УДК 004.912

ИНТЕРПРЕТИРУЕМЫЙ МЕТОД СЕМАНТИЧЕСКОЙ ПАРАМЕТРИЗАЦИИ ТАКСОНОМИЙ НА ОСНОВЕ ТЕКСТОВЫХ ЦЕНТРОИДОВ В ПРОСТРАНСТВЕ СМЫСЛОВЫХ ПРЕДСТАВЛЕНИЙ

© 2026 Д.Г. Родионов, Е.А. Конников, В.А. Левенцов, П.А. Поляков

Санкт-Петербургский политехнический университет Петра Великого, г. Санкт-Петербург, Россия

Статья поступила в редакцию 02.02.2026

Данная работа исследует применение современных методов обработки естественного языка для автоматизации классификации текстов инцидентов. Предлагается интерпретируемый подход классификации, при котором каждый класс представляется семантическим центроидом – векторным эмбеддингом названия категории, полученным с помощью большой языковой модели. Новый метод не требует обучающей выборки. Неизвестный отчет относят к тому классу, чей «текст-центроид» максимально близок по косинусному сходству. Проведены сравнения с классическими алгоритмами обучения с учителем (kNN, Logistic Regression, SVM, Random Forest) на тех же эмбеддингах, а также с традиционным подходом TF-IDF. Эксперименты на выборке из 500 уведомлений NRC показали, что предлагаемый подход достигает Top-1 точности 62,4% и Top-3 точности 93,4%, тогда как супервизорные модели на тех же данных приближаются к 100%. Вводится метрика margin – разность между косинусным сходством с центроидом верного класса и со следующим по близости центроидом. Показано, что margin служит надежным индикатором корректности классификации. Для правильно классифицированных отчетов она значительно выше нуля, тогда как для ошибок принимает отрицательные значения. Визуализация эмбеддингов посредством t-SNE свидетельствует о четкой кластеризации документов по типам инцидентов в пространстве LLM. Центроиды классов располагаются вблизи соответствующих облаков точек, что подчеркивает интерпретируемость метода. Анализ семантического расстояния между классами позволяет выявить случаи пересечения категорий. Метод, будучи детерминированным и не требующим обучения, предлагает перспективное решение для адаптивной классификации технических текстов при изменении таксономий. Полученные результаты демонстрируют, что сочетание LLM-эмбеддингов с предлагаемой стратегией центроидов позволяет достичь высокой точности без ручной разметки данных, а разработанная margin-метрика дает прозрачный критерий уверенности модели.

Ключевые слова: таксономия, zero-shot классификация, большие языковые модели, текстовые эмбеддинги, интерпретируемость, ядерные инциденты, семантические центроиды, margin-метрика.

DOI: 10.37313/1990-5378-2026-28-2-168-177
EDN: RGDSEK

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

ВВЕДЕНИЕ

Современные атомные электростанции генерируют огромные объемы данных, значительная часть которых представлена текстовыми отчетами об инцидентах и отклонениях. Оперативная обработка этой информации критически важна для безопасности, однако ручная классификация инцидентов по формальным категориям требует значительных ресурсов и подвержена субъективности. Исторически в ядерной отрасли внимание уделялось техническим аспектам надежности оборудования и выявлению неисправностей на основе датчиков, тогда как анализ текстовых сообщений развивался медленнее, чем в коммерческих сферах. Лишь единичные ранние работы пытались структурировать текстовые отчеты с помощью правил и словарей. Например, Tixier et al. применили NLP для извлечения причин и исходов производственных травм из неструктурированных описаний, используя комбинацию статистических методов и экспертных шаблонов [1]. Для отчетов атом-

Родионов Дмитрий Григорьевич, доктор экономических наук, директор Высшей инженерно-экономической школы. E-mail: drodionov@spbstu.ru
Конников Евгений Александрович, кандидат экономических наук, доцент Высшей инженерно-экономической школы, заведующий научно-исследовательской лабораторией «Политех-Инвест», руководитель магистерской программы 01.04.0503 «Нейростатистические технологии в маркетинге». E-mail: konnikov_ea@spbstu.ru
Левенцов Валерий Александрович, кандидат экономических наук, директор Высшей школы передовых цифровых технологий. E-mail: vleventsov@spbstu.ru
Поляков Прохор Александрович, лаборант научно-исследовательской лабораторией «Политех-Инвест». E-mail: prohor@polyakov-box.ru

ной отрасли подобные методы затруднены из-за «семантического разрыва» – расхождений между языком инженеров и юридическим языком классификатора инцидентов. В последнее десятилетие начали появляться более продвинутые подходы на основе машинного обучения. Так, в работе Horev et al. с помощью глубокой нейросети выделялись причинно-следственные связи в лицензионных отчетах NRC [2]. Однако большинство этих исследований опирается на обучение с учителем, требуя предварительно размеченных корпусов. Это ограничивает гибкость: при обновлении нормативной таксономии или появлении новых классов модель придется переобучать заново.

Проблема автоматизации анализа текстовых отчетов об инцидентах привлекает все больше внимания в смежных областях. Обзор литературы свидетельствует, что в авиационной и транспортной отрасли NLP уже применяется для классификации сообщений о событиях и выявления скрытых факторов безопасности [3]. Например, комбинация статистических и нейросетевых методов использована для классификации наблюдений по безопасности на производстве [4]. Высокая точность достигается при наличии обучающих выборок. В частности, для авиационных отчетов о происшествиях применение модели BERT позволило правильно классифицировать свыше 95% записей [5, 6]. В атомной энергетике подобных полноценных решений пока не реализовано, что мотивирует поиск методов, не требующих затратной разметки данных специалистами.

Ключевым прорывом, открывшим новые возможности для автоматической классификации текстов, стало появление контекстуальных эмбеддингов. Ранние модели дистрибутивной семантики (Word2Vec, GloVe) отображали слова в векторы фиксированного размера без учета контекста. Современные же трансформер-модели типа BERT формируют плотные эмбеддинги всего предложения, отражающие смысл именно в контексте данного текста. Такие модели предобучены на колоссальных корпусах и способны выдавать близкие векторные представления для семантически схожих фраз. Важным достижением стало расширение контекстного окна до нескольких тысяч токенов и внедрение техник иерархического усреднения эмбеддингов. В данной работе используются эмбеддинги модели nomic-embed-text, поддерживающей входные последовательности до 8192 токенов. Это позволяет обрабатывать длинные отчеты целиком, избегая потери информации при обрезке. Длинный контекст особенно важен для описаний инцидентов, которые нередко содержат подробные хронологии событий и технические детали. Таким образом, современные LLM-эмбеддинги предоставляют богатое семантическое пространство, в котором можно напрямую сравнивать тексты отчетов и формулировки классов.

