УДК 004.912
DOI: 10.26102/2310-6018/2025.50.3.010

Алгоритмы кластеризации неструктурированных текстов и их реализация в программных системах

В.С. Кондаков✉, А.В. Кузнецова

Южно-Российский государственный политехнический университет (НПИ) имени М.И. Платова, Новочеркасск, Российская Федерация

Резюме. Актуальность исследования обусловлена стремительным ростом объема неструктурированных текстов в цифровой среде и необходимостью их систематического анализа. Отсутствие универсальных и легко воспроизводимых решений по группировке текстовой информации затрудняет ее интерпретацию и ограничивает возможности применения в различных прикладных сферах, включая здравоохранение, образование, маркетинг и корпоративный сектор. В связи с этим данная статья направлена на выявление ключевых алгоритмических подходов к кластеризации неструктурированных текстов, а также на анализ программных систем, реализующих соответствующие методы. Ведущий метод исследования основан на сравнительно-аналитическом подходе, позволившем обобщить и классифицировать современные алгоритмы машинного обучения, применяемые для обработки текстовых данных. В работе рассмотрены как традиционные методы кластеризации, так и современные архитектуры, использующие обучение без учителя, числовые векторные представления и нейросетевые модели. Проанализированы программные инструменты, демонстрирующие различные уровни точности, интерпретируемости и адаптивности. В результате систематизированы критерии выбора методов под конкретные задачи, выявлены ограничения существующих подходов и обозначены перспективные направления развития. Материалы статьи могут быть полезны специалистам, занимающимся проектированием и внедрением программных решений для автоматической обработки и анализа текстовой информации.

Ключевые слова: кластеризация текстов, неструктурированные данные, тематическое моделирование, машинное обучение, векторные представления, алгоритмы без учителя, программные фреймворки, интеллектуальный анализ текста.

Для цитирования: Кондаков В.С., Кузнецова А.В. Алгоритмы кластеризации неструктурированных текстов и их реализация в программных системах. *Моделирование, оптимизация и информационные технологии.* 2025;13(3). URL: https://moitvivt.ru/ru/journal/pdf?id=1970 DOI: 10.26102/2310-6018/2025.50.3.010

Clustering algorithms for unstructured texts and their implementation in software systems

V.S. Kondakov✉, A.V. Kuznetsova

Platov South-Russian State Polytechnic University (NPI), Novocherkassk, the Russian Federation

Abstract. The relevance of this study is driven by the rapid growth of unstructured textual data in the digital environment and the pressing need for its systematic analysis. The lack of universal and easily reproducible methods for grouping textual information complicates interpretation and limits practical application across various domains, including healthcare, education, marketing, and the corporate sector. In response to this challenge, the present article aims to identify key algorithmic approaches to clustering unstructured texts and to analyze software systems implementing these methods. The primary research strategy is based on a comparative and analytical approach that enables the generalization and classification of contemporary machine learning algorithms applied to text data processing. The study

reviews both traditional clustering techniques and advanced architectures incorporating unsupervised learning, numerical vector representations, and neural network models. Software tools are examined with a focus on their levels of accuracy, interpretability, and adaptability. As a result, the study systematizes criteria for selecting methods according to specific tasks, highlights limitations of existing approaches, and outlines promising directions for further development. The findings are intended to support professionals engaged in designing and deploying software solutions for the automatic processing and analysis of textual information.

Keywords: text clustering, unstructured data, topic modeling, machine learning, vector representations, unsupervised algorithms, software frameworks, text mining.

For citation: Kondakov V.S., Kuznetsova A.V. Clustering algorithms for unstructured texts and their implementation in software systems. *Modeling, Optimization and Information Technology*. 2025;13(3). (In Russ.). URL: https://moitvivt.ru/ru/journal/pdf?id=1970 DOI: 10.26102/2310-6018/2025.50.3.010

Введение

Примерно 80 % информации, циркулирующей в рамках деятельности современных организаций, представлено в неструктурированном виде, что существенно затрудняет ее интерпретацию и использование без специальных методов обработки [1]. Стремительное увеличение объема цифровых текстов, ежедневно генерируемых в глобальной сети, приводит к необходимости разработки и внедрения инструментов, способных обрабатывать и анализировать такие массивы данных без участия человека [2]. Эта динамика усиливает интерес научного сообщества к задачам интеллектуального анализа текстов и стимулирует поиск новых решений для работы с растущими объемами информации.

Наряду с числовыми данными, текстовые массивы содержат значительное количество значимой информации, нуждающейся в систематизации, извлечении и интерпретации. Разнообразие и сложность естественного языка создают уникальные вызовы при попытках формализованной обработки текстов, что требует применения специальных алгоритмических подходов [3]. Текстовые данные сегодня лежат в основе целого спектра практикоориентированных направлений, включая анализ пользовательских мнений, тематическое моделирование, создание рекомендаций, генерацию поисковых подсказок и автоматическое выделение ключевых смысловых единиц [4]. В последние годы особое распространение получили подходы, основанные на разбиении программных систем на логически обособленные сегменты. В этом контексте активно развиваются такие методы, как кластеризация и шаблонное сопоставление, которые зарекомендовали себя как универсальные инструменты анализа структурных и семантических характеристик данных [5]. Принцип группировки на основе внутренней схожести объектов нашел применение в различных прикладных задачах, от анализа программного кода до обработки текстовой информации.

