МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО»**

Кафедра социальной информатики

КЛАСТЕРИЗАЦИЯ СОЦИОЛОГИЧЕСКИХ ДАННЫХ С ПОМОЩЬЮ HDBCSAN И UMAP

(автореферат бакалаврской работы)

студента 5 курса 531 группы
направления 09.03.03 - Прикладная информатика
профиль Прикладная информатика в социологии
Социологического факультета
Магомедова Магомедхабиба Сайпудиновича

Научный руководитель
доцент, кандидат физико-математических наук _____ Л. Б. Тяпаев
подпись, дата

Зав. кафедрой
кандидат социологических наук, доцент _____ И.Г. Малинский
подпись, дата

Саратов 2025

ВВЕДЕНИЕ

Современные социологические исследования всё чаще сталкиваются с необходимостью анализа больших массивов данных, генерируемых в цифровой среде, особенно в социальных сетях. Такие данные многомерны, содержат шум и не имеют чёткой структуры, что делает их сложными для обработки традиционными методами. Кластеризация, как один из ключевых подходов машинного обучения, позволяет выявлять скрытые закономерности и группировать объекты по их сходству без предварительного задания категорий. Однако классические алгоритмы кластеризации, такие как k-средние, требуют указания числа кластеров заранее, что ограничивает их применимость к социологическим данным, где структура групп неизвестна. Высокая размерность текстовых данных требует применения методов снижения размерности для упрощения анализа и визуализации.

В последние годы алгоритмы HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) и UMAP (Uniform Manifold Approximation and Projection) продемонстрировали высокую эффективность в обработке сложных и шумных данных. HDBSCAN автоматически определяет число кластеров и устойчив к выбросам, а UMAP эффективно снижает размерность, сохраняя локальные и глобальные структуры данных. Их сочетание перспективно для анализа текстовых социологических данных, таких как комментарии в социальных сетях, где важно учитывать смысловую и эмоциональную близость высказываний.

Актуальность исследования обусловлена потребностью в новых методах анализа социологических данных, позволяющих выявлять группы пользователей на основе их текстовых сообщений без предварительных предположений о числе групп. Это важно для изучения общественного мнения, социальных трендов и поведения в цифровой среде, где данные часто неоднородны и содержат шум.

Цель работы: разработать и апробировать метод кластеризации социологических текстовых данных с использованием HDBSCAN и UMAP для

выявления устойчивых групп пользователей по смыслу их высказываний без необходимости задания числа кластеров.

Задачи исследования:

- Изучить теоретические основы методов кластеризации и снижения размерности данных.
- Провести сравнительный анализ существующих алгоритмов кластеризации и их применимости к социологическим данным.
- Обосновать выбор HDBSCAN и UMAP для анализа текстовых данных.
- Разработать метод обработки и кластеризации текстов, включающий предобработку, векторизацию, снижение размерности и группировку.
- Провести эксперимент на реальных данных из социальных сетей и оце-нить качество кластеризации.
- Сравнить результаты с традиционными методами кластеризации (k-средние, DBSCAN, иерархическая кластеризация).

Объект исследования: многомерные социологические данные, представленные текстовыми сообщениями пользователей социальных сетей, отражающими их поведение, эмоции и мнения.

Предмет исследования: процесс кластеризации текстовых данных с использованием HDBSCAN и UMAP для выделения групп пользователей по смысловому содержанию их сообщений.

Эмпирическая база. Исследование основано на корпусе из 1200 анонимизированных комментариев пользователей социальных сетей, собранных из открытых источников. Данные включают сообщения разной тональности (позитивные, негативные, нейтральные) и прошли предварительную обработку: удаление стоп-слов, лемматизацию и векторизацию с использованием TF-IDF.

Научная новизна. Впервые для анализа русскоязычных текстов социальных сетей предложено и реализовано сочетание HDBSCAN и UMAP. Этот подход позволяет автоматически определять число кластеров и выявлять

подгруппы в данных, что ранее не применялось в подобных задачах на русском языке. Разработанный метод интегрирует предобработку текстов и визуализацию результатов, повышая интерпретируемость кластеризации.

