УДК 004.822
DOI: 10.22213/2410-9304-2026-1-13-25

Автоматическая категоризация текстовых обращений с использованием дообученных языковых моделей

А. Н. Исенбаев, аспирант, ИжГТУ имени М. Т. Калашникова, Ижевск, Россия
И. М. Янников, доктор технических наук, ИжГТУ имени М. Т. Калашникова, Ижевск, Россия

Службы поддержки различных организаций ежедневно получают сотни и тысячи обращений от пользователей. Ручная сортировка этих заявок занимает значительное время и часто приводит к ошибкам маршрутизации, что снижает скорость и качество обслуживания клиентов. Автоматизация процесса категоризации обращений является актуальной задачей для компаний любого профиля: IT-поддержка, медицинские учреждения, банки, государственные службы, интернет-магазины. В данной работе предложен универсальный метод автоматической сортировки текстовых обращений по категориям с использованием дообученной нейросетевой модели Sentence-BERT (SBERT). Исследована проблема низкой эффективности предобученных языковых моделей при работе с текстами узкоспециализированных предметных областей. Для решения этой проблемы применено контрастивное дообучение модели на предметно-ориентированных данных, что позволило существенно улучшить качество векторных представлений текстов. Проведено систематическое сравнение четырех подходов: базовая модель без дообучения, контрастивное обучение без учителя на неразмеченных данных, дообучение с учителем с использованием критерия Cosine Similarity Loss и дообучение с критерием Multiple Negatives Ranking Loss (MNRL). Эксперименты проведены на наборе из 6500 обращений на русском языке, из которых 1119 были размечены по 16 категориям. Для оценки качества кластеризации использованы как внутренние метрики (Silhouette Score, Davies-Bouldin Index), так и внешние (Purity, NMI, ARI). Лучший результат показал метод MNRL: качество кластеризации по метрике Purity выросло на 123 %, по NMI – на 233 %, по ARI – на 658 % по сравнению с базовой моделью. Предложен механизм оценки уверенности классификации на основе индивидуального Silhouette Score для каждого обращения, позволяющий направлять неуверенные случаи на ручную обработку. Разработанный подход универсален и может быть адаптирован для автоматизации обработки обращений в любой предметной области при наличии 10–20 % размеченных данных.

Ключевые слова: кластеризация текстов, автоматическая категоризация, SBERT, контрастивное обучение, служба поддержки, обработка естественного языка.

Введение

В современном мире службы поддержки являются неотъемлемой частью любой организации, работающей с клиентами или сотрудниками. IT-поддержка, медицинские консультации, банковское обслуживание, государственные услуги, техподдержка интернет-магазинов – все эти системы ежедневно обрабатывают тысячи текстовых обращений. По данным исследований, среднее время первичной обработки одной заявки составляет от 2 до 5 минут, а крупные компании получают от 500 до 10 000 обращений в день [1]. При ручной обработке это требует штата из десятков сотрудников первой линии поддержки. Ручная сортировка заявок имеет ряд существенных недостатков. Во-первых, операторы тратят значительную часть рабочего времени на категоризацию вместо решения проблем. Во-вторых, при большом потоке заявок неизбежны ошибки в категоризации, что приводит к неправильной маршрутизации. В-третьих, разные операторы могут относить похожие заявки к разным категориям, что нарушает единообразие классификации. Наконец, в периоды пиковой нагрузки время ожидания клиентов существенно возрастает. Автоматизация процесса категоризации позволяет решить эти проблемы, что особенно актуально в условиях роста объемов цифрового взаимодействия с клиентами и перехода многих организаций на удаленный формат работы.

Традиционные методы классификации текстов требуют большого объема размеченных данных – обычно от нескольких тысяч до десятков тысяч примеров для каждой категории. На практике такие данные редко доступны: организации либо не ведут историю категоризации, либо используют устаревшую или непоследовательную систему меток. Сбор размеченных данных вручную – трудоемкий и дорогостоящий процесс. Кроме того, стандартные предобученные языковые модели (BERT, GPT и др.) обучены на общих текстах и плохо понимают специфическую терминологию конкретной предметной области: сокращения, названия продуктов, технические термины.

В данной работе исследуется гибридный подход, сочетающий преимущества кластеризации и дообучения с учителем. Предложенный метод включает три ключевых компонента: контрастивное дообучение без учителя (модель обучается на неразмеченных данных, выявляя семантически близкие тексты), дообучение с учителем на небольшом объеме разметки (используется всего ~17 % размеченных данных) и кластеризацию с отказом от классификации (неуверенные предсказания направляются на ручную обработку). Метод универсален и может быть адаптирован для любой предметной области: IT-поддержка, медицина, банковский сектор, государственные услуги, электронная коммерция и др.

Объект исследования – процесс автоматической категоризации текстовых обращений в системах поддержки пользователей.

Предмет исследования – методы дообучения языковых моделей для повышения качества кластеризации предметно-ориентированных текстов.

Цель исследования – экспериментальная оценка эффективности различных методов дообучения языковых моделей для задачи автоматической категоризации текстовых обращений.

Задачи исследования:

1. Сравнить методы векторного представления текста для задачи кластеризации текстовых обращений.

2. Исследовать влияние контрастивного дообучения на качество кластеризации.

3. Оценить эффективность дообучения с учителем при ограниченном объеме размеченных данных.

4. Провести сравнительный анализ методов с использованием внутренних и внешних метрик качества кластеризации.

Гипотезы исследования

– Гипотеза 1: Дообучение языковой модели на предметно-ориентированных данных улучшает качество кластеризации по сравнению с использованием базовой предобученной модели.

– Гипотеза 2: Контрастивное обучение с учителем (на размеченных данных) даёт лучшие результаты, чем контрастивное обучение без учителя.

– Гипотеза 3: Выбор функции потерь влияет на качество дообучения — MNRL (Multiple Negatives Ranking Loss) эффективнее CosineSimilarityLoss для задачи кластеризации.

1. Теоретические основы

Задача автоматической обработки текстов на естественном языке активно развивается в последние десятилетия. Появление сети Интернет и бурный рост доступной текстовой информации значительно ускорили развитие этой научной области [2]. В рамках данного направления предложено множество подходов к кластеризации и классификации текстовых документов, которые можно разделить на классические методы, основанные на статистических признаках, и современные нейросетевые методы.

