Научная статья

УДК 81.33, 81.322.2
DOI 10.25205/1818-7935-2022-20-4-90-106

Об одном подходе к автоматической суммаризации потребительских отзывов

Надежда Сергеевна Чечнева

Санкт-Петербургский государственный университет
Санкт-Петербург, Россия

chechnevanadegda@mail.ru, https://orcid.org/0000-0003-1987-0244

Аннотация

Потребительские отзывы являются важным аспектом электронной коммерции, они влияют на коммуникативное поведение потенциального адресата (покупателей). Пользователи часто читают их, чтобы получить общее либо конкретное представление о продукте или услуге. Компании анализируют отзывы клиентов для усовершенствования своих продуктов и/или для корректировки маркетинговой стратегии. Однако из-за слабой структурированности отзывов и быстро возрастающего количества, их изучение становится трудоемким процессом как для пользователей, так и для компаний. Вследствие этого повышается актуальность проблемы автоматической суммаризации потребительских отзывов. В данной статье предлагается новый подход к решению этой задачи, разработанный на основе проведенного и апробированного исследования отзывов о цифровой и бытовой технике. Он позволяет структурировать и расположить тексты отзывов в соответствии с тематической иерархией аспектов, предоставить сводную информацию об их тональности (количестве положительных и отрицательных упоминаний тех или иных аспектов), а также показать наиболее релевантные предложения. Исследование строится на материале текстов отзывов, собранных с интернет-ресурса «Яндекс.Маркет». Подробно описан процесс суммаризации отзывов, включающий несколько этапов: 1) экспертное формирование перечня тематических классов аспектных терминов; 2) классификация предложений по заданным классам аспектов; 3) распределение предложений на два класса – положительные и отрицательные – с подсчетом количества положительных и отрицательных предложений внутри каждого класса аспектов; 4) этап ранжирования предложений; 5) этап визуализации результатов, полученных на предыдущих шагах. Качество работы алгоритма по созданию резюме из большой коллекции отзывов было протестировано на пяти моделях товаров из следующих категорий: кофемашины, роботы-пылесосы, электронные книги, телевизоры, стиральные машины.

Ключевые слова

компьютерная лингвистика, машинное обучение, суммаризация отзывов, анализ тональности, векторизация предложений, ранжирование предложений

Благодарности

Мы благодарим «Яндекс.Маркет» и «YM Сканнер» за материалы для исследования.

Для цитирования

Чечнева Н. С. Об одном подходе к автоматической суммаризации потребительских отзывов // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2022. Т. 20, № 4. С. 90–106. DOI 10.25205/1818-7935-2022-20-4-90-106

An Approach to Summarizing Product Reviews

Nadezhda S. Chechneva

Saint Petersburg State University
Saint Petersburg, Russia

chechnevanadegda@mail.ru, https://orcid.org/0000-0003-1987-0244

Abstract

Product reviews are an important feature of e-commerce because they influence the communicative behavior of a potential addressee (customers). Users often read online product reviews to get either general or specific information about a product or service. Besides, companies analyze customer reviews to improve their product quality or to adjust their marketing strategy. However, many reviews are unstructured and long. With the number of product reviews growing rapidly, reading a large number of reviews becomes a time-consuming process for both users and companies. Therefore, review summarization becomes a serious issue. In this paper, we propose a new approach to summarizing reviews of electronics and household appliances. The proposed approach makes it possible to structure information for every aspect category; it provides calculation sentiment score for each aspect category, and shows the most relevant sentences for each aspect. We used product reviews from Yandex.Market as target data. Our task was performed in five main phases: 1) expert identification of thematic aspect categories; 2) classifying sentences into the predefined aspect categories; 3) sentence classification into two classes—positive and negative—with calculating the number of both type sentences within each aspect category; 4) sentence ranking; 5) visualization of the results obtained in the previous phases. The quality of the algorithm for creating a resume from a large collection of reviews has been tested on five models of products from the following categories: coffee machines, robot vacuum cleaners, e-books, TV-sets, and washing machines.

Keywords

computational linguistics, machine learning, review summarization, sentiment analysis, sentence embeddings, sentence ranking

Acknowledgements

We are grateful to Yandex.Market and YM Scanner for the research materials.