Методы исследования. Работа использует методы машинного обучения (кластеризация, снижение размерности, обработка текстов), включая TF-IDF векторизацию, алгоритмы HDBSCAN и UMAP, а также оценку качества кластеризации с помощью силуэтного коэффициента. Реализация выполнена на языке Python с использованием библиотек scikit-learn, hdbscan, umap-learn, pandas и Plotly.

Практическая значимость. Разработанный метод может применяться для анализа общественного мнения, сегментации аудитории социальных сетей и выявления социальных трендов. Результаты кластеризации полезны для социологов, маркетологов и специалистов по анализу данных.

Структура работы. Автореферат состоит из введения, трёх разделов, заключения и списка литературы. Первый раздел посвящён теоретическим основам кластеризации и снижения размерности. Второй описывает разработанный метод. Третий содержит результаты эксперимента, их анализ и сравнение с другими методами

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

В первой главе «Теоретические основы методов кластеризации и уменьшения размерности» рассмотрена кластеризация — это метод машинного обучения, направленный на группировку объектов на основе их сходства без использования обучающих данных. Она широко применяется в социологии для анализа текстов, поведения и предпочтений пользователей. Основная цель кластеризации — выявить скрытые структуры в данных, где заранее неизвестно, сколько групп должно быть. В зависимости от подхода к группировке алгоритмы кластеризации делятся на несколько категорий: центроидные, иерархические и плотностные.

Центроидные методы, такие как k-средние, минимизируют сумму квадратов расстояний между объектами и центрами кластеров. Алгоритм k-средних, предложенный MacQueen в 1967 году [3], эффективен для данных с чёткой сферической структурой. Однако он требует задания числа кластеров зара-нее, что проблематично для социологических данных, где структура групп неизвестна. Кроме того, k-средние чувствительны к шуму и выбросам, что может приводить к некорректному объединению объектов, особенно в слу-чае текстовых данных с неравномерной плотностью. Например, в анализе комментариев из социальных сетей k-средние могут смешивать позитивные и нейтральные высказывания из-за наличия общих слов.

Иерархические методы кластеризации строят дерево кластеров (дендрограмму), позволяя анализировать данные на разных уровнях детализации. Они делятся на агломеративные (объединение мелких кластеров в крупные) и дивизионные (разделение большого кластера на меньшие). Иерархическая кластеризация полезна для небольших наборов данных, но её вычислительная сложность ($O(n^2)$ или выше) делает её неподходящей для больших корпусов, таких как комментарии в социальных сетях. Кроме того, выбор уровня разбиения дендрограммы требует экспертной оценки, что усложняет автоматизацию.

Плотностные методы, такие как DBSCAN и HDBSCAN, группируют объекты по плотности их расположения, что делает их устойчивыми к выбросам. DBSCAN [4] требует настройки двух параметров: радиуса окрестности () и минимального числа точек (minPts). Это ограничивает его применимость, так как выбор параметров зависит от данных. HDBSCAN [5] улучшает DBSCAN, автоматически определяя число кластеров и адаптируясь к данным разной плотности. Алгоритм строит иерархию плотностных кластеров и использует минимальный размер кластера для выделения устойчивых групп. Это делает HDBSCAN идеальным для анализа текстовых данных, где присутствуют шумные записи, такие как неформальные комментарии или саркастические высказывания.

Высокая размерность текстовых данных, получаемых после векторизации, требует применения методов снижения размерности. Эти методы уменьшают число признаков, сохраняя ключевые характеристики данных. Рассмотрим три основных подхода:

Анализ главных компонент (PCA) — линейный метод, который проецирует данные на пространство меньшей размерности, максимизируя дисперсию. PCA эффективен для данных с высокой корреляцией, но теряет нелинейные связи, что ограничивает его применимость к текстам. Например, в случае текстовых векторов PCA может игнорировать семантическую близость слов.

t-SNE [10] сохраняет локальные структуры данных, минимизируя расхождение Кульбака-Лейблера между распределениями в исходном и проекционном пространствах. Однако t-SNE медленный (сложность $O(n^2)$) и менее точен в сохранении глобальных структур, что делает его менее подходящим для больших наборов данных.

