольшее преимущество наблюдается для объекта «Недостатки существующего положения», где модель выявила шестнадцать уникальных пар, не найденных другими моделями. Это подтверждает более высокую чувствительность Giga-Embeddings к выявлению смысловой близости текстов.

Для качественной оценки результатов модели Giga-Embeddings представлены примеры текстов, демонстрирующих семантическое сходство по каждому из шести объектов рационализаторского предложения.

Для объекта **«Наименование предложения»** сопоставлены тексты: «Способ отключения наружного освещения цеха №» и «Система отключения наружного освещения КЦ №».

Для объекта **«Цель предложения»** выявлено сходство между текстами: «Повысить стабильность работы оборудования» и «повышение надежности и стабильности оборудования».

Для объекта **«Сущность предложения»** рассмотрены два семантически схожих текста. Первый текст: «В ЛПУМГ «х» для записи телефонных звонков используется специализированный АРМ EarDog 8 (занято 4 из 8 каналов). АРМ WDS Tetra для проигрывания звука использует колонки с AUX выходом. Предлагаем изготовить кабель согласно Приложения № 1».

Второй текст: «В ЛПУМГ «х» для записи телефонных звонков используется отдельный АРМ EarDog 8 (занято 4 из 8 каналов). АРМ WDS Tetra для проигрывания звука использует колонки с AUX выходом. Предлагаем изготовить кабель согласно Приложения № 1».

Для объекта **«Существующее положение»** рассмотрены следующие семантически схожие тексты. Первый текст: «В цехе № по производству термического технического углерода циклоны «х» оснащены ворошителями, предназначенными для очистки стенок от налипшего технического углерода, его разрыхления и свободного падения т/у в шлюзовый затвор для дальнейшей дозированной выгрузки в транспортную систему. Привод ворошителей осуществлен посредством передачи крутящего момента от привода через червячную передачу».

Второй текст: «На производстве термического технического углерода установлены в циклоне ворошители, состоящие из П образных ножей закрепленных при помощи болтов к валу и приводимых через червячную передачу эл.мотором и служат для срезания со стенок т/у и его разрыхления для свободного падения в шлюзовой затвор».

Для объекта **«Обоснование эффекта»** рассмотрены следующие тексты. Первый текст: «Использование данного предложения позволит более оперативно предпринимать действия для восстановления передачи данных блока подготовки газа факела в цехе №, повысит информативность скада системы WinCC АРМ 1, 2 оператора по наливу СК и СГ».

Второй текст: «Использование данного предложения позволит более оперативно предпринимать действия для восстановления измерительных каналов программируемого логического контроллера блока управления факелом цеха №, повысит информативность скада системы WinCC».

Для объекта **«Недостатки существующего положения»** выявлено сходство между текстами: «Существующий прототип отсутствует» и «прототип отсутствует».

Приведённые примеры демонстрируют способность модели выявлять семантическую близость текстов даже при существенных различиях в используемой лексике, структуре предложений и объёме описания.

Эволюционная оптимизация модели

Шесть текстовых объектов рационализаторского предложения вносят различный вклад в итоговую оценку семантической близости. Для повышения качества классификации была выполнена оптимизация весовых коэффициентов на основе критерия оптимальности — максимизации качества разделения дублирующих и недублирующих пар предложений в рамках конкретной обучающей выборки — экспертно размеченной совокупности пар предложений, используемой для оценки каждой комбинации весов.

Поскольку целевая функция не является гладкой и может содержать локальные максимумы, для поиска оптимальных весов применён эволюционный алгоритм (в соответствии с рисунком 3, алгоритм 2).

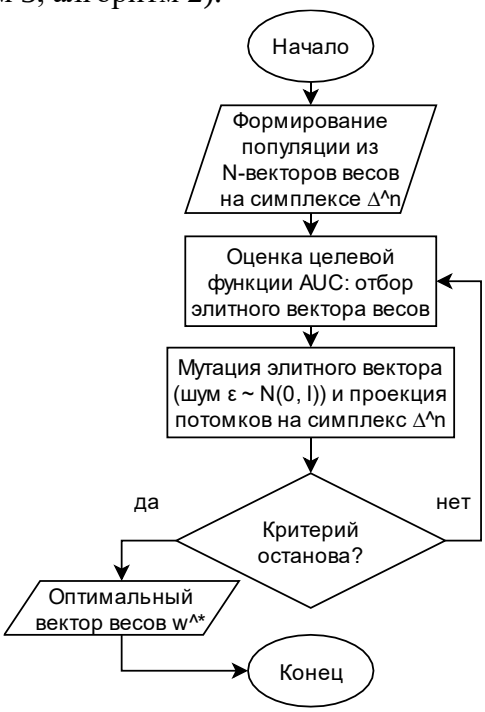


Рисунок 3 – Блок-схема работы эволюционного алгоритма

Алгоритм 2 – Оптимизация весов текстовых объектов рационализаторского предложения эволюционным алгоритмом