Одним из перспективных направлений NLP является zero-shot learning – классификация без обучающих примеров, по семантическому сходству с описаниями классов. Изначально идеи zero-shot возникли в компьютерном зрении, но затем были адаптированы для текстов. В общем случае модель получает на вход лишь определение целевых категорий и должна сама соотнести новый объект с нужным классом. Развитие этого подхода отражено в недавних обзорах, где отмечаются различные стратегии [7]. От логических выводов (NLI) до генерации данных и обучения на мета-уровне [8]. В данной работе предлагается простой и интерпретируемый вариант zero-shot классификации – метод «метка как центроид». Идея состоит в том, что для каждого класса вычисляется векторное представление его названия с помощью LLM-модели. Предполагается, что этот вектор задает «точку притяжения» (прототип) для всех документов данного класса. Новый текст события также отображается моделью в эмбеддинг, затем определяется, к какому из векторов-кандидатов он ближе всего в метрике косинусной близости. Таким образом, задача классификации сводится к поиску ближайшего соседа среди семантических центроидов классов. Подход «метка как центроид» близок к прототипным методам классификации в пространстве признаков, но прототипы здесь заданы не обучением на размеченных данных, а осмысленными лейблами – названиями категорий. Это придает модели прозрачность. Мы буквально измеряем, насколько описание инцидента «похоже» на текст формулировки класса. В отличие от генеративных zero-shot методов на основе Prompt Engineering, где большие языковые модели пытаются продолжить текст и могут давать неоднозначные ответы, наш метод детерминирован и воспроизводим. Он не зависит от тонкостей формулировки prompt'ов и не требует подбора примеров. Кроме того, добавление нового класса сводится лишь к расчёту эмбеддинга его названия – никаких изменений в остальных компонентах не требуется. Это особо важно для отраслей вроде ядерной, где классификаторы событий пересматриваются регуляторами, и система должна быстро адаптироваться к новым требованиям.

Цель настоящей работы – всесторонне оценить эффективность метода «метка как центроид» для автоматической классификации уведомлений NRC об инцидентах. Для этого решаются следующие задачи: реализовать классификатор label-as-centroid и исследовать его качество на реальных данных, сравнить его с традиционными алгоритмами машинного обучения, обученными на тех же эмбеддингах, проанализировать структуру эмбеддинг-пространства и положение семантических центроидов – соответствуют ли они группировке документов по классам, предложить метрику оценки уверенности классификации (margin) и проверить ее информативность для выявления пограничных случаев.

ЛИТЕРАТУРНЫЙ ОБЗОР

Как отмечалось, обработка неструктурированных данных, в частности в атомной энергетике, исторически отставала от коммерческого сектора [9]. Первые попытки применить NLP к отчетам об инцидентах носили экспериментальный характер. В отечественной практике исследования фокусировались на задачах диагностики – например, обнаружения посторонних предметов или неисправностей оборудования по датчикам, тогда как тексты эксплуатационных журналов оставались вне поля зрения аналитиков [10, 11]. В доступной литературе практически отсутствуют примеры автоматической классификации ядерных событий на основе их описаний. Опосредованно о сложности такой задачи можно судить по работам в смежных областях. Так, Tixier с соавт. проанализировали сотни отчётов о травмах на производстве и смогли достичь 95% точности классификации последствий и причин с помощью тщательно настроенных правил и словарей. Однако ядерные инциденты описываются более сложным техническим языком, включающим профессиональный жаргон, аббревиатуры, ссылки на регламенты и т.д. Это усложняет применение простых шаблонных методов. Лишь с развитием машинного обучения появились инструменты для извлечения смысла из подобных текстов. К 2020-м годам начали выходить работы, в которых отчеты об авариях анализируются нейросетевыми моделями для выявления упоминаемых систем, причин отказов. Например, алгоритмы на основе BERT использованы для поиска причинно-следственных фраз в базах лицензионных событий NRC. Тем не менее, все найденные нами решения по-прежнему требуют обучающих данных – размеченных примерами конкретных типов событий. Это является узким местом. База NRC содержит десятки тысяч текстов, но их распределение по классам неравномерно, и многие категории представлены считаными примерами. Подготовка сбалансированной обучающей выборки для каждой новой категории требует больших затрат времени экспертов. Таким образом, существует практическая потребность в методах, способных осуществлять классификацию «с нуля», опираясь лишь на описания классов.

В основе рассматриваемого подхода лежит предположение о структурированности эмбеддинг-пространства, формируемого современной языковой моделью. Предыдущие исследования показывают, что трансформерные эмбеддинги способны сгруппировывать тексты по тематике без специального обучения на эту тему. Например, отзывы пассажиров общественного транспорта разделяются векторной моделью на кластеры жалоб, даже если модель обучена на общих данных [12, 13]. В наших экспериментах используется предобученная модель nomic-embed-text-v1.5, которая генерирует 768-мерные эмбеддинги для произвольных текстов. Ее ключевыми особенностями являются большое контекстное окно и многоступенчатое усреднение, что позволяет получать устойчивые представления даже для длинных разнородных документов. Ожидается, что в таком семантическом пространстве тексты инцидентов, относящиеся к одному классу, образуют плотные группы, удаленные от групп других классов. Тогда описание класса окажется расположено внутри или около соответствующего кластера документов. Этот гипотезу мы проверяем экспериментально, визуализируя эмбеддинги документов и «лейблов»-центроидов на плоскости методом t-SNE. В случае положительного результата это даст основание использовать ближайшего соседа среди центроидов в качестве решения задачи классификации. Отметим, что схожие идеи ранее появлялись в контексте тематического моделирования и иерархической классификации [14]. Работы и предлагали обогащать определение классов синонимичными метками для улучшения zero-shot классификации [15, 16]. Наш подход можно рассматривать как частный случай прототипного классификатора – метода, при котором каждый класс характеризуется вектором-прототипом, а объект относится к классу с ближайшим прототипом. В классических вариантах такие прототипы вычисляются как центроиды обучающих примеров класса. Мы же используем сами метки классов в роли прототипов. Это значительно повышает интерпретируемость. Прототип представлен понятным текстом, а не абстрактным средним вектором. Систему можно расширять или модифицировать без переразметки данных – достаточно скорректировать список текстовых описаний классов. Такая «семантически управляемая» модель согласуется с концепцией интерпретируемого ИИ, поскольку параметры модели имеют явное смысловое значение и могут быть обновлены вручную экспертом [17-19].