Классические подходы к кластеризации текстов основаны на представлении документов в виде разреженных векторов TF-IDF (Term Frequency – Inverse Document Frequency) с последующим применением алгоритмов кластеризации [3]. К основным методам относятся: K-Means, иерархическая агломеративная кластеризация, DBSCAN, а также вероятностные модели, такие как латентное размещение Дирихле (LDA). Методы агломеративной иерархической кластеризации используют различные способы измерения расстояния между кластерами (ближнего соседа, дальнего соседа, группового среднего), а алгоритм k-средних имеет множество модификаций [4].

Основной недостаток методов на основе TF-IDF – учет только лексического, а не семантического сходства. Тексты «не работает принтер» и «принтер не печатает» будут иметь низкое сходство при использовании TF-IDF, несмотря на идентичный смысл. Эту проблему частично решают методы дистрибутивной семантики: Word2Vec [5] позволяет получать плотные векторные представления слов, учитывающие контекст их употребления. Однако для получения представления целого предложения или документа требуются дополнительные методы агрегации.

Прорыв в области обработки естественного языка произошел с появлением архитектуры Transformer и модели BERT (Bidirectional Encoder Representations from Transformers) [6]. BERT использует механизм внимания (attention) и обучается на задачах маскированного языкового моделирования, что позволяет получать контекстно-зависимые представления слов. Модель показала лучшие на момент публикации результаты на множестве задач обработки естественного языка, включая классификацию текстов, ответы на вопросы и определение семантического сходства.

Однако прямое использование BERT для задач кластеризации и семантического поиска имеет существенные ограничения. Поиск наиболее похожей пары среди 10 000 предложений требует около 65 часов вычислений с BERT, поскольку необходимо вычислять попарное сходство [7]. Для решения этой проблемы была предложена модель Sentence-BERT (SBERT) – модификация BERT с сиамской архитектурой, которая позволяет получать фиксированные векторные представления предложений. Это сокращает время поиска с 65 часов до нескольких секунд при сохранении качества.

Для русского языка разработано несколько предобученных моделей. RuBERT был получен путем дообучения многоязычной модели BERT на русскоязычных текстах «Википедии» и новостных данных [8]. Модель rubert-tiny2 представляет собой компактную версию (около 45 МБ), полученную методом дистилляции из нескольких больших моделей [9]. Она генерирует векторные представления (эмбеддинги) размерностью 312 компонентов, поддерживает последовательности длиной до 2048 токенов и работает примерно в 10 раз быстрее базовой версии BERT. Эмбеддинги rubert-tiny2 аппроксимируют LaBSE и могут использоваться для задач кластеризации и семантического поиска.

Контрастивное обучение (contrastive learning) – современный подход к обучению представлений, при котором модель учится различать «позитивные» пары (семантически близкие примеры) и «негативные» пары (семантически далёкие примеры). Эффективность данного подхода была продемонстрирована для визуальных представлений в рамках архитектуры SimCLR [10]. Multiple Negatives Ranking Loss (MNRL) – метод обучения, при котором модель учится отличать похожие тексты от непохожих: в процессе обучения модель должна определить, какие пары текстов связаны по смыслу, а какие – нет [11]. Это позволяет эффективно обучаться и показывает значительное улучшение качества представлений.

Для оценки качества кластеризации используются внутренние и внешние метрики. Silhouette Score [12] измеряет, насколько объекты похожи на другие объекты в своем кластере по сравнению с соседними кластерами; значения варьируются от –1 до 1. Normalized Mutual Information (NMI) [13] оценивает взаимную информацию между полученными кластерами и истинными метками, нормализованную для обеспечения сравнимости. Adjusted Rand Index (ARI) [14] измеряет согласованность двух разбиений с поправкой на случайное совпадение.

Ограничением большинства существующих подходов является необходимость большого объема размеченных данных – обычно от нескольких тысяч до десятков тысяч примеров для каждой категории. На практике такие данные редко доступны.

В данной работе исследуется гибридный подход, сочетающий преимущества кластеризации и дообучения с учителем на небольшом объеме размеченных данных. Предложенный метод универсален и может быть адаптирован для любой предметной области.

2. Методология

2.1. Набор данных

Исследование проведено на наборе данных, содержащем 6500 обращений на русском языке. Каждое обращение состоит из заголовка (краткое описание проблемы) и тела (подробное описание). Набор данных сформирован на основе типичных обращений в IT-службу поддержки: часть данных получена из реальной системы службы поддержки (с анонимизацией персональных данных), часть – сгенерирована синтетически для расширения выборки и обеспечения баланса категорий.

Важно отметить, что предложенный метод не зависит от конкретной предметной области. Аналогичным образом можно обработать обращения пациентов в поликлинику, запросы клиентов банка, заявления граждан в государственные органы или любые другие текстовые заявки.

Для оценки качества кластеризации 1119 обращений были размечены по 16 категориям (табл. 1).

Таблица 1. **Категории IT-обращений**

Table 1. **Categories of IT requests**

Категория	Описание	Примеры
Сетевые проблемы	Проблемы с интернетом, локальной сетью	«Нет доступа к сети», «Медленный интернет»
Принтеры	Проблемы с печатью	«Принтер не печатает», «Замятие бумаги»
Электронная почта	Проблемы с почтой	«Не приходят письма», «Ошибка Outlook»
Программное обеспечение	Установка, обновление ПО	«Установить Office», «Программа зависает»
Оборудование	Неисправности hardware	«Не включается компьютер», «Сломался монитор»
Учётные записи	Пароли, доступы	«Забыл пароль», «Заблокирован аккаунт»
Резервное копирование	Бэкапы, восстановление	«Восстановить файлы», «Настроить бэкап»

2.2. Архитектура системы

Разработанная система автоматической категоризации обращений реализована в виде веб-сервиса и состоит из нескольких функциональных модулей (рис. 1).

Модуль препроцессинга выполняет предварительную обработку входного текста: приведение к нижнему регистру, удаление лишних пробелов и специальных символов, нормализацию пунктуации. На вход модуль получает текстовое обращение, состоящее из заголовка (краткое описание проблемы) и тела (подробное описание). Заголовок и тело объединяются в единую строку для последующей обработки.