1. Вход: экспертно размеченная выборка пар рационализаторских предложений (дубликаты и недубликаты); размер популяции N.
2. Сформировать начальную популяцию из N векторов весов текстовых объектов, лежащих на симплексе (сумма компонент вектора равна 1).
3. Оценить целевую функцию (AUC) для каждого вектора весов популяции на размеченной выборке; отобрать элитный вектор весов с максимальным значением AUC.
4. Сформировать потомков путём мутации элитного вектора весов (добавление шума ~ N(0, I)) с последующей проекцией потомков на симплекс.
5. Повторять шаги 3–4, пока не выполнится критерий остановки (достижение максимального числа итераций).
6. Выход: оптимальный вектор весов, соответствующий элитному решению последней итерации.

Целевой функцией оптимизации выступает площадь под ROC-кривой (AUC, Area Under the ROC Curve). ROC-кривая строится путём варьирования порога классификации θ и отображает зависимость доли истинно положительных классификаций (TPR) от доли ложно положительных (FPR) [20], [21]:

$$TPR(\theta) = TP(\theta) \,/\, (TP(\theta) + FN(\theta)), FPR(\theta) = FP(\theta) \,/\, (FP(\theta) + TN(\theta)),$$

где $TPR(\theta)$ и $FPR(\theta)$ — доли верно и ошибочно определённых дубликатов при пороге θ; $TP(\theta)$, $FN(\theta)$, $FP(\theta)$, $TN(\theta)$ — количества верно и ошибочно классифицированных пар (дубликат / недубликат) при пороге θ.

AUC вычисляется по формуле Уилкоксона–Манна–Уитни (3) как вероятность того, что модель присвоит случайной паре-дубликату более высокую оценку схожести, чем случайной паре-недубликату:

$$AUC = P\big(S(A^+, B^+;\, w) > S(A^-, B^-;\, w)\big), \quad (3)$$

где $S(A, B;\, w)$ — итоговая оценка семантической близости пары предложений A и B при векторе весов w; (A^+, B^+) — дублирующая пара, (A^-, B^-) — недублирующая пара.

Значение AUC = 1.0 соответствует идеальному разделению классов, AUC = 0.5 — случайному угадыванию; метрика выбрана как целевая функция, поскольку не зависит от порога θ.

В результате оптимизации получена комбинация весовых коэффициентов, обеспечивающая наилучшее разделение дублирующих и недублирующих пар. Оптимальные значения приведены в таблице 2, а сравнение качества классификации при начальных и оптимизированных весах представлено в соответствии с рисунком 4.

Таблица 2 – Оптимизированные весовые коэффициенты объектов рационализаторского предложения

Объект рационализаторского предложения	**Оптимизированный вес**
Наименование предложения	0.000
Цель предложения	0.121
Сущность предложения	0.014
Существующее положение	0.667
Обоснование эффекта	0.030
Недостатки существующего положения	0.167

