К. С. Хилиманюк, И. Ю. Карякин
Тюменский государственный университет, г. Тюмень
УДК 004.8

СЕМАНТИЧЕСКИЙ АНАЛИЗ ОТЗЫВОВ МАРКЕТПЛЕЙСОВ С ПРИМЕНЕНИЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Аннотация. В статье представлен поэтапный алгоритм и набор Python библиотек для текстового анализа отзывов на маркетплейсах. В рамках каждого этапа кратко описывается каждый алгоритм. Результатом является модель, способная классифицировать содержательные и бессодержательные отзывы для дальнейшего анализа и поиска ключевых характеристик.

Ключевые слова: анализ отзывов на маркетплейсах; предварительная обработка данных; анализ текстов; алгоритмы текстовой классификации; машинное обучение.

Введение. По данным исследований рынок интернет-торговли в России демонстрирует устойчивый рост [1]. В структуре рынка все более значимое место занимают крупнейшие универсальные маркетплейсы — Wildberries, Ozon, Яндекс Маркет и др.

Одной из основных проблем, отмечаемых продавцами, является неправильный выбор товара [2]. Для решения данной проблемы существует широкий спектр приложений по аналитике рынка [3].

Однако необходимо отметить, что в большинстве из этих приложений специфика рынка интернет-торговли не учитывается. Те немногие специфические данные, такие как отзывы, анализируются с точки зрения рейтинга (количество "звезд" на маркетплейсе) или количества, а сам текст отзывов либо игнорируется, либо просматривается экспертами вручную.

В научной литературе существует ряд статей по теме анализа отзывов в других предметных областях [4-8]. Но целью авторов этих статей является анализ тональности и выявление положительных и отрицательных аспектов в отзывах потребителей. Анализ тональности не обеспечит выявление качественных характеристик товаров.

Проблема исследования. Существующие приложения по аналитике рынка на маркетплейсах не учитывают текстовую информацию, содержащуюся в отзывах.

Целью данной работы является подбор алгоритмов машинного обучения для семантического анализа отзывов на маркетплейсах. Для этого необходимо изучить существующие алгоритмы предобработки и алгоритмы классификации текстов.

Материалы и методы. Текстовый анализ начинается с предварительной обработки данных, для которой существует множество алгоритмов [9-10].

Предобработку текстов можно разделить на несколько этапов. Основными являются: очистка данных, извлечение признаков и уменьшение размерности.

По результатам предобработки набор текстов трансформируется в набор векторов, которые можно использовать для обучения модели классификации текстов.

Существует множество алгоритмов классификации [11]. Одними из самых распространенных являются К-ближайших соседей и Наивный байесовский классификатор. Они легки для понимания и реализации.

Также существуют более сложные алгоритмы, такие как Метод опорных векторов, Дерево решений и Случайный лес. Каждый из этих алгоритмов имеет свои достоинства для определенных наборов данных.

Результаты. Проведен анализ алгоритмов предобработки и классификации текстов для семантического анализа отзывов на маркетплейсах. По итогу, выбраны следующие решения:

1. Для очистки данных был применен алгоритм лемматизации [11] (выделения основы слова), были удалены все стоп-слова [11] (предлоги, союзы, частицы и т. д.), пунктуация и повторы.

Также, для дальнейшего анализа, все отзывы были разбиты на синтаксические конструкции, так как один и тот же отзыв может одновременно содержать как полезную для исследования часть, так и бесполезную. Для этого был проведен морфологический разбор каждого слова (выделение части речи, рода, числа, падежа и т. д.) и синтаксический анализ каждого предложения (разбиение предложения на подлежащее, сказуемое, дополнение и т. д.). Все алгоритмы, описанные в данном пункте, применялись с помощью библиотеки

Python под названием «Natasha» [12]. Результат можно видеть на Рис. 1.