Настоящая работа направлена на освещение существующих алгоритмических решений в области машинного обучения, применяемых для анализа текстов. В фокусе внимания – механизмы кластеризации, позволяющие идентифицировать скрытые структуры в массиве текстов и организовать их в осмысленные группы. Также рассматриваются программные инструменты, реализующие соответствующие алгоритмы, и оценивается их эффективность с точки зрения обработки неструктурированных текстовых данных.

Материалы и методы

Настоящее исследование основано на систематическом обзоре современных теоретических и прикладных разработок в области кластеризации неструктурированных

текстов, а также на анализе программных инструментов, реализующих соответствующие алгоритмы. Методология работы включает многоэтапную процедуру, позволяющую всесторонне охарактеризовать существующие решения, классифицировать их по типологическим признакам и выявить тенденции развития в данной области. В качестве основного исследовательского метода был избран критико-аналитический подход, сочетающий качественный анализ алгоритмических схем с эмпирическим рассмотрением практических реализаций в программных системах. Объектом анализа выступили научные публикации, документация библиотек с открытым исходным кодом, описания прикладных решений и отчеты о внедрении кластеризующих систем в реальных условиях. Все источники были отобраны по принципу релевантности, новизны и цитируемости, преимущественно из журналов, индексируемых в Scopus, Web of Science. Методология статьи ориентирована на полноту охвата современных решений, критический разбор их свойств и предоставление базиса для их дальнейшего использования и сравнения в исследовательских и практических целях.

Результаты и обсуждение

Кластеризация текстов представляет собой одно из ключевых направлений в области анализа неструктурированной информации, направленное на объединение объектов по признаку внутренней близости. Основной задачей кластерного анализа является идентификация взаимосвязей между элементами данных с последующим выделением относительно однородных групп [5]. При этом особое внимание уделяется выявлению архитектурных и содержательных характеристик объектов, которые могут быть организованы в структуры с четко выраженными центрами притяжения признаков. В контексте обработки текстов без предварительной разметки особую актуальность приобретает задача распределения текстов на группы, обладающие высоким уровнем внутреннего сходства и низкой степенью перекрытия между собой. Это позволяет формировать компактные и интерпретируемые кластеры, внутри которых тексты характеризуются близостью по содержанию, стилю или тематике [4]. Независимо от объема и формы представления, тексты обрабатываются на основе их признакового пространства, в котором вычисляются меры сходства и различия.

Существенную роль в построении эффективных моделей играет тематическое моделирование – подход, относящийся к методам обучения без учителя, направленный на выявление латентных семантических закономерностей в текстовых массивах. Такие модели предназначены для автоматического обнаружения повторяющихся тем и концептов, скрытых за поверхностным лексическим разнообразием [2]. Используя вероятностные предположения и статистические методы, тематические алгоритмы позволяют выявить устойчивые смысловые паттерны, что особенно ценно при работе с большими корпусами документов. Одним из ключевых механизмов, лежащих в основе тематических моделей, является возможность формирования групп документов, объединенных по смыслу. Это не только способствует систематизации информации, но и позволяет находить содержательные связи между, на первый взгляд, разнородными текстами [6]. Кластеризация, реализуемая в рамках таких подходов, позволяет выявить структуры, которые не очевидны при поверхностном чтении, что делает ее важным инструментом в прикладных и исследовательских задачах.

Важное значение в современной текстовой аналитике имеют векторные представления, отражающие семантику текстов в числовом виде. Эти представления позволяют проектировать тексты в многомерное пространство признаков, в котором схожие тексты оказываются в относительной близости друг к другу [7]. Такая геометризация смысла открывает возможности для применения математических методов

оценки расстояний между объектами, что, в свою очередь, делает возможной реализацию кластеризации на основе чисто формализованных признаков. Применение кластерного анализа на уровне документов позволяет осуществлять структурирование текстовой коллекции по критерию смысловой близости. Это особенно важно при работе с большими объемами данных, когда ручная сортировка и категоризация становятся неэффективными [8]. Подобная автоматизация делает возможным разбиение текстов на логически обоснованные группы, каждая из которых может быть дополнительно проанализирована для извлечения содержательной информации.

Множество алгоритмов кластеризации, применяемых к неструктурированным текстам, различаются по способу обработки данных, формированию кластеров и степени интерпретируемости результатов. Один из часто используемых подходов – латентное размещение Дирихле (LDA), основан на предположении, что каждый документ включает в себя множество тем с различной степенью выраженности. Однако такая гипотеза не всегда отражает реальное тематическое распределение, особенно в случаях, когда тексты фокусируются на одном ключевом вопросе. В противоположность этому, латентно-семантический анализ (LSA) проигрывает по части точности, так как при преобразовании текстов в векторное пространство полностью игнорирует порядок слов, что в ряде случаев приводит к потере контекстуальных связей между понятиями. Одним из ключевых критериев оценки качества тематической кластеризации является показатель когерентности, который используется для количественного сравнения моделей. Высокие значения этого параметра свидетельствуют о согласованности тем внутри модели и позволяют утверждать о ее пригодности для анализа [9]. Некоторые современные системы, ориентированные на тематическое моделирование, демонстрируют стабильные показатели когерентности, что создает основу для проведения дальнейших теоретических и прикладных исследований.