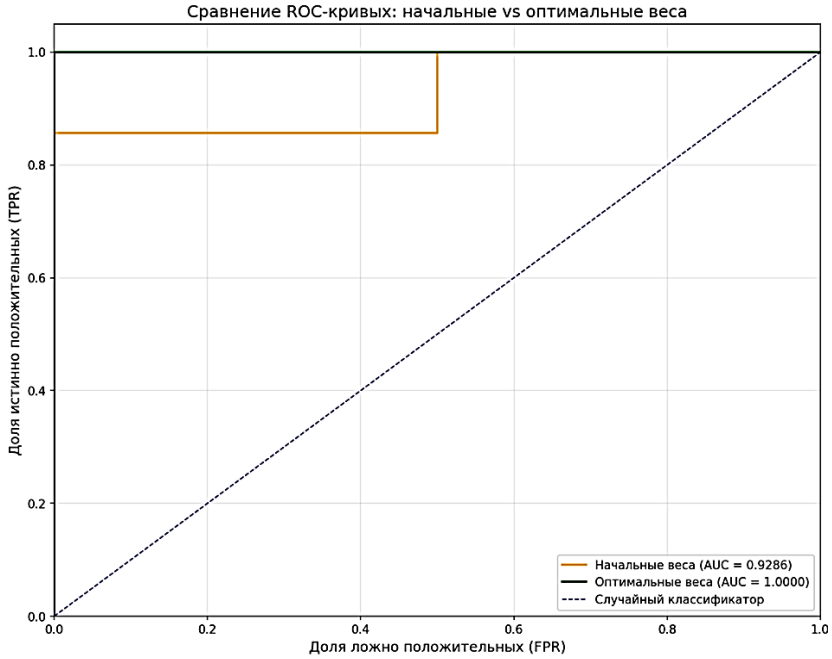


Рисунок 4 – Сравнение ROC-кривых при начальных и оптимальных весах

Наибольший вес (0.667) получил объект «существующее положение». Описание текущей проблемы наиболее информативно для выявления дублирования независимо от формулировки остальных полей. Наименьший вес (0.000) присвоен объекту «название предложения», которое формулируется свободно и потому ненадёжно как признак дублирования.

Экономическое обоснование

Экономическая эффективность внедрения модуля строится на двух уровнях: прямом сокращении затрат и повышении качества работы с инновациями.

Прямой эффект получается за счет сокращения трудозатрат экспертов на проверку новизны предложений (в среднем 5 минут, или 0.08 ч, на инициативу): 1.6 млн руб. в год;

Поскольку модуль работает автономно и лишь уведомляет о гипотезе дублирования, не останавливая процесс, эксплуатационные расходы близки к нулю; с учётом налога на прибыль (25%) **чистый годовой поток** составляет 1.2 млн руб.;

При ставке дисконтирования 12% и начальных инвестициях 1.41 млн руб. (2026 г.) накопленный **чистый дисконтированный поток** выходит в положительную зону в 2028 году и достигает 2.92 млн руб. к 2031 году (в соответствии с рисунком 5).

Таким образом, **чистая приведённая стоимость** проекта (NPV) за пятилетний период составляет 2.92 млн руб., что подтверждает экономическую целесообразность внедрения модуля.

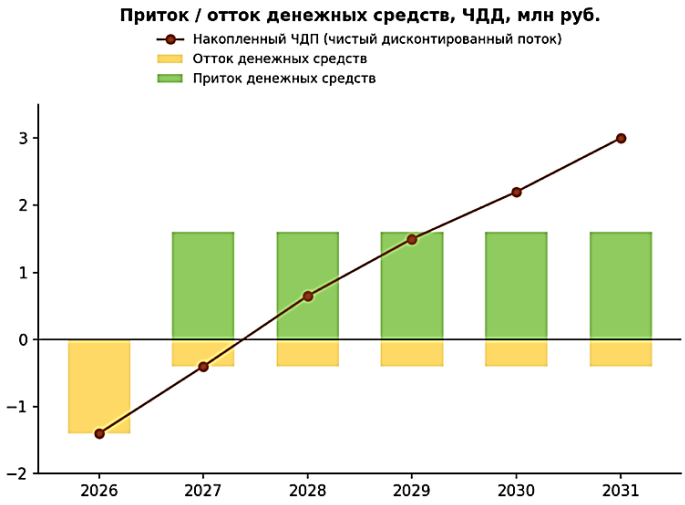


Рисунок 5 — Приток / отток денежных средств, ЧДД, млн руб.

Заключение

В работе решена задача автоматического выявления дубликатов рационализаторских предложений в информационной системе управления рационализаторской деятельностью с использованием embedding-модели и косинусного сходства для оценки семантической близости текстов.

Сравнение трёх моделей на выборке из 221 предложения показало, что наилучшие результаты продемонстрировала модель GigaEmbeddings. Она выявила наибольшее количество семантически близких пар, включая уникальные совпадения, не обнаруженные другими моделями. Для повышения точности выполнена оптимизация весов шести текстовых объектов с применением эволюционного алгоритма и целевой функцией AUC на экспертно размеченной выборке. Наибольший вес получил объект «существующее положение» (0.667), а наименьший — «наименование предложения» (0.000).