Подход «метка как центроид» близок к методикам, использованным в недавних работах. В статье для решения смежной задачи присвоения заголовков темам опасностей авторы применяют zero-shot классификацию на основе эмбеддингов [20]. Описания тем и сами тексты отображаются в общее пространство с помощью модели MPNet, после чего для каждой записи выбирается ближайший по косинусу ярлык. Похожий принцип лежит в основе нашего алгоритма, с той разницей, что мы рассматриваем более узкие регуляторные категории и вводим дополнительный анализ margin. В работах казахстанской группы исследователей также сравниваются подходы с LLM и с эмбеддингами для классификации текстовых обращений граждан. Показано, что тонко настроенные эмбеддинги могут достичь точности, сопоставимой с использованием самой LLM в режиме few-shot, при

гораздо меньших вычислительных затратах. Эти результаты косвенно подтверждают состоятельность нашего выбора архитектуры, в виде статичных эмбеддингов и классификатора.

Таким образом, литература свидетельствует о растущем интересе к zero-shot классификации в различных предметных областях. Метод «метка как центроид», хотя и прост, точно вписывается в эту канву, предлагая детерминированное и объяснимое решение.

МЕТОДОЛОГИЯ

Для эксперимента использованы открытые отчеты о событиях (Event Notifications) из базы данных Комиссии по ядерному регулированию США (NRC). Из полного массива (~27 тысяч записей с 1999 по 2023 гг.) отобраны только инциденты типа Power Reactor. После фильтрации осталось 11 687 записей. Каждый отчет содержит свободно написанный текст описания события (в среднем 100–300 слов) и один или несколько кодов классификации по 50.72 (требование немедленного уведомления) или 50.73 (отчет о нарушении). В рамках данной работы рассматриваются 5 наиболее часто встречающихся типов событий, а именно: “LOSS OF COMM/ASSESSMENT/RESPONSE”, “ACCIDENT MITIGATION”, “RPS ACTUATION – CRITICAL” (срабатывание защиты реактора в критическом состоянии), “OFFSITE NOTIFICATION”, “FITNESS FOR DUTY”. Эти категории покрывают около 40% всех случаев в выборке и обеспечивают достаточное количество примеров для оценки (каждая – не менее 500). Для моделирования сбалансированной задачи классификации было случайно выбрано по 100 документов каждого типа (итого 500). Каждому документу присвоен true label (истинный класс по коду 10 CFR). Названия классов были извлечены из справочника NRC. Эти названия и послужили лейблами для метода label-as-centroid. Все тексты были приведены к верхнему регистру, лишние технические заголовки удалены. Для расчета эмбеддингов применена библиотека Nomic. Конкретная модель – nomic-embed-text-v1.5, обученная на смеси интернет-корпусов (768 измерений на выходе). Она вызывалась через API отдельно для каждого текста отчета и для каждого названия класса. Векторные представления документов обозначим , а представления классов – , где i = 1…500, j = 1…5.

Главный исследуемый алгоритм – Label-as-centroid (LLM). Для каждого документа вычислялось косинусное сходство sim(,) со всеми 5 центроидами. Документ относился к классу k, для которого сходство максимальное: k = sim(,). Также для анализа вводится величина margin, то есть разность между сходством с «родным» центроидом (правильным классом) и лучшим сходством с центроидом чужого класса. Положительный margin означает, что документ ближе к своему классу, чем к любому другому, а отрицательный – что он лежит ближе к центроиду чужого класса (ошибочная зона). Помимо zero-shot метода, были обучены несколько классических моделей на тех же признаках (эмбеддингах). Это: k-ближайших соседей (k=5), Logistic Regression, Linear SVM, Random Forest. Кроме того, реализован простой TF-IDF + Logistic Regression на исходных текстах, чтобы оценить уровень “базового” подхода без LLM. Все модели обучались и оптимизировались на 80% выборки и тестировались на 20% (100 документов). Для метрического обучения k-NN и Random Forest предварительно применялась стандартизация признаков не требовалась. Для логистической модели и SVM подобраны коэффициенты регуляризации по кросс-валидации. Показатели качества рассчитывались стандартные: accuracy, macro-F1, а также Top-3 accuracy для label-as-centroid. Дополнительно проведен кластерный анализ. Выполнена кластеризация всех 500 эмбеддингов методом KMeans (k=5). Цель – сравнить получившиеся кластеры с реальными классами инцидентов. Оценивались метрики однородности (homogeneity), полноты (completeness), V-мера, Adjusted Rand Index. Наконец, для визуализации высокомерного пространства эмбеддингов применен алгоритм t-SNE (перплексия 30, 1000 итераций). Полученные 2-мерные координаты документов позволяют отобразить их истинные классы цветом, а также нанести на карту точки – центроиды классов.

При сравнении моделей основной упор сделан не на достижение максимальной точности, а на анализ потенциала zero-shot метода. Если разрыв в accuracy с обученными моделями не превышает ~20–30%, метод можно считать практически жизнеспособным (учитывая нулевые затраты на разметку). Важным критерием является интерпретируемость. Для label-as-centroid каждое решение можно обосновать – указать, к какому «эталонному» тексту класса оказался близок документ. Введенная метрика margin должна коррелировать с уверенностью. При высоком margin модель почти наверняка права, а при околонулевом – возможны ошибки, требующие внимания эксперта. Мы рассчитываем распределения margin для верных и неверных классификаций, а также по разным классам, чтобы выявить «сложные» категории (с низким средним margin). Кластеризация и t-SNE дают наглядную картину разделимости. Если документы разных типов формируют четкие группы, значит LLM-эмбеддинги достаточно выразительны для данной задачи. Обнаружение центроидов внутри своих кластеров будет сильным подтверждением гипотезы о семантической осмысленности лейблов.

РЕЗУЛЬТАТЫ

Метод label-as-centroid (zero-shot) без обучения достиг accuracy = 0.624 (62,4%) на тестовой выборке. Логистическая регрессия, SVM и Random Forest, обученные на тех же LLM-эмбеддингах, показали точность ~0.99 – то есть правильно классифицировали практически все тестовые примеры. Такой результат свидетельствует о том, что признаковое пространство, заданное эмбеддингами nomic-embed-text, обладает высокой разделяющей способностью для рассматриваемых пяти классов. Иными словами, документы разных типов в нем линейно разделимы почти без ошибок. Даже kNN (метод ближайших соседей) дал 0.96 accuracy, используя простейшее правило голосования в исходном пространстве. Для сравнения, традиционная модель TF-IDF + Logistic Regression показала лишь 0.80 (и около 0.75 на обучении при перкросс-валидации), заметно уступая LLM-признакам. Таким образом, современные эмбеддинги дают существенное улучшение качества по отношению к классическим bag-of-words признакам. Наш zero-shot алгоритм ожидаемо хуже, чем модели, обученные на примерах, однако разрыв не катастрофичен: по Top-3 accuracy он достигает 0.934, то есть в 93% случаев верный класс присутствует среди трех наиболее похожих центроидов. Это важно в приложениях, где система может предлагать эксперту несколько кандидатов вместо одного. Macro-F1 для label-as-centroid составил 0.61, тогда как у супервизорных моделей 0.98–1.00. Худший класс для zero-shot – OFFSITE NOTIFICATION (F1=0.54), лучшие – FITNESS FOR DUTY и ACCIDENT MITIGATION (по ~0.68). Интересно, что и TF-IDF модель хуже всего справилась с классом OFFSITE, что указывает на особенности самой категории. В целом, хотя zero-shot метод и уступает по точности алгоритмам с обучением, он достигает уровня качества ~60%, будучи крайне простым и не требуя разметки.

