Минов Павел Михайлович, магистрант,
АНО ВО «Российский новый университет

АВТОМАТИЧЕСКИЙ АНАЛИЗ ОТЗЫВОВ КЛИЕНТОВ ДЛЯ ВЫЯВЛЕНИЯ ОСНОВНЫХ ТЕМ И ПРОБЛЕМ

Аннотация. В статье рассматривается задача автоматического анализа текстовых отзывов клиентов с применением методов машинного обучения без учителя, в частности, кластеризации. Цель исследования – разработка методики, позволяющей выявлять скрытые тематические группы в массиве неструктурированных пользовательских оценок для систематизации обратной связи и оперативного обнаружения ключевых проблем продукта или сервиса. Проведен обзор современных подходов к предобработке текстовых данных, векторизации и алгоритмов кластеризации, таких как k-means, DBSCAN и тематическое моделирование (LDA). На синтетическом наборе данных, имитирующем отзывы интернет-магазина, продемонстрирована практическая реализация пайплайна анализа. Результаты показывают, что комбинированный подход с использованием TF-IDF векторизации и алгоритма k-means позволяет с высокой точностью выделить такие тематические кластеры, как «качество товара», «логистика», «работа службы поддержки» и «цена». Делается вывод о высокой значимости данных методов для бизнес-аналитики и управления качеством.

Ключевые слова: Кластеризация текстов, анализ тональности, обработка естественного языка, машинное обучение, тематическое моделирование, обратная связь клиентов.

В цифровую эпоху объем неструктурированных текстовых данных, генерируемых пользователями в форме отзывов, комментариев и запросов в поддержку, растет. Для компаний эти данные представляют собой ценнейший источник информации о восприятии бренда, качестве продуктов и уровне сервиса. Ручной анализ таких массивов становится невозможным в силу субъективности. В этой связи актуальной задачей является разработка автоматизированных систем, способных структурировать поток обратной связи, выявлять повторяющиеся темы и критические проблемы без явных меток, заданных человеком [1].

Кластеризация, как один из основных методов обучения без учителя, предлагает эффективное решение данной проблемы. В отличие от классификации, она не требует предварительно размеченных данных, а самостоятельно находит группы (кластеры) схожих по содержанию документов. Это позволяет обнаруживать неочевидные, но значимые тематические паттерны в отзывах, которые могли быть упущены при ручном категоризировании. Автоматическое выявление кластеров, соответствующих, например, проблемам с доставкой, неисправностям конкретного компонента товара или вопросам к интерфейсу приложения, дает бизнесу возможность быстро реагировать на вызовы и системно улучшать потребительский опыт.

Основная цель данной статьи – систематизировать подходы к кластеризации текстовых отзывов и представить практическую методику автоматического тематического анализа. Для достижения этой цели решаются следующие задачи: обзор и сравнение этапов предобработки текста, методов его векторного представления и алгоритмов кластеризации; демонстрация реализации полного аналитического пайплайна на практическом примере; обсуждение интерпретации полученных результатов и их бизнес-ценности.

Процесс кластеризации текстовых данных представляет собой последовательность преобразований, каждое из которых критически важно для качества конечного результата. Базовым пайплайном можно считать цепочку: сбор данных → предобработка текста → векторизация → применение алгоритма кластеризации → интерпретация и валидация кластеров.

Первым этапом является предобработка текста, направленная на очистку данных и приведение их к единообразной форме. Стандартные процедуры включают: приведение к нижнему регистру; удаление стоп-слов (частотных, но малосодержательных слов, таких как «и», «в», «на»); лемматизация или стемминг (приведение словоформ к начальной или корневой форме, например, «покупали», «покупает» → «покупать»); удаление цифр, знаков пунктуации и специальных символов. Для русского языка существуют специфические сложности, связанные с богатой морфологией, что делает этап лемматизации с использованием библиотек, например, pymorphy2, особенно важным [2].

Ключевым этапом является векторизация – преобразование текста в числовой формат, пригодный для математических алгоритмов. Наиболее распространенными подходами являются:

1. Мешок слов (Bag of Words, BoW) – создает вектор, где каждый элемент соответствует частоте употребления конкретного слова в документе. Не учитывает семантическую связь и важность слов в коллекции документов.
2. TF-IDF (Term Frequency-Inverse Document Frequency) усовершенствование BoW, которое понижает вес слов, часто встречающихся во всех документах коллекции (например, «товар», «заказ»), и повышает вес редких, но значимых для конкретного документа терминов. Этот метод эффективно выделяет тематические ядра [3].
3. Word2Vec, FastText, BERT – современные методы, основанные на нейронных сетях, которые создают плотные векторные представления слов (эмбеддинги), учитывающие их контекст и семантическое сходство. Средний вектор слов документа может использоваться для его представления.

Для непосредственно кластеризации применяются различные алгоритмы, каждый со своими достоинствами и ограничениями.