Экономический эффект внедрения модуля связан с сокращением трудозатрат. Чистая приведенная стоимость проекта (NPV) за пятилетний период составляет 2.92 млн руб. при начальных инвестициях 1.41 млн руб. в 2026 году.

Список источников

1. Алещенко В.В. Методы автоматического выявления дублирующих документов в корпоративных информационных системах / В.В. Алещенко, В.В. Козлов // Вестник компьютерных и информационных технологий. 2022. № 3. С. 24–33. DOI: 10.14489/vkit.2022.03. Pp.024-033.
2. Aggarwal C.C. Mining Text Data / C.C. Aggarwalю, C. Zhai. New York: Springer, 2012. 524 p. DOI: 10.1007/978-1-4614-3223-4.
3. Никитенко К.А. Интерпретируемость нейросемантических моделей обработки естественного языка / К.А. Никитенко, А.В. Звягинцева // Проблемы искусственного интеллекта. 2025. № 2 (37). С. 3–18.
4. Семантический анализ текста с использованием искусственных нейронных сетей на основе нейроподобных элементов с временным суммированием сигналов / Харламов А.А., Самаев Е.С., Кузнецов Д.Ю., Пантюхин Д.В. // Проблемы искусственного интеллекта. 2023. № 3 (30). С. 45–56.
5. Харламов А.А. Перспективы развития семантических представлений, основывающиеся на тенденциях нейроинформатики / А.А. Харламов // Проблемы искусственного интеллекта. 2020. № 3 (18). С. 8–19.
6. Reimers N. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks / N. Reimers, I. Gurevych // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP). Hong Kong, 2019. P. 3982–3992. DOI: 10.18653/v1/D19-1410.
7. Дорохина Г.В. Требования к информационной технологии цифрового сбора, обработки и анализа данных / Г.В. Дорохина // Проблемы искусственного интеллекта. 2020. № 3 (18). С. 44–55.
8. Efficient Estimation of Word Representations in Vector Space [Электронный ресурс] / Mikolov T., Chen K., Corrado G., Dean J. // arXiv preprint arXiv:1301.3781. 2013. URL: https://arxiv.org/abs/1301.3781 (дата обращения: 26.04.2026).
9. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / Devlin J., Chang M.-W., Lee K., Toutanova K. // Proceedings of NAACL-HLT 2019. Minneapolis, 2019. P. 4171–4186. DOI: 10.18653/v1/N19-1423.
10. Snegirev T. ruMTEB: Russian Massive Text Embedding Benchmark [Электронный ресурс] / B. Malashenko, D.I. Ignatov // arXiv preprint arXiv:2408.12503. 2024. URL: https://arxiv.org/abs/2408.12503 (дата обращения: 26.04.2026).
11. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation [Электронный ресурс] / Chen J., Xiao S., Zhang P., Luo K. // arXiv preprint arXiv:2402.03216. 2024. URL: https://arxiv.org/abs/2402.03216 (дата обращения: 26.04.2026).
12. Malashenko B. USER2: Universal Sentence Encoder for Russian [Электронный ресурс] / B. Malashenko, E. Spirin, A. Sokolov 2025. URL: https://huggingface.co/deepvk/USER2-base (дата обращения: 26.04.2026).
13. Ванюшин Е.С. Адаптация русскоязычных текстов с сохранением семантики / Е.С. Ванюшин, Д.В. Белоусова // Проблемы искусственного интеллекта. 2024. № 3 (34). С. 37–48.
14. Kolodin E. GigaEmbeddings – Efficient Russian Language Embedding Model [Электронный ресурс] / E. Kolodin, A. Ianina // Proceedings of the 10th Workshop on Slavic Natural Language Processing (Slavic NLP 2025). Vienna, 2025. С. 17–24. URL: https://aclanthology.org/2025.bsnlp-1.3 (дата обращения: 26.04.2026).
15. Text Embeddings by Weakly-Supervised Contrastive Pre-training [Электронный ресурс] / Wang L., Yang N., Huang X. [et al.] // arXiv preprint arXiv:2212.03533. 2022. URL: https://arxiv.org/abs/2212.03533 (дата обращения: 26.04.2026).
16. Universal Sentence Encoder [Электронный ресурс] / Cer D., Yang Y., Kong S.-Y. [et al.] // arXiv preprint arXiv:1803.11175. 2018. URL: https://arxiv.org/abs/1803.11175 (дата обращения: 26.04.2026).
17. Беловодский В.Н. Об интерполяции и аппроксимации функций с использованием нейронных сетей / В.Н. Беловодлский // Проблемы искусственного интеллекта. 2024. № 3 (34). С. 74–83.
18. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models [Электронный ресурс] / Thakur N., Reimers N., Rücklé A. [et al.] // Proceedings of the Neural Information Processing Systems (NeurIPS) 2021 Track on Datasets and Benchmarks. 2021. URL: https://arxiv.org/abs/2104.08663 (дата обращения: 26.04.2026).
19. Hansen N. Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES) / N. Hansen, S.D. Müller, P. Koumoutsakos // Evolutionary Computation. 2003. Vol. 11. No. 1. P. 1–18. DOI: 10.1162/106365603321828970.
20. Fawcett T. An introduction to ROC analysis / T. Fawcett // Pattern Recognition Letters. 2006. Vol. 27. No. 8. P. 861–874. DOI: 10.1016/j.patrec.2005.10.010.
21. Hand D.J. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems / D.J. Hand, R.J. Till // Machine Learning. 2001. Vol. 45. No. 2. P. 171–186. DOI: 10.1023/A:1010920819831.