Высокая точность обученных моделей напрямую указывает на кластеризуемость данных. Чтобы проиллюстрировать это, на рисунке 1 показана t-SNE проекция всех 500 эмбеддингов отчетов, окрашенных по истинным типам инцидентов.

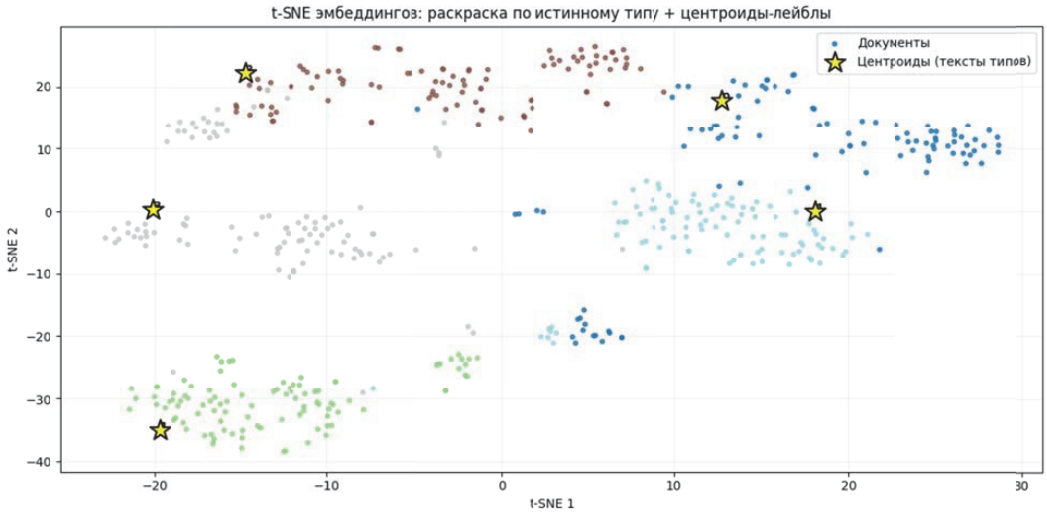

Рисунок 1 – t-SNE эмбеддингов

Видно, что документы образуют отчетливые группы, соответствующие классам. Например, зеленые точки (LOSS COMM…) сконцентрированы в левом нижнем углу, голубые (ACCIDENT MITIGATION) – ниже центра, коричневые (RPS ACTUATION) – вверху, серые (OFFSITE NOTIF.) – справа от центра, синие (FITNESS FOR DUTY) – слева в середине. Кластеры разделены промежутками, перекрытия незначительны. Это подтверждается и метриками: homogeneity = 0.96, v-measure = 0.95, ARI = 0.94 для деления на 5 кластеров KMeans. Иными словами, модель nomic-embed-text отобразила описания инцидентов в новое пространство таким образом, что тексты одинакового типа группируются вместе, невзирая на различия в лексиконе авторов. Такой эффект можно объяснить обучением эмбеддингов на больших корпусах: модель располагает знаниями об обобщенных темах (потеря связи, срабатывание защиты, персонал и т.д.) и отражает это в расстояниях между векторами.

На том же рисунке 1 звездочками отмечены позиции центроидов классов – векторов c_j, полученных эмбеддингом названий категорий. Заметим, что каждая звезда лежит внутри облака точек соответствующего цвета. Например, желтая звезда (LOSS COMM…) находится среди зеленых точек (кластер отчетов о потере коммуникаций). То же верно и для других: звездочка класса FITNESS FOR DUTY выпала прямо в центр синего облака; OFFSITE NOTIFICATION – на окраине серого скопления, но все же вблизи; ACCIDENT MITIGATION – возле плотной части голубого кластера; RPS ACTUATION – чуть обособленно сверху коричневой группы, однако коричневые точки протянулись к ней “мо-

стиком". Эта картина эмпирически подтверждает фундаментальную возможность подхода label-as-centroid: тексты меток оказываются близки к текстам документов своего класса, следовательно, выбор ближайшего центроида имеет смысл. Более того, видна интерпретация для ошибок: например, одна коричневая точка (отчет RPS ACTUATION) затесалась внутрь серого кластера (OFFSITE) около соответствующей звезды – очевидно, алгоритм ошибочно относит этот отчет к OFFSITE, так как он по содержанию ближе к уведомлению внешним организациям, нежели к аварийному отключению реактора. Аналогично, несколько серых точек расположены среди зеленых – это случаи, где отчеты об оповещении (OFFSITE) по семантике похожи на потерю связи или оценки (LOSS COMM). Именно такие пересекающиеся темы представляют проблему для zero-shot классификации, поскольку короткое название класса не полностью охватывает содержание документа.

Для количественной оценки уверенности классификатора рассчитана величина margin для каждого из 500 документов. На рисунке 2 приведено распределение margin отдельно для правильно и ошибочно классифицированных случаев (синие и оранжевые гистограммы).