UMAP [6] использует топологический подход, основанный на теории многообразий. Он сохраняет как локальные, так и глобальные структуры данных, что делает его эффективным для визуализации текстовых данных. UMAP быстрее t-SNE и поддерживает различные метрики расстояния, такие как косинусное расстояние, подходящее для текстов. Например, UMAP может выделить группы комментариев с похожей эмоциональной окраской, сохраняя их положение относительно других групп.

HDBSCAN и UMAP выбраны для исследования благодаря их способности обрабатывать сложные, шумные данные и предоставлять интерпретируемые результаты. HDBSCAN не требует задания числа кластеров, что идеально для социологических данных, а UMAP обеспечивает качественное снижение размерности, упрощая визуализацию и анализ.

Для оценки качества кластеризации использовались следующие метрики:

- Силуэтный коэффициент [1]: измеряет, насколько объекты внутри кластера ближе друг к другу, чем к объектам других кластеров. Значения близкие к 1 указывают на хорошее разделение.
- Индекс Дэвиса-Болдина: оценивает компактность и разделимость кла-стеров.

Внутренняя дисперсия: измеряет разброс внутри кластеров.

Для текстовых данных важна предобработка. Метод TF-IDF [7] преобразует тексты в числовые векторы, учитывая частоту слов и их значимость в корпусе. Лемматизация и удаление стоп-слов позволяют сократить шум и выделить ключевые термины. Например, в позитивных комментариях часто встречаются слова «спасибо», «отлично», а в негативных — «проблема», «критика».

Кластеризация текстовых данных в социологии имеет свои особенности. Тексты из социальных сетей содержат сарказм, неформальную лексику и эмоциональную окраску, что требует устойчивых алгоритмов. HDBSCAN и UMAP хорошо справляются с такими задачами, так как учитывают плотность дан-ных и шум. В сравнении с другими подходами, такими как Latent Dirichlet Allocation [9], HDBSCAN и UMAP не требуют предположений о распределении данных, что делает их более гибкими. Пример применения HDBSCAN и UMAP: анализ отзывов пользователей о продукте позволяет выделить группы с позитивными, негативными и ней-нии данных, что делает их более гибкими.

Пример применения HDBSCAN и UMAP: анализ отзывов пользователей о продукте позволяет выделить группы с позитивными, негативными и нейтральными отзывами. Это полезно для маркетинговых исследований и сегментации аудитории. В социологии такие методы помогают изучать общественное мнение, например, реакцию на политические события.

Для углубления анализа рассмотрены ограничения каждого метода. K-средние плохо работают с неравномерной плотностью данных, DBSCAN чувствителен к параметрам, а иерархическая кластеризация требует больших вычислений. HDBSCAN и UMAP решают эти проблемы, но требуют тщательной

настройки параметров, таких как min_cluster_size для HDBSCAN и n_neighbors для UMAP. Также рассмотрены альтернативные подходы, такие как кластеризация на основе графов, но они менее эффективны для текстовых данных.

Анализ текстовых данных требует учёта их семантической структуры. Например, слова с одинаковым написанием, но разным смыслом (омонимы) могут влиять на качество кластеризации. TF-IDF частично решает эту проблему, учитывая контекст, но более сложные методы, такие как word embeddings, могут улучшить результаты. Однако для русскоязычных текстов TF-IDF остаётся эффективным из-за простоты и доступности.

Вторая глава «Разработка метода кластеризации текстовых данных» описывает Разработанный метод предназначен для кластеризации текстовых социологических данных, таких как комментарии в социальных сетях, с использованием HDBSCAN и UMAP. Он включает несколько этапов: сбор и предобработку данных, векторизацию, снижение размерности, кластеризацию и анализ результатов.