Отдельную проблему при реализации кластеризации составляет этап нормализации входных данных. Применение метода min-max, чувствительного к выбросам, может существенно исказить значения признаков, что снижает точность итоговых кластеров. Это связано с тем, что крайние значения оказывают чрезмерное влияние на масштабирование, приводя к потере информации, потенциально значимой для классификации текстов. Среди направлений развития методов текстовой кластеризации все большую актуальность приобретают модели глубокого обучения, использующие контекстуальные эмбеддинги. Одной из перспективных альтернатив классическим техникам является использование векторных представлений, полученных с помощью языковых моделей типа BERT [10]. По сравнению с TF-IDF такие векторы сохраняют семантические связи между словами, что повышает точность группировки и способствует более содержательному разделению документов.

К числу методов, основанных на спектральной теории, относится спектральная кластеризация. Ее основная идея заключается в вычислении собственных значений и собственных векторов лапласиана графа сходства между объектами. Это позволяет получить вложение исходных данных в пространство меньшей размерности, в котором тексты легче поддаются разбиению на смысловые группы [8]. Такой подход особенно эффективен при наличии сложной структуры данных и пересекающихся тем. Современные алгоритмы, применяемые для кластеризации текстов, принято классифицировать по трем основным типам: иерархические, разбиения (или partitional) и алгоритмы на основе плотности. Иерархические методы, получившие широкое распространение, реализуют поэтапное объединение или разбиение объектов на основе мер близости, таких как косинусное сходство, коэффициенты Жаккара или Дайса [11]. Эти алгоритмы позволяют строить дендрограммы, отражающие многослойную структуру данных и обеспечивающие гибкость в выборе уровня детализации кластеров.

Наиболее известной и используемой техникой среди алгоритмов разбиения является метод k-средних. Он предполагает, что количество кластеров заранее задано, а алгоритм итеративно перераспределяет элементы по группам с учетом минимизации внутрикластерного расхождения. Главной целью такого подхода является нахождение центров кластеров, минимизирующих среднее квадратичное расстояние между объектами и соответствующим центром тяжести [12, 13]. Широкое распространение этого метода объясняется его относительной простотой и высокой скоростью работы на больших объемах данных. Дополнением к классическим алгоритмам разбиения служит нечеткий вариант – метод c-средних (Fuzzy C-Means). В отличие от жесткого присваивания, FCM позволяет одному объекту принадлежать сразу к нескольким кластерам с различной степенью уверенности [10], что делает его особенно полезным при анализе текстов с перекрывающимися тематиками.

В исследовании Х.С. Ши и Т. Сакаи предлагаются инновационные архитектуры, использующие принципы контрастивного обучения для безразметочной кластеризации. Такой подход опирается на обучение представлений, способных выделять тонкие различия между группами, и может применяться для кластеризации встраиваемых представлений предложений и абзацев [4]. Согласно научной работе Н. Мюнихоффа и соавторов одна из моделей, реализованная в мини-пакетном режиме, может обучаться на компактных поднаборах данных, позволяя эффективно группировать фрагменты текста в осмысленные категории даже при ограниченных вычислительных ресурсах [14]. По результатам сопоставления различных моделей в исследовании Д. Воскерьяна, Р. Джаюси, М. Юсефа установлено, что вероятностные алгоритмы, такие как LDA и LSI, обеспечивают высокие значения F1-метрики при минимальном числе признаков [2]. Это делает их привлекательными для задач, в которых требуется баланс между вычислительной эффективностью и качеством тематической группировки. Иерархические методы в научной работе Х. Янь, Л. Гуй и Ю. Хэ показали высокую степень адаптивности к разнородным данным и обеспечили хорошую интерпретируемость структуры текстовых массивов [15].

Современные технологии обработки неструктурированных текстовых данных активно развиваются за счет интеграции специализированных программных решений, предназначенных для кластеризации, тематического моделирования и визуального анализа. Одним из таких инструментов является VOSviewer – платформа, предназначенная для построения и изучения библиографических карт. Несмотря на свою популярность, данный пакет не предоставляет пользователям функциональности для анализа динамики по временным срезам и не позволяет визуализировать данные в форматах, таких как иерархические деревья, географические схемы или спектрограммы, что ограничивает его применимость в задачах с расширенным спектром аналитических требований. Альтернативные подходы ориентированы на более глубокую обработку текстов, включая токенизацию, векторизацию и тематическое разбиение. Примером может служить исследование У.А. Букар и соавторов, в котором в результате процедуры токенизации было сформировано 128 файлов, представляющих текстовые единицы. На их основе были построены карты со-встречаемости терминов, отражающие структуру кластеризации понятийных единиц, что позволило выявить закономерности в распределении лексических признаков [16].