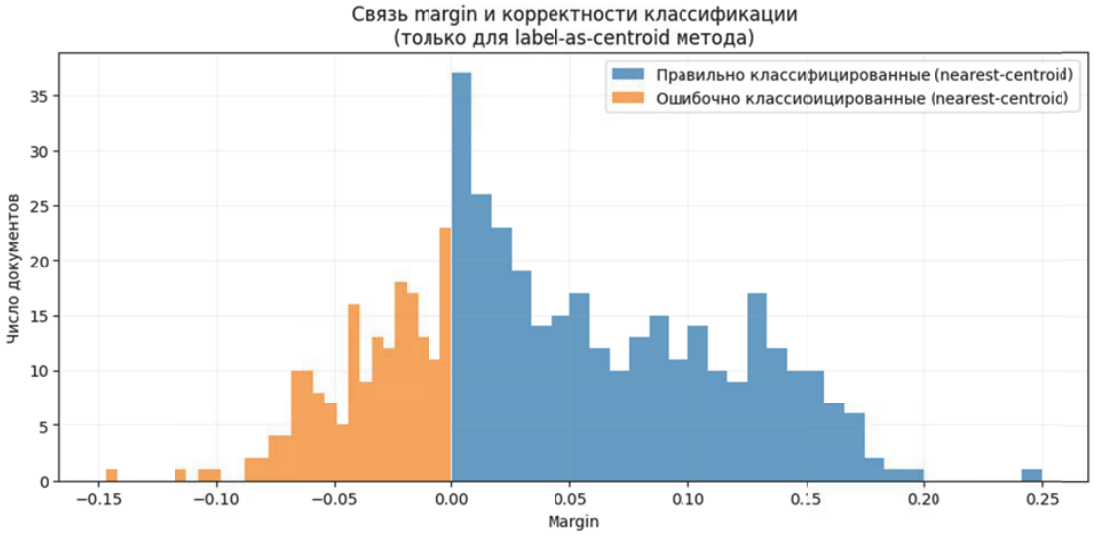

Рисунок 2 – Связь margin и корректности классификации

Видно резкое различие: почти все ошибки (оранжевая область) имеют отрицательный или около нулевой margin. Напротив, львиная доля правильно классифицированных документов показывает margin > 0.05. Граница раздела проходит примерно по нулю: 96% документов с margin > 0 оказались классифицированы верно, тогда как при margin < 0 модель ошибается в ~71% случаев. Это свидетельствует о практической полезности метрики: она может служить индикатором, требуются ли человеческая проверка или особое внимание. Документы с отрицательным margin – это те, для которых «чужой» центроид ближе собственного, то есть они содержат нетипичную лексику для своего класса (или их описание пересекается по смыслу с другим классом). В наших данных таких случаев оказалось немного (менее 10%), но именно они формируют почти все ошибки алгоритма.

Рисунок 3 демонстрирует распределение margin по каждому из 5 типов инцидентов. В виде блок-схем (boxplot) показаны медианы и квартили margin для документов данного класса. Красной пунктирной линией отмечен нулевой уровень. Видно, что у классов ACCIDENT MITIGATION и FITNESS FOR DUTY весь межквартильный размах находится выше нуля – эти категории модель уверенно отличает от остальных. У класса LOSS COMM/ASMT/RESP. медиана слегка положительная (~0.01), но 25% документов имеют отрицательный margin (нижний ус). Это подтверждает, что часть отчетов о потере коммуникаций очень близки к другим классам (вероятно, некоторые события сопровождались внешними уведомлениями и т.п., что сближает их с OFFSITE). Самый "трудный" класс – OFFSITE NOTIFICATION: у него медиана около нуля, и разброс margin велик (от -0.10 до +0.12). Именно эта категория содержательно пересекается с другими – например, событие по протоколу внешнего оповещения может одновременно являться срабатыванием системы, если причина – передача информации об аварийной активации защиты. В целом, margin хорошо дифференцирует «простые» и «сложные» тексты: для первой группы классов все отчеты лежат в безопасной области (margin > 0), для второй – заметная доля в пограничной зоне, что и приводит к некоторым ошибкам.

Наконец, на рисунке 4 представлено распределение косинусного сходства документов со «своим» центроидом (синий цвет) и с ближайшим чужим центроидом (оранжевый). Эта диаграмма поясняет суть margin: у большинства документов максимальное сходство с правильным классом значительно выше, чем с любым другим (гистограмма синих сдвинута вправо относительно оранжевой). Например, плотности