Модуль генерации эмбеддингов преобразует очищенный текст в векторное представление фиксированной размерности. В качестве базовой модели используется rubert-tiny2 – компактная русскоязычная модель на основе архитектуры BERT [4]. Модель генерирует векторы размерностью 312 компонентов. Для повышения качества представлений модель дообучается на предметно-ориентированных данных с использованием метода MNRL (Multiple Negatives Ranking Loss) [12].

Модуль кластеризации группирует обращения на основе косинусного сходства их векторных представлений. Используется алгоритм агломеративной иерархической кластеризации с методом связи average [9]. Число кластеров задается как гиперпараметр (в экспериментах – 16, по числу категорий).

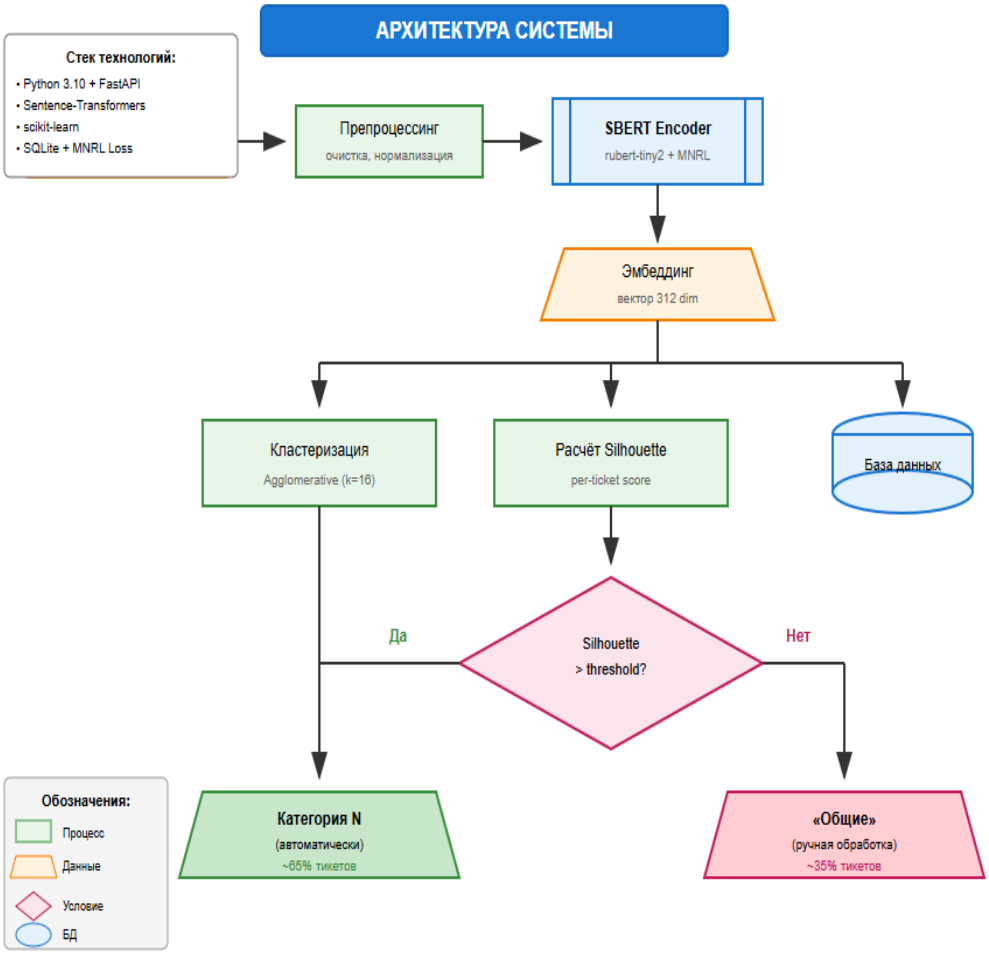

Рис. 1. Общая схема работы системы

Fig. 1. General diagram of the system operation

Модуль оценки уверенности вычисляет индивидуальный Silhouette Score для каждого обращения [13]. Эта метрика показывает, насколько обращение «хорошо» принадлежит своему кластеру по сравнению с соседними кластерами. Значения варьируются от –1 (обращение ошибочно отнесено к кластеру) до 1 (обращение находится в центре плотного кластера). На основе порогового значения (threshold) система принимает решение: при высокой уверенности обращение автоматически направляется в соответствующую категорию, при низкой – передается на ручную обработку в категорию «Общие».

Модуль хранения данных обеспечивает персистентность: сохранение обращений, их эмбеддингов, результатов кластеризации и ручных меток. Реализован на основе реляционной СУБД SQLite.

REST API предоставляет программный интерфейс для взаимодействия с системой: добавление новых обращений, запуск кластеризации, получение результатов, управление метками. Реализован на основе фреймворка FastAPI.

2.3. Схема эксперимента

Для проверки гипотез проведено последовательное сравнение четырех методов представления текста. Каждый последующий этап основывается на результатах предыдущего, что позволяет оценить вклад каждого компонента в итоговое качество кластеризации.

Этап 1. Базовая модель. На первом этапе используется предобученная модель rubert-tiny2 без какой-либо дополнительной настройки. Модель генерирует 312-мерные эмбеддинги для всех 6500 обращений, после чего выполняется кластеризация и вычис-

ляются метрики качества. Этот этап служит точкой отсчета для оценки эффективности дообучения.

Этап 2. Контрастивное дообучение без учителя. На втором этапе модель дообучается методом контрастивного обучения без использования ручной разметки [11, 19]. Для формирования обучающих пар применяется следующий подход: для каждого обращения с помощью TF-IDF-векторов находятся наиболее похожие тексты из корпуса, которые используются как позитивные примеры. Модель обучается сближать представления семантически близких текстов и отдалять представления непохожих. После дообучения выполняется повторная кластеризация и оценка метрик.

Этап 3. Дообучение с учителем (CosineSimilarityLoss). На третьем этапе модель дообучается на вручную размеченных данных (1119 обращений с категориями). Используется критерий обучения Cosine Similarity Loss, который минимизирует косинусное расстояние между эмбеддингами текстов одной категории и максимизирует расстояние между текстами разных категорий. Обучающие пары формируются из обращений с одинаковыми метками (позитивные пары) и разными метками (негативные пары).