Одним из наиболее технологически продвинутых фреймворков в этой области выступает BERTopic. Его архитектура построена по многоступенчатой схеме: на первом этапе извлекаются эмбеддинги документов, затем происходит уменьшение размерности векторного пространства, после чего осуществляется кластеризация. Финальным шагом является формирование тем при помощи модифицированного TF-IDF, учитывающего значимость слов в пределах кластера, а не корпуса в целом. Разработчикам,

ориентированным на оценку и сравнение различных моделей тематического анализа, доступен пакет OCTIS (Optimizing and Comparing Topic models is Simple). Это open-source решение, предназначенное для экспериментирования с тематическими моделями и выбора оптимальных параметров на основании заданных метрик [7]. С его помощью можно реализовывать гибкие пайплайны без необходимости глубокого погружения в архитектуру используемых алгоритмов. Дополнительный интерес представляет система MTEB (Massive Text Embedding Benchmark), доступная с открытым исходным кодом. Она позволяет тестировать широкое множество моделей эмбеддингов при минимальных усилиях со стороны исследователя [14]. Для интеграции новых данных и запуска оценки достаточно внести менее десятка строк программного кода, что делает платформу удобной для быстрого прототипирования.

В качестве основного инструментария для построения кластеризующих систем чаще всего применяется язык программирования Python, обладающий широкой экосистемой библиотек. Наиболее часто используемыми являются: Scikit-learn – для построения и обучения моделей машинного обучения; NLTK и spaCy – для лингвистического препроцессинга; Gensim – для тематического моделирования (в том числе реализация алгоритмов LDA и Doc2Vec); NetworkX – для анализа графов; Yellowbrick – для визуализации результатов моделирования [17]. Сочетание этих инструментов позволяет выстраивать многоуровневые цепочки обработки, охватывающие все этапы работы с текстом: от чистки до интерпретации итогов кластеризации. Для морфологического разбора и нормализации русскоязычных текстов активно используется анализатор pymorphy2, который нередко включается в связке с Gensim и NLTK при построении предварительной обработки корпусов [18]. Такой подход обеспечивает высокую степень согласованности между синтаксическим и семантическим уровнями анализа. Среди решений, предлагающих функциональность для аннотирования и структурирования текстов, можно выделить GATE (General Architecture for Text Engineering). Эта платформа может применяться как в виде интегрированной среды, так и как библиотека в составе более сложных систем [19]. Ее архитектура предусматривает использование различных ресурсов: языковых, визуальных, аналитических, что обеспечивает расширенные возможности по метатегированию и извлечению информации из текстовых данных.

Хранение и управление большими массивами неструктурированной информации требует применения гибких СУБД. В этом контексте NoSQL-технологии, в частности, MongoDB, оказываются особенно эффективными [20]. Они предоставляют динамическую структуру хранения, отсутствие фиксированных схем и поддержку масштабируемых запросов, что делает их пригодными для аналитики в реальном времени и быстрой агрегации текстовых фрагментов по тематическим признакам. На прикладном уровне популярны схемы, включающие несколько специализированных компонентов. Например, в одном из проектов В. Мехты, С. Бавы, Дж. Сингха структура приложения включала Elasticsearch (поисковый движок), Doc2Vec (для генерации векторных представлений) и модуль кластеризации, отвечающий за группировку данных [21]. Такой модульный подход обеспечивает масштабируемость и легкую адаптацию к различным предметным областям.

Пример успешной реализации кластеризационного алгоритма представлен в системе, разработанной И.Э. Арнарссоном и соавторами, использующей последовательное применение ряда модулей: очистка данных (TextCleaner), преобразование в эмбеддинги (Embedder), первичная группировка (DBSCAN), а затем уточняющая кластеризация с использованием агломеративного метода [1]. Комбинирование нескольких уровней кластеризации позволяет повысить точность итоговой сегментации. При анализе связей между терминами в больших текстовых массивах нередко применяется методика

построения карт совместной встречаемости. Она реализована в VOSviewer, где сила связи между словами отображается в виде визуальных кластеров [22], что позволяет исследователю быстро выявить тематические ядра и характерные паттерны в распределении терминов по корпусу. Материалы настоящего исследования обобщены в Таблице 1.

Таблица 1 – Кластеризация текстов: алгоритмы и программные решения
Table 1 – Text clustering: algorithms and software solutions

Тип алгоритма	Краткая характеристика	Примеры программной реализации
Иерархическая кластеризация	Формирует древовидную структуру групп на основе последовательного объединения или разделения объектов по показателю близости	«Scikit-learn», «SciPy», «NetworkX», GATE (при структурировании текстов)
Метод k-средних (k-Means)	Требует предварительного задания числа кластеров; минимизирует разброс внутри групп; прост в реализации и эффективен при больших объёмах	«Scikit-learn», «Elasticsearch» в модульных решениях, «Yellowbrick» для визуального анализа
Нечеткий метод c-средних (Fuzzy C-Means)	Позволяет одному элементу одновременно принадлежать нескольким кластерам с различной степенью уверенности, подходит для неоднозначных данных	«scikit-fuzzy», интеграции с «NumPy», «pandas», а также ручные реализации для кастомных пайплайнов
LDA (Latent Dirichlet Allocation)	Использует вероятностный подход для выявления скрытых тематик в текстах; применим к корпусам документов с пересекающимися темами	«Gensim», «OCTIS», «MALLET», «Scikit-learn», «BERTopic»
LSA (Latent Semantic Analysis)	Опирается на матричную декомпозицию; снижает размерность признаков, но теряет порядок слов, что может приводить к потере контекста	«Scikit-learn», «Gensim», «SVDLib», библиотеки линейной алгебры («NumPy», «SciPy»)
DBSCAN (кластеризация по плотности)	Определяет кластеры как области повышенной плотности точек, устойчив к шуму и не требует задания числа групп	«Scikit-learn», «ELKI», используется в связке с «spaCy», «TextCleaner», «Embedder» в пользовательских пайплайнах
Спектральная кластеризация	Работает на основе спектральной теории графов; снижает размерность через собственные векторы, затем применяет методы разбиения	«Scikit-learn», «SpectralClustering» API, «BERTopic», специализированные модули «NumPy»/ «SciPy»