For citation

Chechneva N. S. An Approach to Summarizing Product Reviews. *Vestnik NSU. Series: Linguistics and Intercultural Communication*, 2022, vol. 20, no. 4, pp. 90–106. (in Russ.) DOI 10.25205/1818-7935-2022-20-4-90-106

Введение

В условиях информационного общества сфера электронной коммерции с каждым годом расширяется. Одним из ее неотъемлемых атрибутов являются отзывы на продукцию, которые выступают в качестве проводника между пользователями и компаниями. В них пользователи делятся своим опытом использования товара, рассказывают, по каким критериям они его выбирали, указывают на достоинства и недостатки. Компании, в свою очередь, могут использовать отзывы как источник обратной связи от покупателей для усовершенствования своих продуктов или для корректировки стратегии. Отзывы богаты такого рода информацией, но сложность заключается в ее извлечении. Дело в том, что такие тексты обычно слабо структурированы. Кроме того, если какая-то модель товара популярна, то только на одну эту модель могут быть написаны десятки тысяч отзывов. Все это затрудняет получение необходимой информации как для пользователей, так и для компаний.

Вышеописанные трудности привели к возникновению такого направления, как суммаризация отзывов (product reviews summarizing). Первые работы по суммаризации отзывов появились примерно в то же время, когда зародилась новая область, получившая название «Анализ тональности» (Sentiment Analysis), то есть в начале 2000-х гг. В отличие от аспектного анализа тональности (aspect-based sentiment analysis), задача которого заключается в определении отношения к конкретному аспекту основного объекта, суммаризация отзывов предполагает некое обобщение мнений людей относительно аспектов на основании нескольких документов.

Задача суммаризации отзывов проистекает из традиционной задачи суммаризации текстов. Однако, согласно статье [Mithun, 2009], суммаризация субъективных документов[1] сложнее, чем суммаризация объективных документов. Так, в статье приводится сравнение суммаризации новостей и блогов, и результаты для блогов – субъективных текстов – хуже, чем для новостей. Это объясняется тем, что в субъективных текстах люди выражают мнения неформальным (то есть обиходно-разговорным) языком, отвлекаясь на темы, не имеющие, строго говоря, отношения к предмету обсуждения (хотя сам пишущий может быть уверен в обратном). Все это усложняет суммаризацию.

Примером суммаризации может послужить задача извлечения наиболее часто упоминаемых аспектов и вычисление их усредненных оценок. Возможен и другой вариант – автоматическое создание краткого резюме на основе множества отзывов на определенный товар. Реализуя подобные задачи, онлайн-сервисы значительно упрощают процесс ознакомления с отзывами. Так, сервис для выбора и покупки товаров «Яндекс.Маркет» реализовал алгоритм, который определяет, какие достоинства и недостатки данного товара упоминают чаще всего.

Таким образом, существует потребность в структурной организации отзывов. Одним из возможных вариантов решения этой задачи могло бы стать, с одной стороны, выявление перечня тематических классов аспектов верхнего уровня, присущих широкому кругу товаров из разных категорий, а с другой – детализация и адаптация полученной системы тем под конкретную категорию или модель товара. В данной работе предложен аспектно-ориентированный алгоритм суммаризации отзывов на товары из нескольких категорий.

1. Подходы к суммаризации отзывов

В работах по суммаризации отзывов можно выделить два подхода:

- *экстрактивный*, заключающийся в извлечении из исходных текстов наиболее «значимых» блоков;
- *абстрактивный*, целью которого является генерация нового текста, обобщающего исходные документы.

Сравнение экстрактивного и абстрактивного подходов к суммаризации дается в работе [Condori, 2017]. Мы же в своем обзоре сфокусируемся на экстрактивном, поскольку он наиболее близок к методологии нашего исследования.

Один из первых экстрактивных подходов к суммаризации отзывов, ставший к настоящему времени классическим, был предложен в работе [Hu, Liu, 2004]. Авторы ставят задачу суммаризации отзывов, мотивируя ее возрастанием количества онлайн-отзывов, что затрудняет их прочтение потенциальным покупателем, которому нужно принять взвешенное решение о покупке. Авторы подчеркивают, что от традиционной суммаризации текстов задача суммаризации отзывов отличается тем, что для последней значение имеют только те характеристики продукта, по которым потребители выразили свое мнение в той или иной форме. Данную задачу они решают в три этапа: 1) извлечение аспектов продукта, наиболее часто упоминаемых в отзывах; 2) идентификация положительных и отрицательных предложений, относящихся к каждому аспекту; 3) составление итогового резюме.

Результат такой суммаризации для цифровых камер они иллюстрируют в виде схемы, представленной на рисунке 1, которая основана на характеристиках.