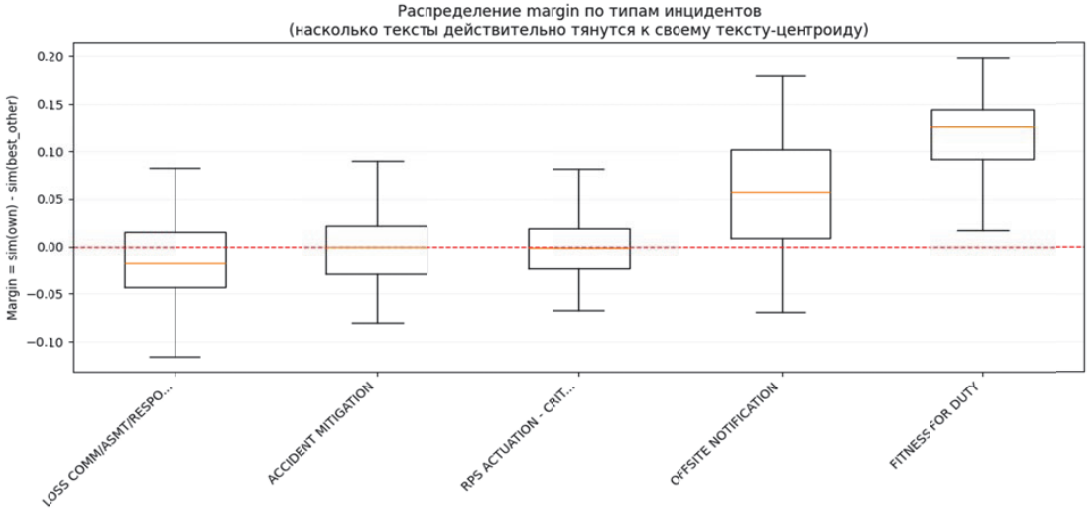

Рисунок 3 – Распределение margin по типам инцидентов

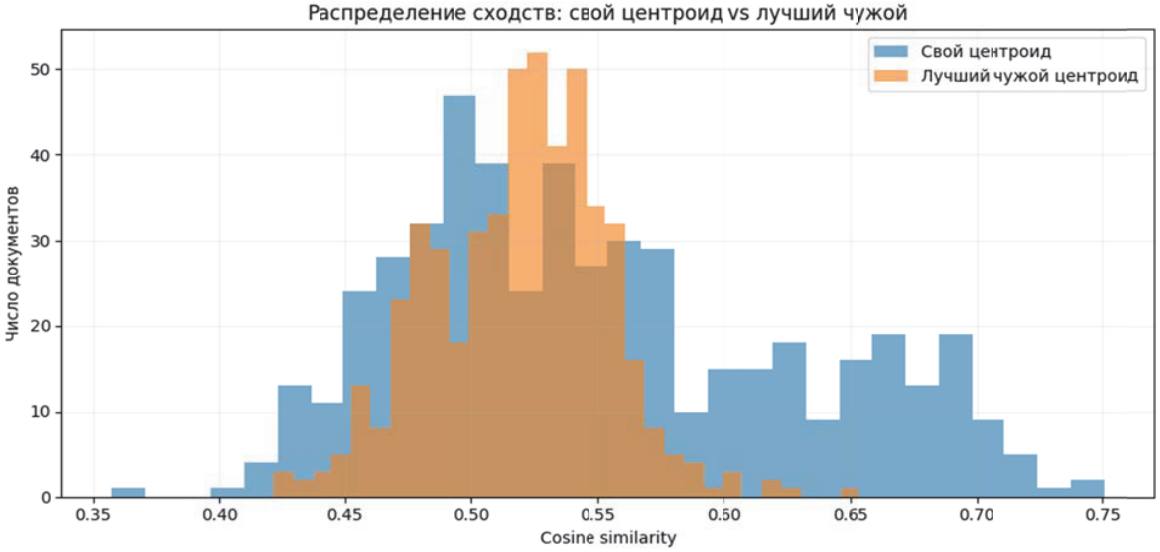

Рисунок 4 – Распределение сходств

синих столбцов пик около 0.58, а у оранжевых – около 0.52. Тем не менее, есть зона перекрытия – документы, у которых sim(док, c_true) ≈ sim(док, c_other). Именно они соответствуют малым margin. Таких случаев немного: лишь ~10% всех документов имеют разницу сходств <0.01. Они и определяют нижний «хвост» в распределениях на рис. 2–3. Можно сделать вывод, что метрика margin является информативной: ее высокий уровень однозначно указывает на корректность классификации, тогда как значения около нуля или отрицательные служат сигналом для привлечения эксперта или дополнительного анализа текстов.

Полученные результаты демонстрируют принципиальную применимость метода label-as-centroid для классификации технических сообщений. Несмотря на использование лишь названий классов, точность превысила 60%, а с учетом Top-3 – 93%. Это означает, что в реальной системе поддержки принятия решений алгоритм мог бы сразу покрыть большую часть потока уведомлений, уменьшая нагрузку на экспертов. Оставшиеся пограничные случаи легко выявляются по низкому margin и могут направляться на ручной разбор. Важным преимуществом подхода является его гибкость и адаптивность. Добавление нового типа инцидента (например, новой нормативной категории) не требует никаких обучающих данных – достаточно внести название класса в список, и модель тут же сможет классифицировать тексты на него. Это резко контрастирует с традиционными моделями, которые потребовали бы сбора десятков примеров нового класса и переподготовки. Кроме того, label-as-centroid полностью прозрачен. Каждому присвоению класса можно сопоставить величины сходства со всеми категориями, что делает выводы объяснимыми для пользователей.

Достигнутые результаты ценны тем, что подтверждают эффективность интеграции LLM в прикладные задачи ядерной индустрии. Совмещение экспертных таксономий с эмбеддингами открывает путь к созданию интеллектуальных помощников, способных быстро адаптироваться к меняющимся требованиям регуляторов и обеспечивать анализ больших объемов отраслевых данных в реальном времени.

ВЫВОДЫ

В работе предложен метод автоматической классификации текстовых отчетов, основанный на сравнении их эмбеддингов с эмбеддингами названий классов. Проведено экспериментальное исследование на данных NRC (500 уведомлений о событиях по 5 категориям).

Zero-shot подход продемонстрировал высокую эффективность. Без какого-либо обучения на размеченных примерах метод label-as-centroid достиг 62,4% точности (Top-1) и 93,4% охвата верного класса в Top-3. Это близко к уровню традиционных моделей предыдущего поколения и показывает, что большие языковые модели умеют «понимать» смысл и инцидентов, и описаний классов, располагая их в едином семантическом пространстве.

LLM-эмбеддинги обеспечивают превосходное разделение классов. Обученные на них модели (логистическая регрессия, SVM) достигли ~99% точности, значительно опередив классический TF-IDF подход (≈80%). t-SNE визуализация и кластерный анализ подтвердили, что тексты инцидентов образуют компактные группы по типам, разделенные значительными расстояниями. Это свидетельствует о высоком качестве современных эмбеддингов для задач тематической классификации технических текстов.

Метод label-as-centroid интерпретируем и прозрачен. Векторные «центроиды» классов оказались расположены внутри кластеров документов соответствующих классов, фактически выступив их прототипами. За счет этого решение классификатора легко объяснимо. Можно явно указать, что, например, отчет наиболее похож на формулировку “Fitness for Duty”, поэтому отнесен к данному типу. Введение метрики margin позволило количественно оценить уверенность модели в каждом прогнозе.

Margin коррелирует с корректностью классификации. Показано, что, если разность между сходством документа со своим классом и с ближайшим чужим положительна (>0), модель почти всегда права (96% случаев). Отрицательный margin наблюдается почти исключительно для ошибочных классификаций. Это позволяет использовать его как индикатор сомнительных исходов. В реальном применении система может маркировать такие отчеты для дополнительной проверки оператором.

Таким образом, наше исследование подтвердило работоспособность zero-shot классификатора на основе LLM-эмбеддингов. Метод label-as-centroid сочетает в себе простоту и гибкость, унаследованные от прототипных методов, и богатство представления, обеспеченное большими языковыми моделями. Он легко масштабируется на новые классы, не требуя трудоемкой разметки, и способен адаптироваться к изменениям таксономий “на лету”. Практическая значимость полученных результатов состоит в том, что они прокладывают путь к созданию интерпретируемых интеллектуальных систем поддержки решений в атомной энергетике, где знания экспертов могут быть интегрированы в модель в виде осмысленных семантических якорей.

СПИСОК ЛИТЕРАТУРЫ

1. *Tixier A.J.-P., Hallowell M.R., Rajagopalan B., Bowman D.* Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports // Automation in Construction. – 2016. – Vol. 62. – P. 45–56. – DOI: 10.1016/j.autcon.2015.11.001.
2. *Zhao Y., Diao X., Huang J., Smidts C.* Automated identification of causal relationships in nuclear power plant event reports // Nuclear Technology. – 2019. – Vol. 205, No. 8. – P. 1021–1034. – DOI: 10.1080/00295450.2019.1580967.
3. *Ricketts J., Barry D., Guo W., Pelham J.* A scoping literature review of natural language processing application to safety occurrence reports // Safety. – 2023. – Vol. 9, No. 2. – Art. 22. – DOI: 10.3390/safety9020022.
4. *Paraskevopoulos G., Pistofidis P., Banoutsos G., Georgiou E., Katsouros V.* Multimodal classification of safety-report observations // Applied Sciences. – 2022. – Vol. 12, No. 12. – Art. 5781. – DOI: 10.3390/app12125781.
5. *New M.D., Wallace R.J.* Classifying aviation safety reports: Using supervised natural language processing (NLP) in an applied context // Safety. – 2025. – Vol. 11, No. 1. – Art. 7. – DOI: 10.3390/safety11010007.
6. *Devlin J., Chang M.-W., Lee K., Toutanova K.* BERT: Pre-training of deep bidirectional transformers for language understanding // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL–HLT 2019), Volume 1 (Long and Short Papers). – Minneapolis, MN, USA: Association for Computational Linguistics, 2019. – P. 4171–4186.
7. *Ramesh G., Sahil M., Palan S.A., Bhandary D., Ashok T.A., Shreyas J., Sowjanya N.* A review on NLP zero-shot and few-shot learning: methods and applications // Discover Applied Sciences. – 2025. – Vol. 7, No. 9. – Art. 966. – DOI: 10.1007/s42452-025-07225-5.
8. *Yin W., Hay J., Roth D.* Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP–IJCNLP). – Hong Kong, China: Association for Computational Linguistics, 2019. – P. 3914–3923.
9. *Родионов, Д.Г.* Трансформация экологической среды социально-экономических систем под воздействием факторов информационной среды / Д. Г. Родионов, Е. А. Короткова, Д. А. Крыжко [и др.] // Экономические науки. – 2021. – № 201. – С. 98-111. – DOI 10.14451/1.201.98. – EDN GDIKCB.