Таблица 1 (продолжение)
Table 1 (continued)

Контрастивное обучение (Unsupervised)	Позволяет группировать без аннотированных данных, обучая встраивания, чувствительные к различиям между парами примеров	Кастомные архитектуры на «PyTorch», «Transformers», «BERT», «SimCLR», «MTEB»
BERT и контекстуальные эмбеддинги	Преобразуют текст в контекстно-зависимые векторы, улучшающие качество кластеризации благодаря сохранению семантических связей между словами	«BERTopic», «SentenceTransformers», «Hugging Face Transformers», интеграция с «UMAP» и «HDBSCAN» для кластеризации
Комплексные пайплайны (модульные схемы)	Используют комбинации: очистка текста → эмбеддинги → кластеризация (например, DBSCAN + агломеративная кластеризация)	Сборки на «Python»: «TextCleaner», «Embedder», «Scikit-learn», «pymorphy2», «Gensim», «NLTK», в связке с «Elasticsearch» или «MongoDB»

Применение алгоритмов кластеризации в сфере обработки текстов охватывает широкий круг практических задач и отраслей, где требуется анализ больших объемов неструктурированных данных. Одной из наиболее активно развивающихся областей выступает анализ пользовательских отзывов, особенно в контексте управления объектами недвижимости и инфраструктурными системами (facility management). Кластеризация позволяет систематизировать и группировать жалобы и предложения пользователей, облегчая адаптацию обслуживающих процессов к актуальным потребностям. Например, при использовании алгоритма GSDMM (Gibbs Sampling Dirichlet Multinomial Mixture) Ч.Ю. Паку и соавторам удалось выделить типовые тематические направления в обращениях: вопросы, связанные с персоналом, качеством обслуживания, состоянием торговых помещений и функциональностью пространств [23]. Не менее значимыми становятся технологии интеллектуального анализа текстов в корпоративной среде, где они служат инструментом мониторинга рисков, прогнозирования критических ситуаций и совершенствования коммуникаций с клиентами [9]. Сегментация отзывов и обращений по тематическим кластерам позволяет своевременно выявлять потенциальные угрозы для репутации компании, а также формировать индивидуальные предложения, соответствующие ожиданиям конкретных групп потребителей.

Особое место в прикладной практике занимает использование кластеризации в медицине и биомедицинских исследованиях. Здесь востребованы методы, способные обрабатывать сложные и высокоразмерные данные, включая результаты диагностики, медицинские записи и биосигналы. Алгоритмы нечеткой кластеризации, допускающие частичную принадлежность объектов к нескольким группам, хорошо зарекомендовали себя в задачах распознавания патологий и построения диагностических моделей. В частности, в исследовании Дж. Рашид и соавторов, при классификации специализированных биомедицинских наборов данных, таких как MuchMore Springer и Ohsumed, модель MKFTM продемонстрировала преимущество по точности по сравнению с рядом классических алгоритмов [24]. Работа с текстовыми данными в клинической сфере показала, что применение алгоритмов кластеризации к электронным медицинским картам и другим неструктурированным источникам информации

позволяет извлекать ценную информацию для диагностики и прогнозирования. Например, в научной работе К.Х. Го и соавторов один из алгоритмов предсказывал развитие сепсиса за 12 часов до клинических проявлений, демонстрируя улучшенные значения по ключевым метрикам – AUC, чувствительности и специфичности [25].

В сфере тематического моделирования особую ценность представляет использование латентного размещения Дирихле (LDA), позволяющего определять скрытые смысловые структуры в больших массивах текстов. Как показало исследование К.А. Мелтона и соавторов, этот метод оказался особенно полезен в случаях, когда речь идет об анализе крупных текстовых коллекций с неочевидной семантической структурой [26]. LDA помогает извлекать повторяющиеся темы, которые не всегда выражены явно, и тем самым улучшает понимание внутренней логики текстового материала. В маркетинговых исследованиях особое внимание уделяется обработке пользовательского контента в цифровой среде, включая отзывы, размещенные в рамках электронной коммуникации между потребителями (eWOM – electronic word of mouth) [27]. Анализ таких текстов с помощью кластеризации позволяет выявить доминирующие установки, оценить восприятие брендов и адаптировать маркетинговые стратегии к ожиданиям целевых аудиторий.