Этап 4. Дообучение с учителем (MNRL). На четвертом этапе модель дообучается с использованием критерия Multiple Negatives Ranking Loss (MNRL) [12]. В отличие от Cosine Similarity Loss, MNRL оптимизирует относительное ранжирование: для каждого «якорного» текста модель должна поставить позитивный пример выше всех негативных примеров в батче (in-batch negatives). Это более эффективный подход, поскольку каждый элемент батча служит негативным примером для остальных, что увеличивает число обучающих сигналов без дополнительных вычислений.

После каждого этапа все 6500 обращений кластеризуются заново и вычисляются как внутренние (Silhouette, Davies-Bouldin), так и внешние (Purity, NMI, ARI) метрики качества на размеченном подмножестве. Схема эксперимента представлена на рис. 2.

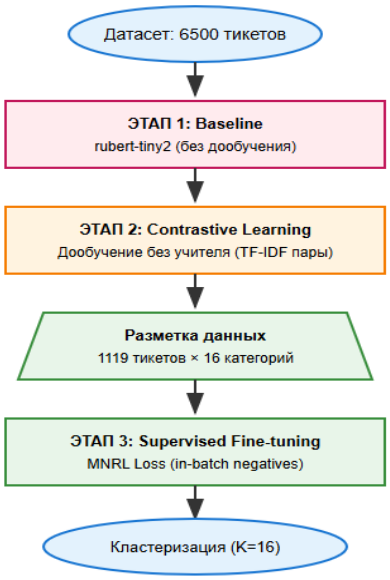

Рис. 2. Схема эксперимента

Fig. 2. Experimental scheme

2.4. Алгоритм кластеризации

Для группировки обращений по категориям использован алгоритм агломеративной иерархической кластеризации [9]. Данный алгоритм относится к классу восходящих иерархических методов и работает следующим образом.

Принцип работы алгоритма. На начальном этапе каждое обращение рассматривается как отдельный кластер. Затем на каждой итерации алгоритм объединяет два наиболее близких кластера в один, формируя иерархическую структуру (дендрограмму). Процесс продолжается до тех пор, пока не будет достигнуто заданное число кластеров K. Метрика расстояния. Для измерения близости между векторными представлениями обращений использована косинусная метрика:

$$d(a, b) = 1 - \frac{a \cdot b}{\|a\| \cdot \|b\|}, \quad (1)$$

где a и b – эмбеддинги двух обращений размерностью 312; $a \cdot b$ – скалярное произведение векторов; $\|a\|$ и $\|b\|$ – евклидовы нормы векторов. Метод связи (linkage). Для определения расстояния между кластерами использован метод среднего связывания (average linkage), при котором расстояние между кластерами вычисляется как среднее расстояние между всеми парами объектов из разных кластеров:

$$D(C_i, C_j) = 1,$$
$$|C_i| \cdot |C_j| \sum_{a \in Ci} \sum_{b \in Cj} d(a, b), \quad (2)$$

где C_i и C_j – объединяемые кластеры; $|C_i|$ и $|C_j|$ – число элементов в кластерах; $d(a, b)$ – расстояние между объектами по формуле (1).

Выбор числа кластеров. Число кластеров K установлено равным 16 – по числу категорий в системе ручной разметки (см. табл. 1). Такой выбор позволяет напрямую сравнивать полученные кластеры с реальными категориями обращений и вычислять внешние метрики качества.

Обоснование выбора алгоритма. Агломеративная кластеризация была выбрана по следующим причинам: (1) алгоритм детерминирован и не зависит от начальной инициализации, в отличие от K-Means; (2) поддерживает произвольные метрики расстояния, включая косинусную; (3) хорошо работает с данными произвольной формы и не требует предположений о сферической структуре кластеров.

Блок-схема алгоритма категоризации обращений представлена на рис. 3.

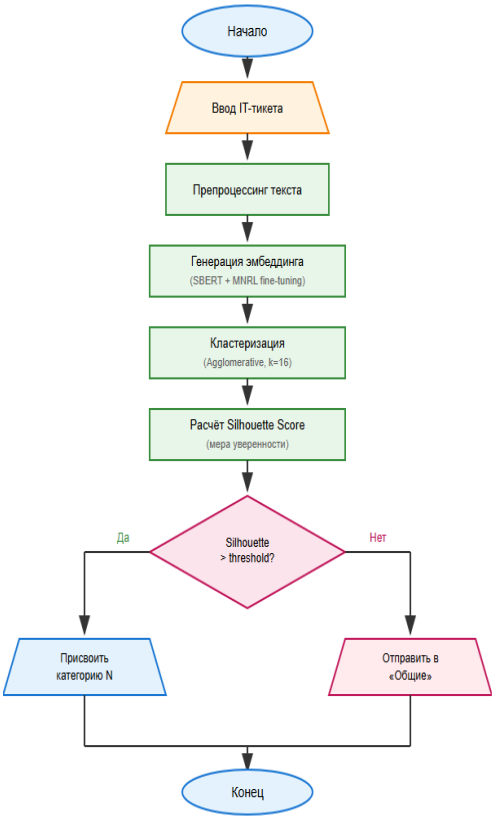

Рис. 3. Блок-схема алгоритма кластеризации

Fig. 3. Flowchart of the clustering algorithm

2.5. Метрики качества

Для оценки качества кластеризации использованы два типа метрик:

Внутренние метрики (не требуют разметки) – измеряют геометрические свойства кластеров:

– Silhouette Score [13] – мера того, насколько объекты похожи на свой кластер по сравнению с другими кластерами. Значения от -1 до 1, где 1 – идеальная кластеризация.

– Davies-Bouldin Index [14] – отношение внутрикластерного разброса к межкластерному расстоянию. Меньшие значения лучше.

Внешние метрики (требуют разметки) – измеряют соответствие кластеров реальным категориям:

– Purity – доля объектов, отнесённых к доминирующей категории в каждом кластере [9].

– Normalized Mutual Information (NMI) [16] – нормализованная взаимная информация между кластерами и истинными метками.

– Adjusted Rand Index (ARI) [15] – скорректированный индекс Рэнда, учитывающий случайное совпадение меток.

Для задачи категоризации IT-обращений приоритетными являются **внешние метрики** (Purity, NMI, ARI), поскольку цель – соответствие кластеров реальным категориям (например, «Принтеры», «Сеть»), а не абстрактное геометрическое качество кластеров.