	text	rel	pos	lemma	syntax	syntax_lemmas
0	Все	Определение	Местоимение	весь	NaN	NaN
1	отверстия	Подлежащее	Существительное	отверстие	NaN	NaN
2	с	Предлог	Предлог	с	NaN	NaN
3	телевизором	Дополнение	Существительное	телевизор	NaN	NaN
4	совпадают	Сказуемое	Глагол	совпадать	Все отверстия с телевизором совпадают.	[весь, отверстие, телевизор, совпадать]
5	и	Союз	Союз	и	NaN	NaN
6	шайбы	Дополнение	Существительное	шайба	NaN	NaN
7	не	Частица	Частица	не	NaN	NaN
8	вываливаются	Сказуемое	Глагол	вываливаться	и шайбы не вываливаются	[шайба, вываливаться]
9	.	Пунктуация	Пунктуация	.	NaN	NaN

Рис. 1. Результат предобработки с помощью библиотеки Natasha

2. Для извлечения признаков и составления набора векторов был использован метод TF-IDF. Метод основан на подсчете количества слов в каждом тексте и присвоении его пространству признаков всех текстов обучающей выборки. Метод хорошо показывает себя в преобразовании текстовой информации в набор векторов, которые в дальнейшем напрямую вносятся в модель машинного обучения.

Для применения метода TF-IDF была использована библиотека Python «sklearn.feature_extraction.text» [11].

3. Для уменьшения размерности был использован метод главных компонент. Метод определяет подпространство, в котором примерно лежат данные. Таким образом происходит поиск новых некоррелированных переменных и максимизация дисперсии, чтобы получить наибольшую вариативность признаков. Данный метод прост для понимания и позволяет сократить размерность за счет исключения слов с высокой корреляцией.

Для уменьшения размерности использовалась библиотека Python «sklearn.decomposition» [11].

4. Для классификации текстов было проанализировано множество алгоритмов. Результаты анализа представлены на Рис. 2. Наилучший показатель в точности в 86% получился у метода Случайных лесов. Причиной является устойчивость метода к переобучению. Одной из основных проблем при анализе является дисбаланс содержательных и бессодержательных отзывов. Так как, зачастую,

пользователи оставляют бессодержательные отзывы, никак не описывающие достоинства и недостатки товаров или какие-либо ключевые характеристики.

Classifier	rank_accuracy	rank_f1	rank_precision	accuracy	f1	precision
Случайный лес	1	1	2	0.865	0.708	0.788
Дерево решений	2	3	4	0.852	0.693	0.733
Метод опорных векторов	3	2	3	0.851	0.694	0.760
K-ближайших соседей	4	5	1	0.823	0.569	0.790
Метод ближайшего центроида	5	4	5	0.816	0.671	0.648

Рис. 2. Метрики сравниваемых моделей

Для обучения модели и ее использования была применена библиотека Python «sklearn.ensemble» [11].

Заключение. Торговля на маркетплейсах является перспективной нишей для продажи товаров. Для увеличения конкурентоспособности дилеров существуют приложения-аналитики, упускающие весомый пласт информации в виде содержимого отзывов.

Для анализа отзывов можно использовать алгоритмы предобработки и классификации текстов.

В процессе предобработки можно разбить отзывы на синтаксические конструкции, очистить данные от повторов, стоп-слов и пунктуации, выделить признаки и сократить размерность.

С помощью модели можно с 86%-ой точностью отсортировать содержательные отзывы от бессодержательных. Таким образом, в дальнейшем, можно будет выявлять качественные характеристики товаров, их положительные и отрицательные аспекты с точки зрения пользователей.

В дальнейшем можно добавить некоторые этапы предобработки или изменить некоторые алгоритмы, также можно пробовать изменять параметры самой модели для разных групп товаров. В данной статье приведен лишь один из возможных способов семантического анализа отзывов на маркетплейсах с применением Искусственного Интеллекта.