Продолжающееся развитие моделей встраивания (embedding) оказывает значительное влияние на результаты кластеризации. Исследование М.Х. Ахмеда и соавторов показывает, что замена традиционных векторных представлений, таких как TF-IDF, на нейросетевые эмбеддинги приводит к более точной группировке и лучшей интерпретации текстов. Это особенно важно в условиях анализа коротких текстов, где классические методы часто демонстрируют низкую производительность. Функциональное разнообразие применения кластеризации охватывает широкий спектр направлений: от информационного поиска и построения рекомендательных систем до анализа данных в Интернете вещей (IoT), биологии, инженерных и климатических систем, а также в медицине и фармацевтике [11]. Каждый из этих доменов предъявляет специфические требования к алгоритмам обработки, но объединяет их необходимость в эффективном и интерпретируемом групповом анализе текстов. Отдельный интерес вызывают задачи, где текстовая кластеризация сочетается с мультимодальными данными, например, с изображениями. Подобные гибридные подходы используются для генерации подсказок в системах, объединяющих визуальные и текстовые элементы [28]. Совместное обучение языковой модели и функции потерь позволяет достичь лучшего качества генеративных решений за счет более точного разделения объектов по смысловым признакам. Кластеризация, в данном случае, выполняет функцию семантического якоря, способствуя структурированию обучающего пространства и улучшению обобщающих способностей модели.

Заключение

Несмотря на значительные успехи в разработке алгоритмов кластеризации текстов, остается ряд нерешенных вопросов, которые снижают точность интерпретации и стабильность результатов. Одной из ключевых проблем является ограниченная способность алгоритмов достоверно отражать структуру и поведение программных компонентов в рамках кластерной модели. На практике это проявляется в расхождении между ожидаемой и фактической архитектурой полученных групп, что свидетельствует о необходимости дальнейшей доработки существующих методов. Характерной особенностью многих современных решений является ограниченная обобщаемость полученных результатов. Это объясняется спецификой используемых данных и контекстной зависимостью моделей. Для большинства исследований, основанных на

эмпирических данных, подобное ограничение оказывается неизбежным, поскольку эффективность кластеризации нередко оказывается тесно привязанной к особенностям конкретного корпуса и применяемой методологии.

В условиях роста интереса к генеративным подходам и их интеграции в системы обработки текстов встает вопрос о сложности архитектурных решений. Большинство генеративных моделей требуют глубокой настройки, значительных вычислительных ресурсов и сложной инженерной поддержки. Несмотря на это, их эффективность в задаче семантического разбиения текстов пока уступает специализированным алгоритмам кластеризации, разработанным с прицелом на структурную интерпретацию данных. Одно из перспективных направлений заключается в развитии инструментов, ориентированных на тематический анализ с высокой степенью автоматизации. В рамках дальнейших исследований предлагается внедрение модулей, способных не только выделять латентные темы, но и оценивать их значимость, устойчивость и релевантность к поставленным задачам. Это может существенно повысить точность результатов и расширить возможности практического применения кластеризации.

Использование обширных корпусов предварительно обученных данных, с одной стороны, облегчает генерацию кластерных структур, но с другой – порождает этические и технологические риски. Некоторые модели, такие как Clus, при работе с чувствительным контентом могут порождать некорректные или предвзятые суждения. В связи с этим особое внимание уделяется механизмам дообучения, направленным на фильтрацию нежелательных эффектов и повышение социальной приемлемости результатов. Устойчивым вызовом остается разрыв между алгоритмами, использующими полностью неконтролируемое обучение, и подходами, основанными на контролируемом машинном обучении. Для достижения приемлемой точности неконтролируемым моделям зачастую необходима дополнительная ручная настройка, что снижает уровень автоматизации и требует участия эксперта на стадии интерпретации результатов.

Несмотря на высокие показатели производительности некоторых моделей, таких как BERTopic, в реальных сценариях наблюдаются ограничения, связанные с их масштабируемостью. При переходе от экспериментальных наборов данных к промышленным объемам информации возникает необходимость в перераспределении вычислительных ресурсов и оптимизации процессов кластеризации, что требует инженерного переосмысления используемых архитектур. Наблюдается рост интереса к тонкой адаптации крупных языковых моделей (LLMs) под конкретные задачи, включая онлайн-образование, анализ пользовательской активности на MOOC-платформах и оценку образовательного контента. Потенциал таких решений значителен, но остается открытым вопрос о выборе метрик и алгоритмов, способных учитывать специфику образовательной среды.

Некоторые инструменты, например, TextNetTopics, демонстрируют высокую эффективность на отдельных датасетах, однако при работе с разнородными текстовыми массивами возможны проблемы с воспроизводимостью и устойчивостью. Такая непоследовательность требует от исследователя глубокого понимания внутренних механизмов алгоритма и навыков настройки параметров в зависимости от структуры и тематики корпуса. Предстоящие этапы исследований предполагают сосредоточение усилий на создании адаптивных, устойчивых и интерпретируемых моделей, способных сохранять высокое качество кластеризации при переносе между различными предметными областями. Особый интерес представляют кросс-дисциплинарные и мультимодальные подходы, где текстовые данные взаимодействуют с визуальными, числовыми и контекстуальными источниками информации, открывая новые перспективы для кластерного анализа в условиях реальной многослойной информации.