В работе [Mukherjee, 2012] решается задача отнесения аспектных терминов к тематическим классам. Авторы предлагают подход, который предполагает описание предметной области в виде исходных слов для каждой категории аспектов. Так, категория STAFF описывается следующими исходными словами: staff, service, waiter, hospitality, upkeep. Авторы проводят эксперименты с разным количеством исходных слов, от двух до пяти, при этом общее число

[1] То есть документов, в которых обычно не соблюдаются нормы официально-деловой речи.

аспектных терминов в каждой категории – 30. В работе приводятся данные о том, как меняется точность соотнесения аспектов с категориями при разном количестве исходных слов.[2]

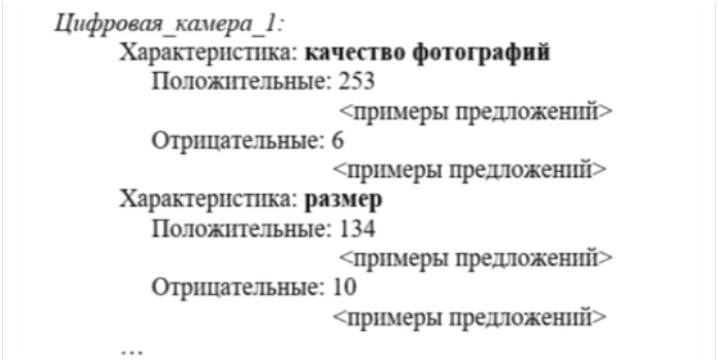

Рис. 1. Пример суммаризации
Fig. 1. Example of a summary

В работе [Zhai et al., 2011] основное внимание уделяется задаче кластеризации аспектных терминов. Поскольку один и тот же аспект продукта люди могут называть множеством разных слов, то целесообразно сгруппировать их в кластеры. Для решения этой задачи в статье используется алгоритм Expectation-Maximization.

В статье [Yu et al., 2011] авторы предлагают собственный подход к иерархическому представлению аспектов продукта. Он состоит из четырех этапов: 1) построение исходной иерархии аспектов товара на основе перечня характеристик товара, представленных в спецификации; 2) извлечение потенциальных аспектов из «достоинств» и «недостатков» отзывов; 3) подсчет семантической близости аспектов; 4) итоговая генерация иерархии аспектов.

У [Kim, 2013] также формулируется гипотеза о том, что отзывы содержат скрытую структуру аспектов, которые естественным образом могут быть организованы в иерархию, и каждый из узлов этой иерархии состоит из аспекта и связанных с ним оценок. Для подтверждения этой гипотезы они ставят задачу порождения такой иерархии. Для этого в работе используется непараметрический байесовский метод «Китайский ресторан».

В работе [Akhtar et al., 2017] представлен подход к суммаризации отзывов на отели. Авторы мотивируют актуальность исследования тем, что при бронировании отелей пользователи опираются на отзывы и рейтинги, однако у большинства туристов нет времени на прочтение всех отзывов, а рейтинги не дают полного представления об отелях. Поэтому авторы предлагают подход, включающий: 1) классификацию предложений из отзывов на десять предопределенных классов, полученных экспертным путем; 2) применение тематического моделирования LDA для выявления тем внутри каждого класса.

В рамках нашего исследования была выбрана реализация экстрактивного подхода. Мы исходили из того, что, во-первых, данный подход более популярен на сегодняшний день в электронной коммерции, во-вторых, такой подход относительно проще адаптировать к разным предметным областям, и, в-третьих, этот подход позволяет получить доступное для интерпретации структурированное резюме. При этом следует отметить, что абстрактивный подход имеет свои преимущества и является популярным и перспективным направлением исследований.

[2] В других терминах это фрейм и его базовые лексические компоненты, ситуативно и ассоциативно связанные с именем фрейма. – Прим. авт.

2. Предложенный подход к суммаризации отзывов и материал исследования

2.1. Характеристика и обоснование предложенного подхода

Мы предлагаем подход к суммаризации отзывов, который является модификацией классического экстрактивного подхода. Основное отличие заключается в том, что в качестве первого этапа происходит не извлечение наиболее упоминаемых аспектов отдельно для каждого товара, а экспертное установление перечня тематических классов аспектов для широкого класса товаров.