Сбор и предобработка данных. Корпус из 1200 комментариев собран из открытых источников (публичные группы в социальных сетях). Данные анонимизированы и разделены по тональности: 400 позитивных, 400 негативных и 400 нейтральных. Предобработка включала:

- Удаление знаков препинания, эмодзи и специальных символов с помощью регулярных выражений.

Удаление стоп-слов (союзы, предлоги) с использованием библиотеки stop-words.

Лемматизация с помощью pymystem3 для приведения слов к начальной форме.

Например, комментарий «Спасибо, всё отлично!» после предобработки превращается в

«спасибо отлично». Это позволяет сократить шум и выделить ключевые слова.

Векторизация текстов. Для преобразования текстов в числовые векторы ис-пользован метод TF-IDF [7] с библиотекой scikit-learn [8]. Параметры:

Максимум 5000 признаков.

Исключение слов, встречающихся реже чем в 5 документах.

Исключение слов, присутствующих более чем в 80% документов.

TF-IDF учитывает частоту слов в документе и их редкость в корпусе, что позволяет выделить ключевые термины, такие как «благодарность» в позитивных комментариях или «критика» в негативных. Например, слово «проблема» имеет высокий вес в негативных текстах, но низкий в позитивных.

Снижение размерности. UMAP [6] применён для преобразования векторов

TF-IDF в двумерное пространство. Параметры:

n_neighbors=15 (число соседей, влияющее на локальную структуру).

min_dist=0.1 (минимальное расстояние между точками в проекции).

metric=’cosine’ (косинусное расстояние, подходящее для текстов).

UMAP позволил визуализировать данные и подготовить их для кластеризации. Косинусное расстояние выбрано из-за его эффективности для разреженных текстовых векторов, где большинство элементов равно нулю.

Кластеризация. HDBSCAN [5] применён с параметрами:

min_cluster_size=10 (минимальный размер кластера).

min_samples=5 (минимальное число точек для формирования плотного региона).

Алгоритм автоматически определил число кластеров и выделил шумовые точки, что важно для текстов с неформальной лексикой и выбросами, такими как спам или нерелевантные сообщения.

Анализ результатов. Результаты кластеризации проанализированы с помощью силуэтного коэффициента и визуального осмотра двумерных проекций. Метод протестирован на подмножестве данных (200 комментариев) для проверки устойчивости. Проведён анализ чувствительности параметров:

Для UMAP варьировалось n_neighbors от 5 до 30.

Для HDBSCAN варьировался min_cluster_size от 5 до 20.

Оптимальные параметры выбраны на основе максимального силуэтного коэффициента (0.52 при n_neighbors=15 и min_cluster_size=10).

Реализация выполнена на Python с использованием библиотек:

scikit-learn для TF-IDF и метрик.

hdbscan для кластеризации.

umap-learn для снижения размерности.

Метод сравнивался с другими подходами, такими как k-средние и DBSCAN, чтобы показать его преимущества. K-средние требуют задания числа кластеров, что не подходит для данных с неизвестной структурой. DBSCAN чувствителен к параметрам, что затрудняет его настройку. HDBSCAN и UMAP решают эти проблемы, обеспечивая автоматическое определение кластеров и качественную визуализацию.

Пример применения метода: анализ комментариев о политическом событии позволил выделить группы пользователей с разными точками зрения (под-держка, критика, нейтральные обсуждения). Это демонстрирует практическую ценность метода для социологических исследований. Например, в кластере критики были выделены подтемы, такие как «экономические проблемы» и «политическая нестабильность», что подтверждает способность метода выявлять детали.

Ограничения метода включают:

Высокую вычислительную сложность для больших корпусов (более 10,000 комментариев).

Зависимость от качества предобработки (ошибки в лемматизации могут повлиять на результаты).