СПИСОК ИСТОЧНИКОВ / REFERENCES

1. Arnarsson I.Ö., Frost O., Gustavsson E., Jirstrand M., Malmqvist J. Natural Language Processing Methods for Knowledge Management–Applying Document Clustering for Fast Search and Grouping of Engineering Documents. *Concurrent Engineering: Research and Applications*. 2021;29(2):142–152. https://doi.org/10.1177/1063293X20982973
2. Voskergian D., Jayousi R., Yousef M. Topic Selection for Text Classification Using Ensemble Topic Modeling with Grouping, Scoring, and Modeling Approach. *Scientific Reports.* 2024;14. https://doi.org/10.1038/s41598-024-74022-2
3. Ковтун Д.Б. Исследование внутриведомственного взаимодействия органов власти РФ на основе документов стратегического планирования с помощью технологии Text Mining. *Московский экономический журнал*. 2021;(2). https://doi.org/10.24412/2413-046X-2021-10119
   Kovtun D.B. Research of the Intradepartmental Authority of the Russian Federation Based on Strategic Planning Using Text Mining Technology. *Moscow Economic Journal.* 2021;(2). (In Russ.). https://doi.org/10.24412/2413-046X-2021-10119
4. Shi H., Sakai T. Self-Supervised and Few-Shot Contrastive Learning Frameworks for Text Clustering. *IEEE Access*. 2023;11:84134–84143. https://doi.org/10.1109/ACCESS.2023.3302913
5. Tulli S.K.C. Enhancing Software Architecture Recovery: A Fuzzy Clustering Approach. *International Journal of Modern Computing.* 2024;7(1):141–153.
6. Khodeir N., Elghannam F. Efficient Topic Identification for Urgent MOOC Forum Posts Using BERTopic and Traditional Topic Modeling Techniques. *Education and Information Technologies.* 2025;30:5501–5527. https://doi.org/10.1007/s10639-024-13003-4
7. Grootendorst M. BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure. arXiv. URL: https://arxiv.org/abs/2203.05794 [Accessed 24th May 2025].
8. Cozzolino I., Ferraro M.B. Document Clustering. *WIREs Computational Statistics*. 2022;14(6). https://doi.org/10.1002/wics.1588
9. Kavitha D., Anandha Mala G.S., Padmavathi B., Varshni S.V. Text Mining: Clustering Using BERT and Probabilistic Topic Modeling. *Social Informatics Journal*. 2023;2(2):1–13. https://doi.org/10.58898/sij.v2i2.01-13
10. Subakti A., Murfi H., Hariadi N. The Performance of BERT as Data Representation of Text Clustering. *Journal of Big Data*. 2022;9. https://doi.org/10.1186/s40537-022-00564-9
11. Ahmed M.H., Tiun S., Omar N., Sani N.S. Short Text Clustering Algorithms, Application and Challenges: A Survey. *Applied Sciences.* 2023;13(1). https://doi.org/10.3390/app13010342
12. Маслова М.А. Автоматизированный подход к отбору предложений для генерации тестовых заданий. *Computational Nanotechnology*. 2024;11(2):29–34. https://doi.org/10.33693/2313-223X-2024-11-2-29-34
    Maslova M.A. An Automated Approach to Selecting Sentences for Test Case Generation. *Computational Nanotechnology*. 2024;11(2):29–34. (In Russ.). https://doi.org/10.33693/2313-223X-2024-11-2-29-34
13. Probierz B., Kozak J., Hrabia A. Clustering of Scientific Articles Using Natural Language Processing. *Procedia Computer Science*. 2022;207:3449–3458. https://doi.org/10.1016/j.procs.2022.09.403
14. Muennighoff N., Tazi N., Magne L., Reimers N. MTEB: Massive Text Embedding Benchmark. arXiv. URL: https://arxiv.org/abs/2210.07316 [Accessed 24th May 2025].