Silhouette Score измеряет компактность и разделимость кластеров, но высокий Silhouette не гарантирует соответствия реальным категориям. Модель может создавать плотные кластеры, которые не совпадают с практически значимыми категориями. Поэтому в данной работе Silhouette используется как вспомогательная метрика, а основные выводы делаются на основе внешних метрик.

3. Эксперименты и результаты

3.1. Сравнение методов представления текста

Эксперименты проведены на полном наборе данных из 6500 обращений. Для каждого метода представления текста выполнена кластеризация с числом кластеров K=16 и вычислены метрики качества на размеченном подмножестве (1119 обращений). Результаты представлены в табл. 2.

Таблица 2. **Сравнение методов представления текста**

Table 2. **Comparison of text presentation methods**

Метод	Silhouette	Davies-Bouldin	Purity	NMI	ARI
rubert-tiny2 (baseline)	0,029	2,23	0,181	0,151	0,021
+Контрастивное (unsupervised)	**0,265**	**1,44**	0,325	0,348	0,064
+ CosineSimilarityLoss	0,053	2,76	0,263	0,245	0,046
+ MNRL	0,153	1,69	**0,404**	**0,501**	**0,162**

Анализ результатов базовой модели. Предобученная модель rubert-tiny2 без дообучения показывает неудовлетворительные результаты по всем метрикам. Значение Silhouette Score = 0,029 свидетельствует о практически случайном распределении объектов по кластерам — кластеры слабо разделены и перекрываются.

Высокое значение Davies-Bouldin Index (2.23) подтверждает низкое качество кластеризации. Внешние метрики также крайне низкие: Purity = 0,181 означает, что лишь 18 % обращений в каждом кластере относятся к доминирующей категории. Значение NMI = 0,151 и ARI = 0,021 указывают на слабую корреляцию между полученными кластерами и реальными категориями.

Эффект контрастивного дообучения без учителя. После контрастивного дообучения на неразмеченных данных наблюдается значительное улучшение всех метрик. Silhouette Score вырос с 0,029 до 0,265 (улучшение в 9 раз), что свидетельствует о формировании более компактных и разделенных кластеров. Davies-Bouldin Index снизился с 2,23 до 1,44, что также указывает на улучшение геометрического качества кластеров. Purity выросла с 0,181 до 0,325 (+79 %), NMI – с 0,151 до 0,348 (+131 %), ARI – с 0,021 до 0,064 (+200 %).

Эти результаты демонстрируют, что даже без использования ручной разметки контрастивное обучение позволяет существенно улучшить качество представлений для задачи кластеризации.

Негативный эффект Cosine Similarity Loss. Дообучение с критерием Cosine Similarity Loss показало результаты хуже, чем контрастивное обучение без учителя. Purity снизилась с 0,325 до 0,263, NMI – с 0,348 до 0,245, ARI – с 0,064 до 0,046. Это объясняется особенностью критерия: Cosine Similarity Loss оптимизирует абсолютное косинусное сходство между парами текстов, что может приводить к «схлопыванию» пространства представлений – все векторы стягиваются к небольшой области, теряя различительную способность [8].

Превосходство метода MNRL. Дообучение с критерием Multiple Negatives Ranking Loss (MNRL) показало лучшие результаты по всем внешним метрикам: Purity = 0,404, NMI = 0,501, ARI = 0,162. В отличие от Cosine Similarity Loss, MNRL оптимизирует относительное ранжирование: модель учится отличать позитивные пары от негативных, используя все остальные элементы батча как негативные примеры (in-batch negatives). Это более эффективный подход, поскольку каждый пример батча участвует в множестве сравнений, что увеличивает количество обучающих сигналов без дополнительных вычислительных затрат.

Интерпретация достигнутых результатов. Значение Purity = 0,404 означает, что в среднем 40 % обращений в каждом кластере относятся к одной (доминирующей) категории. Для задачи автоматической маршрутизации это позволяет с высокой вероятностью определить тематику кластера. NMI = 0,501 свидетельствует о значительной корреляции между полученными кластерами и реальными категориями – модель успешно «выучила» семантическую структуру дан-

ных. ARI = 0,162, хотя и остается относительно низким, показывает улучшение в 7,7 раза по сравнению с базовой моделью.

3.2. Анализ результатов и проверка гипотез

Проверка гипотезы 1 (дообучение улучшает качество): базовая модель rubert-tiny2 показывает низкие значения всех метрик (Purity = 0.181, NMI = 0.151, ARI = 0.021). После контрастивного дообучения без учителя все метрики значительно улучшились: Purity выросла на 79 %, NMI – на 131 %, ARI – на 200 %.

Вывод: гипотеза 1 подтверждена. Дообучение на предметно-ориентированных данных существенно улучшает качество кластеризации.

Проверка гипотезы 2 (supervised лучше unsupervised): сравнение контрастивного дообучения без учителя и дообучения с учителем (MNRL). Результаты сравнения показаны в табл. 3.

Таблица 3. **Сравнение дообучения с учителем и без учителя (гипотеза 2)**

Table 3. **Comparison of supervised and unsupervised retraining (hypothesis 2)**

Метрика	Без учителя	С учителем (MNRL)	Улучшение
Purity	0,325	0,404	+24 %
NMI	0,348	0,501	+44 %
ARI	0,064	0,162	+153 %

Вывод: гипотеза 2 подтверждена. Дообучение с учителем на 1119 размеченных обращениях (~17 % от общего объема) дает значительное улучшение по всем внешним метрикам.

Проверка гипотезы 3 (MNRL лучше CosineSimilarityLoss): дообучение с Cosine Similarity Loss показало результаты хуже, чем контрастивное обучение без учителя (табл. 4).

Таблица 4. **Результаты проверки гипотезы 3**

Table 4. **Results of testing hypothesis 3**

Метрика	CosineSimilarity	MNRL	Разница
Purity	0,263	0,404	+54 %
NMI	0,245	0,501	+104 %
ARI	0,046	0,162	+252 %

Причина такого различия кроется в механизме работы функций потерь. При использовании Cosine Similarity Loss модель стремится максимально сблизить похожие примеры в пространстве представлений. Это приводит к нежелательному эффекту: векторы «схлопываются» в небольшую область, теряя различительную способность [8]. MNRL работает иначе: вместо абсолютного сближения пар он учит модель ранжировать – выбирать правильный пример среди множества негативных [12]. Такой подход сохраняет структуру пространства и лучше подходит для последующей кластеризации.