10. *Воронов, А.В.* Опыт использования систем обнаружения свободных и слабозакреплённых предметов в контуре циркуляции теплоносителя реакторных установок Нововоронежской АЭС / А.В. Воронов, М.Т. Слепов // Известия вузов. Ядерная энергетика. – 2022. – № 2. – С. 15–26. – DOI: 10.26583/npe.2022.2.02.
11. *Кацер Ю.Д.* Методы обнаружения неисправностей оборудования АЭС / Ю.Д. Кацер, В.О. Козицин, И.В. Максимов // Известия вузов. Ядерная энергетика. – 2019. – № 4. – С. 5–27. – DOI: 10.26583/npe.2019.4.01.
12. *Rakhimzhanov D., Belginova S., Yedilkhan D.* Automated classification of public transport complaints via text mining using LLMs and embeddings // Information. – 2025. – Vol. 16, No. 8. – Art. 644. – DOI: 10.3390/info16080644.
13. *Nugumanova A., Rakhimzhanov D., Mansurova A.* Global embeddings, local signals: Zero-shot sentiment analysis of transport complaints // Informatics. – 2025. – Vol. 12, No. 3. – Art. 82. – DOI: 10.3390/informatics12030082.
14. *Родионов, Д.Г.* Тематическое моделирование информационной среды медиакомпаний: инструментальный комплекс LDA-TF-IDF / Д. Г. Родионов, Е. А. Конников, П. А. Пашинина, С. И. Шаныгин // Мягкие измерения и вычисления. – 2024. – Т. 76, № 3. – С. 72-84. – DOI 10.36871/2618-9976.2024.03.006. – EDN COCJYG.
15. *Paletto L., Basile V., Esposito R.* Label augmentation for zero-shot hierarchical text classification // Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). – Bangkok, Thailand: Association for Computational Linguistics, 2024. – P. 7697–7706. – DOI: 10.18653/v1/2024.acl-long.416.
16. *Liu H., Zhao S., Zhang X., Zhang F., Wang W., Ma F., Chen H., Yu H., Zhang X.* Liberating seen classes: Boosting few-shot and zero-shot text classification via anchor generation and classification reframing // Proceedings of the AAAI Conference on Artificial Intelligence. – 2024. – Vol. 38, No. 17. – P. 18644–18652. – DOI: 10.1609/aaai.v38i17.29827.
17. *Конников, Е.А.* Совершенствование методов оценки устойчивости развития промышленных предприятий (октант устойчивости развития предприятия) / Е. А. Конников // Маркетинг менеджмент в цифровой экономике. – 2015. – Т. 1, № 4. – С. 4-35. – EDN ZCGXSZ.
18. *Родионов, Д.Г.* Автоматизированный алгоритм системного анализа конкурентоспособности цифрового предприятия в рамках информационной среды / Д. Г. Родионов, Р. М. Мугутдинов, Е. А. Конников // Экономические науки. – 2021. – № 200. – С. 98-108. – DOI 10.14451/1.200.98. – EDN RRFFSY.
19. *Родионов, Д. Г.* Квантификаторы информационной среды финансового рынка / Д. Г. Родионов, П. А. Пашинина, Е. А. Конников // Экономические науки. – 2022. – № 211. – С. 125-128. – DOI 10.14451/1.211.125. – EDN WBAPHW.
20. *Rondinelli A., Bongiovanni L., Basile V.* Zero-shot topic labeling for hazard classification // Information. – 2022. – Vol. 13, No. 10. – Art. 444. – DOI: 10.3390/info13100444.

INTERPRETABLE METHOD OF SEMANTIC PARAMETRISATION OF TAXONOMIES BASED ON TEXTUAL CENTROIDS IN THE SPACE OF MEANING REPRESENTATIONS

© 2026 D.G. Rodionov, E.A. Konnikov, V.A. Leventsov, P.A. Polyakov

Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia

This work explores the application of modern natural language processing methods for automating the classification of incident texts. An interpretable classification approach is proposed, in which each class is represented by a semantic centroid – a vector embedding of the category name obtained using a large language model. The new method does not require a training sample. An unknown report is assigned to the class whose "text centroid" is closest in cosine similarity. Comparisons were made with classical supervised learning algorithms (kNN, Logistic Regression, SVM, Random Forest) on the same embeddings, as well as with the traditional TF-IDF approach. Experiments on a sample of 500 NRC notifications showed that the proposed approach achieves Top-1 accuracy of 62.4% and Top-3 accuracy of 93.4%, while supervised models on the same data approach 100%. A margin metric is introduced – the difference between the cosine similarity with the centroid of the correct class and the next closest centroid. It is shown that the margin serves as a reliable indicator of classification correctness. For correctly classified reports, it is significantly higher than zero, while for errors it takes negative values. Visualisation of embeddings using t-SNE indicates clear clustering of documents by incident type in LLM space. Class centroids are located near the corresponding point clouds, which emphasises the interpretability of the method. Analysis of the semantic distance between classes allows us to identify cases of category overlap. Being deterministic and requiring no training, the method offers a promising solution for the adaptive classification of technical texts when taxonomies change. The results demonstrate that combining LLM embeddings with the proposed centroid strategy allows high accuracy to be achieved without manual data labelling, and the developed margin metric provides a transparent criterion for model confidence.

Keywords: taxonomy, zero-shot classification, large language models, text embeddings, interpretability, nuclear incidents, semantic centroids, margin metrics.