Для углубления анализа рассмотрены альтернативные подходы к векторизации, такие как word embeddings (Word2Vec, BERT), но TF-IDF выбран из-за простоты и эффективности для русскоязычных текстов. Также обсуждены способы улучшения метода, например использование других метрик расстояния (евклидово, манхэттенское) или интеграция с тематическим моделированием. Например, комбинация HDBSCAN с LDA могла бы улучшить интерпретацию кластеров, но потребовала бы дополнительных ресурсов.

Метод протестирован на различных подмножествах данных, чтобы оценить его устойчивость к изменениям в составе корпуса. Например, удаление 10% случайных комментариев не изменило структуру основных кластеров, что подтверждает надёжность подхода. Также рассмотрены сценарии применения метода в реальных задачах, таких как мониторинг общественного мнения или анализ потребительских предпочтений

Третья глава «Экспериментальное исследование и анализ результатов» рассматривает проводимый эксперимент проводился на корпусе из 1200 комментариев, разделённых по тональности: 400 позитивных, 400 негативных и 400 нейтральных. После предобработки (удаление стоп-слов, лемматизация) и векторизации (TF-IDF) данные были преобразованы с помощью UMAP в двумерное пространство, а затем кластеризованы с использованием HDBSCAN.

Результаты:

Выделено 5 основных кластеров, соответствующих темам:

1.Позитивные отзывы (благодарности, поддержка).

2.Критика (политические и социальные темы).

3.Нейтральные обсуждения (факты, новости).

4.Эмоциональные жалобы (личные проблемы).

5.Информационные сообщения (объявления, вопросы).

Силуэтный коэффициент составил 0.52, что указывает на хорошее качество кластеризации.

Индекс Дэвиса-Болдина составил 1.45, что подтверждает компактность и разделимость кластеров.

Сравнение с другими методами:

K-средние (k=5): показали смешивание кластеров из-за шума. Силуэт-ный коэффициент — 0.38.

DBSCAN: выделил 4 кластера, но многие точки были отнесены к шуму из-за

чувствительности к параметрам. Силуэтный коэффициент — 0.45.

Иерархическая кластеризация: создала 12 мелких групп, что усложнило интерпретацию. Силуэтный коэффициент — 0.40.

HDBSCAN и UMAP продемонстрировали устойчивость к шуму и способность выделять подтемы, например, политическую критику в негативной группе. Результаты интерпретируемы: позитивные кластеры содержали слова «спасибо», «отлично», негативные — «проблема», «критика», нейтральные — «новости», «информация». Это подтверждает применимость метода для социологических исследований.

Время выполнения: HDBSCAN и UMAP завершили обработку за 12.5 секунд, k-средние — за 8.2 секунды, DBSCAN — за 10.3 секунды, иерархическая кластеризация — за 15.7 секунд. HDBSCAN и UMAP показали лучшее соотношение качества и скорости.

Ограничения метода:

Высокая вычислительная сложность для больших корпусов.

Зависимость от качества предобработки.

Для углубления анализа рассмотрены потенциальные улучшения, такие как использование предварительно обученных моделей (BERT) для векторизации или интеграция с другими методами кластеризации.

Например, BERT мог бы улучшить обработку семантической близости, но требует больше ресурсов. Также обсуждены сценарии применения метода в реальных задачах, таких как анализ реакции на социальные кампании

ЗАКЛЮЧЕНИЕ

Разработанный метод кластеризации на основе HDBSCAN и UMAP успешно решает задачу группировки текстовых социологических данных без предварительного задания числа кластеров. Эксперимент на корпусе из 1200 комментариев подтвердил высокое качество разделения (силуэтный коэффициент 0.52) и интерпретируемость результатов. Метод превосходит традиционные подходы (k-средние, DBSCAN, иерархическая кластеризация) по устойчивости к шуму и точности группировки. Полученные результаты могут быть использованы для анализа общественного мнения, сегментации аудитории и выявления социальных трендов. В дальнейшем метод, может быть, адаптирован для других типов данных, таких как опросы или транскрипты интервью.