Вывод: гипотеза 3 подтвердилась. Более того, разница между методами оказалась даже больше ожидаемой: MNRL превосходит Cosine Similarity Loss более чем в два раза по ключевым метрикам.

Таблица 5. **Проверка гипотез исследования**

Table 5. **Testing the research hypotheses**

Гипотеза	Описание	Результат	Доказательство
1	Дообучение улучшает качество кластеризации	Подтверждена	Purity: 0,181 → 0,404 (+123 %)
2	Supervised лучше unsupervised	Подтверждена	NMI: 0,348 → 0,501 (+44%)
3	MNRL лучше CosineSimilarityLoss	Подтверждена	ARI: 0,046 → 0,162 (+252 %)

3.3. Визуализация результатов

Результаты экспериментов наглядно демонстрируют преимущества предложенного подхода. На рис. 4 представлено сравнение трех методов по внешним метрикам качества кластеризации. Базовая модель rubert-tiny2 показывает результаты, близкие к случайному распределению: при 16 кластерах и равномерном распределении по категориям ожидаемое значение Purity составило бы 1/16 ≈ 0,063, фактическое значение 0,181 лишь незначительно выше.

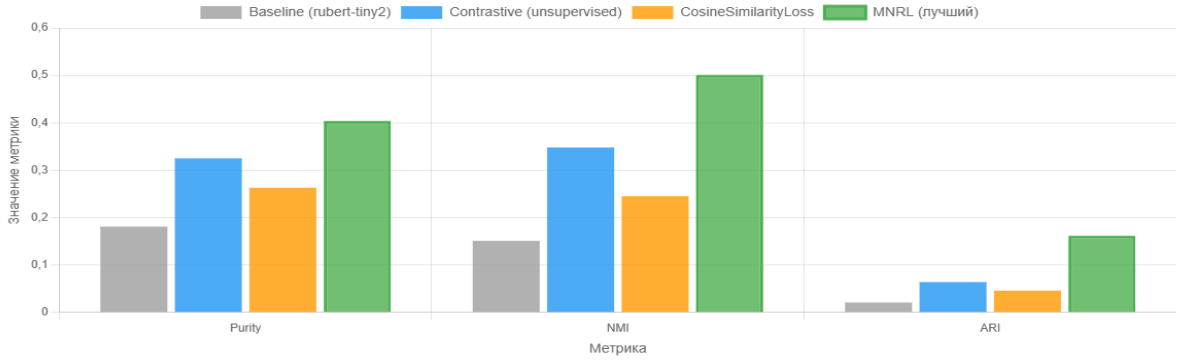

Рис. 4. Сравнение методов представления текста по внешним метрикам

Fig. 4. Comparison of text presentation methods based on external metrics

После контрастивного дообучения без учителя наблюдается существенный рост всех метрик. Модель научилась выделять семантически близкие тексты, хотя и не знала, какие именно категории существуют. Это подтверждает эффективность контрастивного подхода для формирования качественных текстовых представлений даже без разметки.

Наибольший прирост достигнут при дообучении с MNRL: метрика NMI впервые превысила отметку 0,5, что свидетельствует о высокой корреляции между автоматическими кластерами и экспертными категориями. Фактически, модель научилась «понимать» предметную область IT-поддержки и группировать обращения в соответствии с их реальной тематикой.

На рис. 5 показано относительное улучшение метрик по сравнению с базовой моделью. Особенно заметен прирост по метрике ARI: если контрастивное дообучение увеличило её в 3 раза, то MNRL — почти в 8 раз. ARI является наиболее строгой метрикой из рассмотренных, поскольку учитывает не только попадание в правильный кластер, но и корректность попарных отношений между объектами. Столь значительное улучшение говорит о том, что MNRL-модель действительно научилась различать категории обращений, а не просто сформировала произвольные группы.

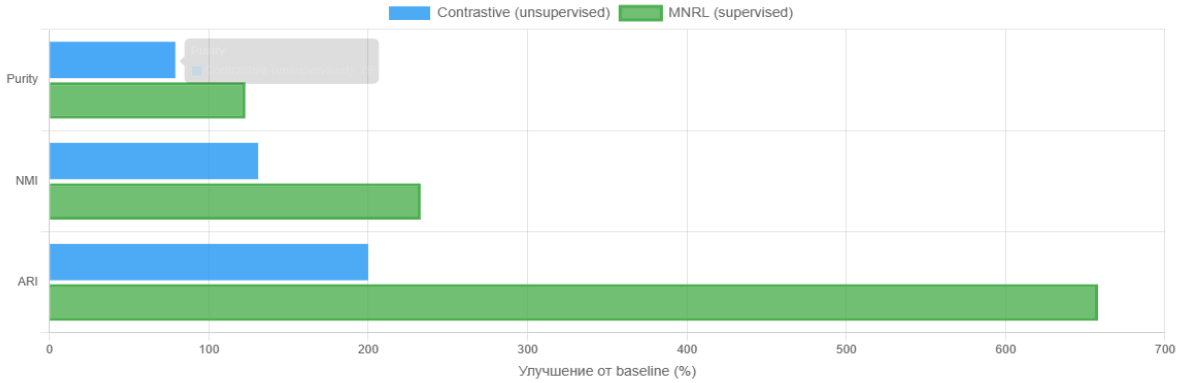

Рис. 5. Относительное улучшение метрик от baseline (%)

Fig. 5. Relative improvement of metrics from baseline (%)

Стоит отметить практическую сторону полученных результатов. При Purity = 0,404 система может присвоить каждому кластеру метку его доминирующей категории, и примерно 40 % обращений окажутся классифицированы правильно без какого-либо участия человека. Оставшиеся 60 % – это либо обращения из менее представленных категорий внутри кластера, либо пограничные случаи, которые сложно отнести к одной категории однозначно. Для таких случаев система может использовать механизм rejection, направляя их на ручную обработку.