СПИСОК ЛИТЕРАТУРЫ

1. Интернет-торговля рынок России (2023) — Текст: электронный // TAdviser: [сайт]. — URL: https://clck.ru/34RmTg (дата обращения: 28.05.2023).
2. 7 типичных ошибок новичков при выходе на маркетплейсы — Текст: электронный // Тинькофф Банк — Бизнес-секреты: [сайт]. — URL: https://secrets.tinkoff.ru/biznes-s-nulya/oshibki-sellera-kak-prodavat-na-marketpleysah/ (дата обращения: 28.05.2023).
3. Аналитика маркетплейсов: 31 сервис для внешнего и внутреннего анализа — Текст: электронный // moysklad.ru: [сайт]. — URL: https://www.moysklad.ru/poleznoe/marketplejsy/analitika-marketpleysov/ (дата обращения: 28.05.2023).
4. Володченко А. О. Повышение конкурентоспособности гостиничного бизнеса с помощью семантического анализа текстовых отзывов / А. О. Володченко, В. В. Комраков — Текст : электронный // Новые математические методы и компьютерные технологии в проектировании, производстве и научных исследованиях : материалы XXIII Республиканской научной конференции студентов и аспирантов, Гомель, 23-25 марта 2020 г. — Гомель: Гомельский государственный университет им. Франциска Скорины, 2020. — URL: https://elibrary.ru/download/elibrary_44879849_29316384.pdf (дата обращения: 28.05.2023).
5. Бойко М. В. Исследование удовлетворенности потребителей в банковской сфере на основе анализа текстовых отзывов / М. В. Бойко. — Текст : электронный // Информационные технологии интеллектуальной поддержки принятия решений : Proceedings of the 2nd International Conference “Information Technologies for Intelligent Decision Making Support” and the Intended International Workshop “Robots and Robotic Systems”, Уфа, 18–21 мая 2014 года / General Program Chair: Guzairov Murat (USATU, Ufa, Russia); General Chair Woman: Yusupova Nafisa (USATU, Ufa, Russia). Т. 3. — Уфа: ГОУ ВПО «Уфимский государственный авиационный технический университет», 2014. — URL: https://elibrary.ru/download/elibrary_25084028_15003945.pdf (дата обращения: 28.05.2023).
6. Герасименко Е. М. Анализ тональности текстовых отзывов с применением тональных словарей и кардинальности нечеткого множества / Е. М. Герасименко, В. В. Стеценко — Текст : электронный // Известия ЮФУ. Технические науки. — 2022. — № 5(229). — URL: https://elibrary.ru/download/elibrary_50072308_73120248.pdf (дата обращения: 28.05.2023).

7. Кузнецов Т. А. Создание системы автоматической классификации текстовых отзывов на русском языке с помощью машинного обучения / Т. А. Кузнецов, С. И. Гавриленков — Текст : электронный // Политехнический молодежный журнал. — 2022. — № 5(70). — DOI 10.18698/2541-8009-2022-5-794. — EDN SDLCZH. — URL: https://elibrary.ru/download/elibrary_48762870_87269265.pdf (дата обращения: 28.05.2023).
8. Конышев Е. В. Методика изучения ментального туристско-рекреационного пространства по отзывам туристов (на примере Кировской области) / Е. В. Конышев — Текст : электронный // Вестник Московского университета. Серия 5: География. — 2022. — № 5. URL: https://elibrary.ru/download/elibrary_49621934_63324563.pdf (дата обращения: 28.05.2023).
9. Акимов А. А. Предварительная обработка данных для машинного обучения / А. А. Акимов, Д. Р. Валитов, А. И. Кубряк. — Текст : электронный // Научное обозрение. Технические науки. — 2022. — № 2. — URL: https://www.elibrary.ru/download/elibrary_48411861_16222054.pdf (дата обращения: 28.05.2023).
10. Татур М. М. «Сырые» данные и некоторые рецепты их «приготовления» / М. М. Татур, В. М. Проровский, Д. В. Куприянова, И. Н. Носырев. — Текст : электронный // Информационные системы и технологии : материалы международного научного конгресса по информатике. — Минск: Белорусский государственный университет. — URL: https://www.elibrary.ru/download/elibrary_49732158_13093697.pdf (дата обращения: 28.05.2023).
11. Text Classification Algorithms: A Survey / Kamran Kowsari, K. J. (2019). — Текст: электронный // github.com: [сайт]. — URL: https://github.com/kk7nc/Text_Classification#term-frequency-inverse-document-frequency (дата обращения 20.11.2023).
12. Лаборатория анализа данных Александра Кукушкина: [сайт]. — URL: https://natasha.github.io/ (дата обращения 20.11.2023). — Текст: электронный.