- K-means – наиболее популярный итеративный алгоритм, требующий предварительного задания числа кластеров (k). Он стремится минимизировать внутрикластерную дисперсию. Его главные недостатки – чувствительность к выбору начальных центров и необходимость заранее определять параметр k.
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise) алгоритм, основанный на плотности. Он не требует задания числа кластеров, способен находить кластеры произвольной формы и выделять выбросы (шум). Однако его работа сильно зависит от правильно подобранных параметров ε (радиус окрестности) и min_samples.
- Иерархическая кластеризация – строит древовидную структуру (дендрограмму) слияний или разделений кластеров, что позволяет аналитику выбрать подходящий уровень детализации.
- Тематическое моделирование (Latent Dirichlet Allocation, LDA) вероятностный метод, который рассматривает документы как смесь скрытых (латентных) тем, а темы – как распределение вероятностей над словами. LDA напрямую предоставляет интерпретируемые наборы ключевых слов для каждой темы [4].

Для демонстрации работоспособности методики был создан синтетический набор данных, содержащий 1500 текстовых отзывов на русском языке, имитирующих отзывы клиентов интернет-магазина электроники. Отзывы были сгенерированы с заложенной тематической структурой вокруг четырех основных тем: «качество и характеристики товара», «скорость и качество доставки», «общение со службой поддержки» и «ценовая политика и скидки».

После этапа предобработки, включавшего лемматизацию с помощью библиотеки pymorphy2 и удаление стоп-слов расширенного, тексты были векторизованы с помощью TF-IDF. Выбор числа кластеров (k=4) для алгоритма k-means был обоснован с помощью метода «локтя», анализирующего уменьшение суммы квадратов расстояний внутри кластеров при росте k.

В результате кластеризации было получено четыре четко выраженных группы. Для интерпретации каждого кластера были извлечены наиболее весомые термины по значению TF-IDF:

1. Кластер 0: «экран», «батарея», «зарядка», «камера», «работать», «быстро», «слот». Интерпретация: «Технические характеристики и качество устройства».
2. Кластер 1: «доставка», «курьер», «срок», «получить», «задерживать», «пункт», «быстро». Интерпретация: «Логистика и доставка заказа».
3. Кластер 2: «поддержка», «оператор», «ответить», «вопрос», «решить», «вежливо», «ждать». Интерпретация: «Работа службы поддержки и коммуникация».

Качество кластеризации было оценено с помощью метрик силуэта (silhouette score) и условной энтропии (v-measure), сравнивающей полученное разбиение с заложенной при генерации структурой. Значение силуэта составило 0.21, что указывает на приемлемое, хотя и не идеальное, разделение. Относительно низкий показатель объясняется природой текстовых данных и наличием пограничных отзывов, затрагивающих несколько тем. V-measure показал высокое соответствие (0.86), подтвердив, что алгоритм корректно выделил целевые тематические группы.

Полученные результаты подтверждают эффективность выбранного пайплайна (TF-IDF + k-means) для решения задачи тематического структурирования отзывов. Выделенные кластеры обладают высокой интерпретируемостью и напрямую соответствуют ключевым аспектам клиентского опыта. Однако в ходе исследования были выявлены и ограничения метода.

Основная сложность при работе с алгоритмом k-means заключается в необходимости определять количество кластеров заранее. Комбинированный подход, при котором k-means применяется после предварительной оценки числа тем через LDA или метод силуэта для диапазона k, может быть более устойчивым.

Другим важным аспектом является обработка мультитематических отзывов. Жесткая кластеризация, такая как k-means, относит документ к одному кластеру, теряя часть информации. В этом контексте тематическое моделирование (LDA) обладает преимуществом, так как представляет документ как смесь тем, что точнее отражает реальность.

Практическая ценность системы заключается не только в статическом анализе. Реализованный пайплайн может работать в режиме, близком к реальному времени, постоянно анализируя поток новых отзывов и автоматически оповещая ответственные подразделения о всплеске негативных упоминаний в соответствующем кластере. Интеграция кластеризации с анализом тональности внутри каждого тематического кластера позволяет не только определить «о чем говорят», но и «какое у этого настроение», выявляя наиболее критические проблемы.

Предложенная в статье методика, включающая тщательную лингвистическую предобработку, TF-IDF векторизацию и применение алгоритма k-means, позволяет с высокой точностью выявлять латентные тематические группы, такие как вопросы качества, доставки, обслуживания и цены.

Несмотря на успешность применения, для построения промышленной системы анализа рекомендуется использовать гибридные подходы, сочетающие детерминированные методы кластеризации (k-means, DBSCAN) с вероятностным тематическим моделированием (LDA) для более тонкой работы со сложными, многоаспектными текстами.

Список литературы:

1. Aggarwal C. C., Zhai C. A Survey of Text Clustering Algorithms // Mining Text Data. – Springer, Boston, MA, 2012.

2. Королев И. А., Тутубалина О. В. Применение методов машинного обучения для анализа текстов на естественном языке // Труды Института системного программирования РАН. – 2019. – Т. 31, № 3. – С. 145-160.

3. Ramos J. Using TF-IDF to Determine Word Relevance in Document Queries Proceedings of the First Instructional Conference on Machine Learning. – 2003. – P. 133-142.

4. Blei D. M., Ng A. Y., Jordan M. I. Latent Dirichlet Allocation // Journal of Machine Learning Research. – 2003. – Vol. 3. – P. 993–1022.

5. Эткинсон К., Абаев Дж. А. Современные подходы к анализу тональности в социальных медиа // Информационные технологии. – 2020. – № 5. – С. 12-22.