Заключение

Все три сформулированные гипотезы исследования получили экспериментальное подтверждение. Базовая модель rubert-tiny2 показала неудовлетворительные результаты (Purity = 0,181), тогда как после дообучения с MNRL метрика Purity достигла 0,404, что соответствует улучшению на 123 %. Общие предобученные модели не способны эффективно кластеризовать тексты узкой предметной области без дополнительной адаптации. При использовании всего 1119 размеченных обращений (около 17 % от общего объема данных) удалось добиться значительного улучшения по всем внешним метрикам – Purity выросла на 24 %, NMI – на 44 %, ARI – на 153 % по сравнению с контрастивным дообучением без учителя, что демонстрирует высокую эффективность supervised-подхода даже при ограниченном объеме разметки.

Особый интерес представляет подтверждение третьей гипотезы. Критерий Cosine Similarity Loss показал результаты хуже, чем обучение без учителя: причина в эффекте «схлопывания» пространства представлений. Критерий MNRL, напротив, оптимизирует относительное ранжирование через механизм in-batch negatives, что сохраняет структуру пространства и обеспечивает превосходство по всем метрикам: Purity выше на 54 %, NMI – на 104 %, ARI – на 252 %. Выбор критерия обучения оказывает критическое влияние на результат, и не всякий supervised-метод гарантирует улучшение по сравнению с unsupervised-подходом.

Разработанный метод может быть применен для автоматизации обработки текстовых обращений в любой предметной области. Для адаптации к новой предметной области достаточно собрать набор обращений, разметить 10–20 % данных по категориям и выполнить дообучение модели с критерием MNRL. При достигнутом значении Purity = 0,404 система способна автоматически определять тематику кластеров и обрабатывать обращения с высокой уверенностью. Анализ распределения Silhouette Score показал, что около 14 % обращений могут быть классифицированы с высокой уверенностью, что позволяет существенно сократить нагрузку на операторов первой линии поддержки.

Направления дальнейших исследований включают апробацию метода на данных из других предметных областей, исследование альтернативных алгоритмов кластеризации (HDBSCAN, Spectral Clustering), оценку статистической значимости результатов через множественные запуски эксперимента, а также сравнение с методами полностью supervised-классификации для определения оптимального баланса между объемом разметки и качеством категоризации.

Библиографические ссылки

1. *Мансур А. М.* Алгоритм на основе трансформеров для классификации длинных текстов // Известия ЮФУ. Технические науки. 2024. № 3 (239). С. 187–196.

2. *Воронцов К. В.* Машинное обучение: курс лекций. Московский физико-технический институт, 2024. URL: http://www.machinelearning.ru/wiki.

3. *Куратов Ю., Архипов М.* Адаптация глубоких двунаправленных многоязычных трансформеров для русского языка // Вычислительная лингвистика и интеллектуальные технологии : труды Международной конференции «Диалог 2019». 2019. С. 333–340.

4. *Колесникова А.* Rubert-tiny2: компактная русскоязычная модель BERT // Hugging Face Model Hub. 2022. URL: https://huggingface.co/cointegrated/rubert-tiny2.

5. *Гареев Р. М., Майоров В. Д.* Автоматическая классификация обращений в техническую поддержку на основе методов машинного обучения // Информационные процессы и математическое моделирование : труды конференции ИПМТ-2022. Уфа, 2022. С. 112–118.

6. Решения Cleverics для автоматизации Service Desk: AID+ – система интеллектуальной маршрутизации // Официальный сайт Cleverics. 2023. URL: https://cleverics.ru.

7. *Ивахин Д. Е., Андиева Е. Ю.* Автоматический анализ текста для выявления профессиональных навыков: гибридный подход на основе TF-IDF и нейросетевых эмбеддингов // Вестник науки. 2025. № 4 (85).

8. *Давлетов А. Р.* Современные методы машинного обучения и технология OCR для автоматизации обработки документов // Вестник науки. 2023. № 10 (67). С. 676–698.

9. *Рави Дж., Кулкарни С.* Методы встраивания текста для эффективной кластеризации данных из Твиттера // Эволюционный интеллект. 2023. Т. 7.

10. Ли Ч., Чжан С., Чжан И., Лонг Д., Се П., Чжан М. К созданию общих текстовых вложений с помощью многоэтапного контрастивного обучения // Препринт arXiv:2308.03281. 2023.

11. Простая структура для контрастивного обучения визуальных представлений / Т. Чен, С. Корнблит, М. Норузи, Г. Хинтон // Труды 37-й Международной конференции по машинному обучению (ICML). PMLR, 2020.

12. *Гао Т., Яо С., Чен Д.* SimCSE: простое контрастивное обучение вложению предложений // Труды EMNLP. 2021.

13. *Руссеу П. Дж.* Силуэты: графическое средство для интерпретации и проверки кластерного анализа // Журнал вычислительной и прикладной математики. 1987. Т. 20. С. 53–65.

14. *Дэвис Д. Л., Боулдин Д. В.* Мера разделения кластеров // Труды IEEE по анализу образов и машинному интеллекту. 1979. Т. PAMI-1, № 2. С. 224–227.

15. *Хуберт Л., Араби П.* Сравнение разбиений // Журнал классификации. 1985. Т. 2. С. 193–218.

16. *Штрель А., Гош Дж.* Кластерные ансамбли – структура повторного использования знаний для объединения нескольких разделов // Журнал исследований в области машинного обучения. 2002. Т. 3. С. 583–617.

17. Автоматизация обработки заявок: взгляд на современные исследования с применением к сценариям многоуровневой классификации / Ф. Коккорас и др. // Экспертные системы с приложениями. 2023.

18. *Длодло Н., Сибанда К.* Подход машинного обучения к автоматической категоризации запросов на ИТ-услуги // Труды Южноафриканской конференции по телекоммуникационным сетям и приложениям. 2020. С. 1–6.

19. *Гао Т., Яо С., Чен Д.* SimCSE: простое контрастное обучение вложению предложений // Труды EMNLP. 2021.

20. Внимание – это все, что вам нужно / А. Васвани, Н. Шазир, Н. Пармар, Дж. Ушкорейт, Л.Джонс, А. Н. Гомес, Л. Кайзер, И. Полосухин // Достижения в области нейронных систем обработки информации. 2017.

References

1. Mansur A.M. [Algorithm based on transformers for classification of long texts]. Bulletin of SFedU. Technical sciences. 2024. No. 3. Pp. 187-196 (in Russ.).