15. Yan H., Gui L., He Yu. Hierarchical Interpretation of Neural Text Classification. *Computational Linguistics*. 2022;48(4):987–1020. https://doi.org/10.1162/coli_a_00459
16. Ali Bukar U., Sayeed M.S., Razak S.F.A., Yogarayan S., Amodu O.A., Mahmood R.A.R. A Method for Analyzing Text Using VOSviewer. *MethodsX*. 2023;11. https://doi.org/10.1016/j.mex.2023.102339
17. Анферова М.С., Белевцев А.М. Разработка алгоритмов интеллектуального сервиса поиска и мониторинга информации. *Известия ЮФУ. Технические науки.* 2021;(3):6–17. https://doi.org/10.18522/2311-3103-2021-3-6-17
Anferova M.S., Belevtsev A.M. Development of Algorithms of Intelligent Service for Information Search and Monitoring. *Izvestiya SFedU. Engineering Sciences*. 2021;(3):6–17. (In Russ.). https://doi.org/10.18522/2311-3103-2021-3-6-17
18. Заббаров З.Р., Волков А.К. Метод выявления актуальных тем тренажерной подготовки пилотов на основе кластеризации отчетов по безопасности полетов. *Научный вестник МГТУ ГА*. 2024;27(4):34–49. https://doi.org/10.26467/2079-0619-2024-27-4-34-49
Zabbarov Z.R., Volkov A.K. A Method for Identifying Relevant Topics of Pilot Simulator Training Based on Clustering of Flight Safety Reports. *Civil Aviation High Technologies*. 2024;27(4):34–49. (In Russ.). https://doi.org/10.26467/2079-0619-2024-27-4-34-49
19. Губанов А.Р., Данилов А.А., Исаев Ю.Н., Губанова Г.Ф. Проблемы извлечения слабоструктурированной текстовой информации на основе технологии Text Mining (на материале русского и чувашского языков). *Филологические науки. Вопросы теории и практики*. 2024;17(9):3085–3090. https://doi.org/10.30853/phil20240437
Gubanov A.R., Danilov A.A., Isaev Y.N., Gubanova G.F. Problems of Extracting Semi-Structured Textual Information Based on Text Mining Technology (Using the Material of the Russian and Chuvash Languages). *Philology. Theory & Practice.* 2024;17(9):3085–3090. (In Russ.). https://doi.org/10.30853/phil20240437
20. Khan W., Kumar T., Zhang Ch., Raj K., Roy A.M., Luo B. SQL and NoSQL Database Software Architecture Performance Analysis and Assessments–A Systematic Literature Review. *Big Data and Cognitive Computing*. 2023;7(2). https://doi.org/10.3390/bdcc7020097
21. Mehta V., Bawa S., Singh J. WEClustering: Word Embeddings Based Text Clustering Technique for Large Datasets. *Complex & Intelligent Systems*. 2021;7(6):3211–3224. https://doi.org/10.1007/s40747-021-00512-9
22. Зеленков Ю.А., Анисичкина Е.А. Динамика исследований в области интеллектуального анализа данных: тематический анализ публикаций за 20 лет. *Бизнес-информатика*. 2021;15(1):30–46. (На англ.). https://doi.org/10.17323/2587-814X.2021.1.30.46
Zelenkov Yu., Anisichkina E. Trends in Data Mining Research: A Two-Decade Review Using Topic Analysis. *Business Informatics.* 2021;15(1):30–46. https://doi.org/10.17323/2587-814X.2021.1.30.46
23. Park Ju.Yo., Mistur E., Kim D., Mo Yu., Hoefer R. Toward Human-Centric Urban Infrastructure: Text Mining for Social Media Data to Identify the Public Perception of COVID-19 Policy in Transportation Hubs. *Sustainable Cities and Society*. 2022;76. https://doi.org/10.1016/j.scs.2021.103524
24. Rashid Ju., Kim Ju., Hussain A., Naseem U., Juneja S. A Novel Multiple Kernel Fuzzy Topic Modeling Technique for Biomedical Data. *BMC Bioinformatics*. 2022;23. https://doi.org/10.1186/s12859-022-04780-1

25. Goh K.H., Wang L., Yeow A.Yo.K., et al. Artificial Intelligence in Sepsis Early Prediction and Diagnosis Using Unstructured Data in Healthcare. *Nature Communications*. 2021;12. https://doi.org/10.1038/s41467-021-20910-4
26. Melton Ch.A., Olusanya O.A., Ammar N., Shaban-Nejad A. Public Sentiment Analysis and Topic Modeling Regarding COVID-19 Vaccines on the Reddit Social Media Platform: A Call to Action for Strengthening Vaccine Confidence. *Journal of Infection and Public Health*. 2021;14(10):1505–1512. https://doi.org/10.1016/j.jiph.2021.08.010
27. Alzate M., Arce-Urriza M., Cebollada J. Mining the Text of Online Consumer Reviews to Analyze Brand Image and Brand Positioning. *Journal of Retailing and Consumer Services.* 2022;67. https://doi.org/10.1016/j.jretconser.2022.102989
28. Fayou S., Ngo H.Ch., Sek Yo.W., Meng Z. Clustering Swap Prediction for Image-Text Pre-Training. *Scientific Reports.* 2024;14. https://doi.org/10.1038/s41598-024-60832-x

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Кондаков Вячеслав Серегеевич, аспирант, ассистент, Южно-Российский государственный политехнический университет (НПИ) имени М.И. Платова, Новочеркасск, Российская Федерация.
e-mail: kondakov_vyacheslavs@mail.ru
ORCID: 0009-0006-4058-9835

Vyacheslav S. Kondakov, Postgraduate, Platov South-Russian State Polytechnic University (NPI), Novocherkassk, the Russian Federation.

Кузнецова Алла Витальевна, кандидат технических наук, доцент, Южно-Российский государственный политехнический университет (НПИ) имени М.И. Платова, Новочеркасск, Российская Федерация.
e-mail: alvitkuz@yandex.ru
ORCID: 0000-0001-5028-0053

Alla V. Kuznetsova, Candidate of Engineering Sciences, Docent, Associate Professor, Platov South-Russian State Polytechnic University (NPI), Novocherkassk, the Russian Federation.

Статья поступила в редакцию 20.05.2025; одобрена после рецензирования 23.06.2025; принята к публикации 03.07.2025.

The article was submitted 20.05.2025; approved after reviewing 23.06.2025; accepted for publication 03.07.2025.