References

1. Aleshchenko V.V., Kozlov V.V. Methods for automatic detection of duplicate documents in corporate information systems. Vestnik komp'yuternykh i informatsionnykh tekhnologiy [Herald of Computer and Information Technologies]. 2022. No. 3. pp. 24–33. DOI: 10.14489/vkit.2022.03.pp.024-033. (In Russian).
2. Aggarwal C.C., Zhai C. Mining Text Data. New York: Springer, 2012. 524 p. DOI: 10.1007/978-1-4614-3223-4.
3. Nikitenko K.A., Zvyagintseva A.V. Interpretability of neurosemantic models for natural language processing. Problemy iskusstvennogo intellekta [Problems of Artificial Intelligence]. 2025. No. 2 (37). pp. 3–18. (In Russian).
4. Kharlamov A.A., Samaev E.S., Kuznetsov D.Yu., Pantyukhin D.V. Semantic text analysis using artificial neural networks based on neuro-like elements with temporal signal summation. Problemy iskusstvennogo intellekta [Problems of Artificial Intelligence]. 2023. No. 3 (30). pp. 45–56. (In Russian).
5. Kharlamov A.A. Prospects for the development of semantic representations based on trends in neuroinformatics. Problemy iskusstvennogo intellekta [Problems of Artificial Intelligence]. 2020. No. 3 (18). pp. 8–19. (In Russian).
6. Reimers N., Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of EMNLP-IJCNLP. Hong Kong, 2019. pp. 3982–3992. DOI: 10.18653/v1/D19-1410.
7. Dorokhina G.V. Requirements for information technology of digital data collection, processing and analysis. Problemy iskusstvennogo intellekta [Problems of Artificial Intelligence]. 2020. No. 3 (18). pp. 44–55. (In Russian).
8. Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781. 2013. URL: https://arxiv.org/abs/1301.3781 (accessed 26.04.2026).
9. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT 2019. Minneapolis, 2019. pp. 4171–4186. DOI: 10.18653/v1/N19-1423.
10. Snegirev T., Malashenko B., Ignatov D.I. ruMTEB: Russian Massive Text Embedding Benchmark. arXiv preprint arXiv:2408.12503. 2024. URL: https://arxiv.org/abs/2408.12503 (accessed 26.04.2026).
11. Chen J., Xiao S., Zhang P., Luo K. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. arXiv preprint arXiv:2402.03216. 2024. URL: https://arxiv.org/abs/2402.03216 (accessed 26.04.2026).
12. Malashenko B., Spirin E., Sokolov A. USER2: Universal Sentence Encoder for Russian. 2025. URL: https://huggingface.co/deepvk/USER2-base (accessed 26.04.2026).
13. Vanyushin E.S., Belousova D.V. Adaptation of Russian-language texts while preserving semantics. Problemy iskusstvennogo intellekta [Problems of Artificial Intelligence]. 2024. No. 3 (34). pp. 37–48. (In Russian).
14. Kolodin E., Ianina A. GigaEmbeddings — Efficient Russian Language Embedding Model. Proceedings of the 10th Workshop on Slavic NLP 2025. Vienna, 2025. pp. 17–24. URL: https://aclanthology.org/2025.bsnlp-1.3 (accessed 26.04.2026).
15. Wang L., Yang N., Huang X. et al. Text Embeddings by Weakly-Supervised Contrastive Pre-training. arXiv preprint arXiv:2212.03533. 2022. URL: https://arxiv.org/abs/2212.03533 (accessed 26.04.2026).
16. Cer D., Yang Y., Kong S.-Y. et al. Universal Sentence Encoder. arXiv preprint arXiv:1803.11175. 2018. URL: https://arxiv.org/abs/1803.11175 (accessed 26.04.2026).
17. Belovodsky V.N. On interpolation and approximation of functions using neural networks. Problemy iskusstvennogo intellekta [Problems of Artificial Intelligence]. 2024. No. 3 (34). pp. 74–83. (In Russian).
18. Thakur N., Reimers N., Rücklé A. et al. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. Proceedings of NeurIPS 2021 Track on Datasets and Benchmarks. 2021. URL: https://arxiv.org/abs/2104.08663 (accessed 26.04.2026).
19. Hansen N., Müller S.D., Koumoutsakos P. Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES). Evolutionary Computation. 2003. Vol. 11. No. 1. pp. 1–18. DOI: 10.1162/106365603321828970.
20. Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters. 2006. Vol. 27. No. 8. pp. 861–874. DOI: 10.1016/j.patrec.2005.10.010.
21. Hand D.J., Till R.J. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. Machine Learning. 2001. Vol. 45. No. 2. pp. 171–186. DOI: 10.1023/A:1010920819831.