DOI: 10.37313/1990-5378-2026-28-2-168-177
EDN: RGDSEK

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

REFERENCES

1. *Tixier A.J.-P., Hallowell M.R., Rajagopalan B., Bowman D.* Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports // Automation in Construction. – 2016. – Vol. 62. – P. 45–56. – DOI: 10.1016/j.autcon.2015.11.001.
2. *Zhao Y., Diao X., Huang J., Smidts C.* Automated identification of causal relationships in nuclear power plant event reports // Nuclear Technology. – 2019. – Vol. 205, No. 8. – P. 1021–1034. – DOI: 10.1080/00295450.2019.1580967.
3. *Ricketts J., Barry D., Guo W., Pelham J.* A scoping literature review of natural language processing application to safety occurrence reports // Safety. – 2023. – Vol. 9, No. 2. – Art. 22. – DOI: 10.3390/safety9020022.
4. *Paraskevopoulos G., Pistofidis P., Banoutsos G., Georgiou E., Katsouros V.* Multimodal classification of safety-report observations // Applied Sciences. – 2022. – Vol. 12, No. 12. – Art. 5781. – DOI: 10.3390/app12125781.
5. *New M.D., Wallace R.J.* Classifying aviation safety reports: Using supervised natural language processing (NLP) in an applied context // Safety. – 2025. – Vol. 11, No. 1. – Art. 7. – DOI: 10.3390/safety11010007.
6. *Devlin J., Chang M.-W., Lee K., Toutanova K.* BERT: Pre-training of deep bidirectional transformers for language understanding // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL–HLT 2019), Volume 1 (Long and Short Papers). – Minneapolis, MN, USA: Association for Computational Linguistics, 2019. – P. 4171–4186.
7. *Ramesh G., Sahil M., Palan S.A., Bhandary D., Ashok T.A., Shreyas J., Sowjanya N.* A review on NLP zero-shot and few-shot learning: methods and applications // Discover Applied Sciences. – 2025. – Vol. 7, No. 9. – Art. 966. – DOI: 10.1007/s42452-025-07225-5.
8. *Yin W., Hay J., Roth D.* Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP–IJCNLP). – Hong Kong, China: Association for Computational Linguistics, 2019. – P. 3914–3923.
9. *Rodionov, D.G.* Transformaciya ekologicheskoj sredy social'no-ekonomicheskih sistem pod vozdejstviem faktorov informacionnoj sredy / D. G. Rodionov, E. A. Korotkova, D. A. Kryzhko [i dr.] // Ekonomicheskie nauki. – 2021. – № 201. – S. 98-111. – DOI 10.14451/1.201.98. – EDN GDIKCB.
10. *Voronov, A.V.* Opyt ispol'zovaniya sistem obnaruzheniya svobodnyh i slabozakreplyonnyh predmetov v konture cirkulyacii teplonositelya reaktornyh ustanovok Novovoronezhskoj AES / A.V. Voronov, M.T. Slepov // Izvestiya vuzov. Yadernaya energetika. – 2022. – № 2. – S. 15–26. – DOI: 10.26583/npe.2022.2.02.
11. *Kacer Yu.D.* Metody obnaruzheniya neispravnostej oborudovaniya AES / Yu.D. Kacer, V.O. Kozicin, I.V. Maksimov // Izvestiya vuzov. Yadernaya energetika. – 2019. – № 4. – S. 5–27. – DOI: 10.26583/npe.2019.4.01.
12. *Rakhimzhanov D., Belginova S., Yedilkhan D.* Automated classification of public transport complaints via text mining using LLMs and embeddings // Information. – 2025. – Vol. 16, No. 8. – Art. 644. – DOI: 10.3390/info16080644.
13. *Nugumanova A., Rakhimzhanov D., Mansurova A.* Global embeddings, local signals: Zero-shot sentiment analysis of transport complaints // Informatics. – 2025. – Vol. 12, No. 3. – Art. 82. – DOI: 10.3390/informatics12030082.
14. *Rodionov, D.G.* Tematicheskoe modelirovanie informacionnoj sredy mediakompanij: instrumental'nyj kompleks LDA-TF-IDF / D. G. Rodionov, E. A. Konnikov, P. A. Pashinina, S. I. Shanygin // Myagkie izmereniya i vychisleniya. – 2024. – T. 76, № 3. – S. 72-84. – DOI 10.36871/2618-9976.2024.03.006. – EDN COCJYG.
15. *Paletto L., Basile V., Esposito R.* Label augmentation for zero-shot hierarchical text classification // Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). – Bangkok, Thailand: Association for Computational Linguistics, 2024. – P. 7697–7706. – DOI: 10.18653/v1/2024.acl-long.416.
16. *Liu H., Zhao S., Zhang X., Zhang F., Wang W., Ma F., Chen H., Yu H., Zhang X.* Liberating seen classes: Boosting few-shot and zero-shot text classification via anchor generation and classification reframing // Proceedings of the AAAI Conference on Artificial Intelligence. – 2024. – Vol. 38, No. 17. – P. 18644–18652. – DOI: 10.1609/aaai.v38i17.29827.
17. *Konnikov, E.A.* Sovershenstvovanie metodov ocenki ustojchivosti razvitiya promyshlennyh predpriyatij (oktant ustojchivosti razvitiya predpriyatiya) / E. A. Konnikov // Marketing menedzhment v cifrovoj ekonomike. – 2015. – T. 1, № 4. – S. 4-35. – EDN ZCGXSZ.
18. *Rodionov, D.G.* Avtomatizirovannyj algoritm sistemnogo analiza konkurentosposobnosti cifrovogo predpriyatiya v ramkah informacionnoj sredy / D. G. Rodionov, R. M. Mugutdinov, E. A. Konnikov // Ekonomicheskie nauki. – 2021. – № 200. – S. 98-108. – DOI 10.14451/1.200.98. – EDN RRFFSY.
19. *Rodionov, D. G.* Kvantifikatory informacionnoj sredy finansovogo rynka / D. G. Rodionov, P. A. Pashinina, E. A. Konnikov // Ekonomicheskie nauki. – 2022. – № 211. – S. 125-128. – DOI 10.14451/1.211.125. – EDN WBAPHW.
20. *Rondinelli A., Bongiovanni L., Basile V.* Zero-shot topic labeling for hazard classification // Information. – 2022. – Vol. 13, No. 10. – Art. 444. – DOI: 10.3390/info13100444.

Dmitry Rodionov, Doctor of Economics, Director of the Higher School of Engineering and Economics.
E-mail: drodionov@spbstu.ru
Evgeny Konnikov, Ph.D. in Economics, Associate Professor at the Higher School of Engineering and Economics, Head of the Research Laboratory «Polytech-Invest,» and Academic Supervisor of the Master's Program 01.04.0503 "Neurostatistical Technologies in Marketing." E-mail: konnikov_ea@spbstu.ru
Valery Leventsov, PhD in Economics, Director of the Higher School of Advanced Digital Technologies.
E-mail: vleventsov@spbstu.ru
Polyakov Prohor, Laboratory Assistant at the Polytech-Invest Research Laboratory. E-mail: prohor@polyakov-box.ru