2. Vorontsov K.V. [Machine learning: a course of lectures]. Moscow Institute of Physics and Technology, 2024. Available at: http:// www.machinelearning.ru/wiki (in Russ.).

3. Kuratov Y., Arkhipov M. [Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language]. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2019". 2019. Pp. 333-340 (in Russ.).

4. Kolesnikova A. [rubert-tiny2: компактная русскоязычная модель BERT]. Hugging Face Model Hub. 2022. Available at: https://huggingface.co/ cointegrated/rubert-tiny2 (in Russ.).

5. Gareev R.M., Mayorov V.D. [Automatic classification of technical support requests based on machine learning methods]. [Information processes and mathematical modeling: proceedings of the IPMT-2022 conference]. Ufa, 2022. P. 112-118 (in Russ.).

6. [Cleverics Service Desk Automation Solutions: AID+ — Intelligent Routing System]. Official Cleverics website. 2023. Available at: https://cleverics.ru (in Russ.).

7. Ivakhin D.E., Andieva E.Yu. [Automatic text analysis to identify professional skills: a hybrid approach based on TF-IDF and neural network embeddings]. Science Bulletin. 2025. No. 4 (in Russ.).

8. Davletov A.R. [Modern methods of machine learning and OCR technology for automation of document processing]. Science Bulletin. 2023. No. 10). Pp. 676-698 (in Russ.).

9. Ravi J., Kulkarni S. [Text embedding techniques for efficient clustering of twitter data]. Evolutionary Intelligence. 2023. Vol. 7 (in Russ.).

10. Li Z., Zhang X., Zhang Y., Long D., Xie P., Zhang M. [Towards General Text Embeddings with Multi-stage Contrastive Learning]. arXiv preprint arXiv:2308.03281. 2023 (in Russ.).

11. Chen T., Kornblith S., Norouzi M., Hinton G. A [Simple Framework for Contrastive Learning of Visual Representations]. [Proceedings of the 37th International Conference on Machine Learning (ICML)]. PMLR, 2020 (in Russ.).

12. Gao T., Yao X., Chen D. [SimCSE: Simple Contrastive Learning of Sentence Embeddings]. [Proceedings of EMNLP]. 2021 (in Russ.).

13. Rousseeuw P.J. [Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis]. Journal of Computational and Applied Mathematics. 1987. Vol. 20. Pp. 53-65 (in Russ.).

14. Davies D.L., Bouldin D.W. A [Cluster Separation Measure]. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1979. Vol. PAMI-1, No. 2. Pp. 224–227 (in Russ.).

15. Hubert L., Arabie P. [Comparing Partitions]. Journal of Classification. 1985. Vol. 2. Pp. 193-218 (in Russ.).

16. Strehl A., Ghosh J. [Cluster Ensembles - A Knowledge Reuse Framework for Combining Multiple Partitions]. Journal of Machine Learning Research. 2002. Vol. 3. Pp. 583-617 (in Russ.).

17. Kokkoras F. et al. [Ticket automation: An insight into current research with applications to multi-level classification scenarios]. Expert Systems with Applications. 2023 (in Russ.).

18. Dlodlo N., Sibanda K. A [Machine Learning Approach to Automatic IT Service Request Categorization]. [Proceedings of the Southern Africa Telecommunication Networks and Applications Conference]. 2020. Pp. 1-6 (in Russ.).

19. Gao T., Yao X., Chen D. [SimCSE: Simple Contrastive Learning of Sentence Embeddings]. [Proceedings of EMNLP] (in Russ.).

20. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł., Polosukhin I. [Attention Is All You Need]. [Advances in Neural Information Processing Systems]. 2017 (in Russ.).

* * *

Automatic Categorization of Textual Messages Using Pre-Trained Language Models

A. N. Isenbaev, graduate student, Izhevsk State Technical University named after M.T. Kalashnikov, Izhevsk, Russia

I. M. Yannikov, Doctor of Engineering Sciences, Professor, Department of Technosphere Safety Izhevsk State Technical University named after M.T. Kalashnikov, Izhevsk, Russia

Help desks at various organizations receive hundreds and thousands of requests from users daily. Manually sorting these requests takes considerable time and often leads to routing errors, reducing the speed and quality of customer service. Automating the request categorization process is a pressing issue for companies of all types, including IT support, medical institutions, banks, government agencies, and online stores. This paper proposes a universal method for automatically sorting text requests into categories using a pre-trained Sentence-BERT (SBERT) neural network model. The low efficiency of pre-trained language models when working with texts from highly specialized subject areas is investigated. To address this issue, contrastive retraining of the model on domain-specific data was applied, significantly improving the quality of vector text representations. A systematic comparison of four approaches was conducted: a baseline model without retraining, unsupervised contrastive learning on unlabeled data, supervised retraining using the CosineSimilarityLoss criterion, and retraining using the Multiple Negatives Ranking Loss (MNRL) criterion. Experiments were conducted on a dataset of 6,500 Russian-language queries, of which 1,119 were labeled into 16 categories. Both internal metrics (Silhouette Score, Davies-Bouldin Index) and external ones (Purity, NMI, ARI) were used to assess clustering quality. The MNRL method demonstrated the best results: clustering quality increased by 123% for Purity, 233% for NMI, and 658% for ARI compared to the baseline model. A mechanism for assessing classification confidence based on an individual Silhouette Score for each query is proposed, allowing uncertain cases to be redirected for manual processing. The developed approach is universal and can be adapted to automate the processing of requests in any subject area with 10–20% of labeled data.

Keywords: text clustering, automatic categorization, SBERT, contrastive learning, help desk, natural language processing.

Получено: 26.01.26

Образец цитирования

Исенбаев А. Н., *Янников И. М.* Автоматическая категоризация текстовых обращений с использованием дообученных языковых моделей // Интеллектуальные системы в производстве. 2026. Т. 24, № 1. С. 13–25. DOI: 10.22213/2410-9304-2026-1-13-25.

For Citation

Isenbaev A.N., Yannikov I.M. [Automatic Categorization of Textual Messages Using Pre-Trained Language Models]. *Intellektual'nye sistemy v proizvodstve.* 2026, vol. 24, no. 4, pp. 13-25 (in Russ.). DOI: 10.22213/2410-9304-2026-1-13-25.