RESUME

I. A. Korshakova

Automating the detection of duplicate rationalization proposals using text similarity analysis

The rationalization management information system of fuel and energy complex enterprises accumulates semantically similar and duplicate initiatives whose identification is complicated by differences in wording.

To address this issue, embedding model and semantic text similarity method were applied. A comparative study of the BAAI/bge-m3, deepvk/USERv2, and GigaEmbeddings models was conducted on a dataset of 221 rationalization proposals.

The results demonstrated the superiority of the GigaEmbeddings model in detecting semantically similar initiatives. In addition, the weights of proposal structural elements were optimized using an evolutionary algorithm.

The implementation of the proposed module will automate the detection of duplicate initiatives and reduce labor costs associated with proposal processing.

РЕЗЮМЕ

И.А. Коршакова

Автоматизация выявления дубликатов рационализаторских предложений за счет анализа текстовой схожести

В информационной системе управления рационализаторской деятельностью предприятия «ТЭК» накапливаются семантически схожие и дублирующие инициативы, выявление которых затруднено различиями в формулировках.

Для решения задачи применяются embedding-модель и метод семантического сравнения текстов. Проведено сравнение моделей BAAI/bge-m3, deepvk/USERv2 и GigaEmbeddings на выборке из 221 рационализаторского предложения.

Исследование показало превосходство модели GigaEmbeddings в выявлении семантически близких инициатив. Дополнительно выполнена оптимизация весовых коэффициентов структурных элементов предложений с использованием эволюционного алгоритма.

Внедрение модуля позволит автоматизировать поиск дублирующих инициатив и сократить трудозатраты при обработке предложений.

Коршакова Ирина Александровна – разработчик Управления моделирования и бизнес-анализа ООО «Газпром информ», 196143, г. Санкт-Петербург, пл. Победы, д. 2, лит. А, пом. 1-Н, каб. 422, телефон +7(931)369-1837, korshakova40@gmail.com. *Область научных интересов*: DevOps-инженерия, нейронные сети, интеллектуальный анализ данных. ORCID: 0009-0003-3591-1509.

Korshakova Irina Aleksandrovna – Developer, Modeling and Business Analysis Department, Gazprom Inform LLC, Office 422, Premises 1-N, Lit. A, 2 Pobedy Square, Saint Petersburg, 196143, Russian Federation. Tel.: +7 (931) 369-1837, e-mail: korshakova40@gmail.com. *Research interests*: DevOps engineering, neural networks, data mining. ORCID: 0009-0003-3591-1509.

Статья поступила в редакцию 03.08.2026;
одобрена после рецензирования 28.08.2026; принята к публикации 02.09.2026
The article was submitted 03.08.2026; approved after reviewing 28.08.2026;
accepted for publication 02.09.2026.