Из характеристики отзыва и огромного количества предметных областей само собой следует, что в данный момент невозможно выработать совершенно законченный и исчерпывающий перечень классов аспектов, подходящий для любого вида отзывов. Тем не менее было бы неверным утверждать, что у отзывов не существует общих тем. Так, интуитивно понятно, что отзывы на фильмы и на смартфоны будут иметь мало общих аспектов, а отзывы на ноутбуки и смартфоны будут иметь больше пересечений. Можно предположить, что потребитель в процессе написания отзыва опирается на определенные представления о качестве продукта, что позволяет выделить наиболее часто встречающиеся группы характеристик. Таким образом, составление тематических классов для всех видов отзывов представляется слишком обширной задачей, поэтому в нашем исследовании мы ограничиваемся конкретным, но достаточно широким классом товаров – бытовой и цифровой техникой.

В общем виде предложенная нами модель включает следующие этапы[3], представленные на рисунке 2.

Рис. 2. Этапы предложенного подхода к суммаризации отзывов
Fig. 2. Steps of the approach to product reviews summarization

2.2. Материал исследования и предобработка данных

В качестве материала для исследования мы использовали отзывы о цифровой и бытовой технике с сервиса для выбора и покупки изделий «Яндекс.Маркет».

Было отобрано пять моделей товаров, относящихся к разным категориям: кофеварки, роботы-пылесосы, стиральные машины, телевизоры, электронные книги.

Разметка отзывов проводилась экспертным путем. Процедура разметки включала следующие шаги:

1. Отзывы автоматически разбивались на предложения при помощи библиотеки nltk.
2. Предложения, не содержащие аспектов, на этапе аннотирования экспертом исключались, поскольку задачу извлечения предложений, содержащих мнения, мы не решаем.
3. Предложения, содержащие два и более аспекта, вручную разделялись (в тех случаях, где это было возможно) на разные предложения. Так, предложение *Отличное качество, прочная сборка* разделялось на два: *отличное качество* и *прочная сборка*. В тех случаях, когда пред-

[3] Исходный код и материалы исследования доступны по ссылке: https://github.com/nchechneva/summarizing-customer-reviews.

ложение нельзя было разделить без изменения синтаксической структуры (переписывания) или добавления новых слов, оно помечалось как содержащее несколько аспектов и в итоговый датасет не включалось.

4. Каждому предложению присваивалась метка класса аспекта из установленного перечня и метка тональности: положительная или отрицательная.

Итоговый размер тренировочной и тестовой коллекции данных для каждой модели товара представлен в таблице 1.

Таблица 1

Размер тренировочной и тестовой коллекции данных для каждой модели товара

Table 1

The Training and Test Dataset Size for Each Product

Товар	**training/test**	**Количество отзывов**	**Количество предложений**
Кофеварка рожковая VT-1518	test	204	596
	train	171	412
Робот-пылесос RV-R304	test	157	597
	train	147	409
Стиральная машина EW6S4R06W	test	173	530
	train	151	352
Телевизор 42PFL5405H	test	121	589
	train	114	402
Электронная книга PRS-T2 2 ГБ	test	117	599
	train	111	402
Итого		1 466	4 888

3. Характеристика этапов процесса суммаризации отзывов

3.1. Установление перечня тематических классов аспектов

Задача установления тематических классов аспектов довольно сложна, поскольку требования и ожидания относительно продуктов у всех пользователей разные, и, следовательно, спектр затрагиваемых аспектов чрезвычайно широк.

Помочь в выделении тематических классов аспектов, на наш взгляд, может обращение к концептуальному представлению рассматриваемого класса изделий, а именно понятия «техника». Согласно В. Ш. Рубашкину [2012, с. 221], объекты техники описываются тремя группами характеристик:

- Функциональные – описание предназначения изделия.
- Структурные – описание конструкции изделия, то есть его подсистем, модулей, деталей, узлов и их пространственной взаимосвязи.
- Специфические параметры, имеющие точное количественное выражение: мощность, номинальное напряжение питания и т. д.

Далее нам необходимо перейти к выделению классов аспектов, затрагиваемых в отзывах.

Вопрос о номенклатуре тематических классов аспектов можно отнести к малоизученным, однако здесь можно провести аналогию с близкой по своей сути проблематикой, обсуждавшейся задолго до появления анализа тональности. В экономике и маркетинге было проведено большое количество исследований в области анализа потребительского поведения. Начиная

с 1950-х годов широкое распространение получили мультиатрибутивные модели товара, основная идея которых заключалась в том, что покупатель формирует целостное отношение к товару по совокупности его отдельных характеристик. М. Розенберг [Rosenberg, 1956], М. Фишбейн [Fishbein, 1963] и К. Ланкастер [Lancaster, 1966] предложили модели количественной оценки отношения к продукту. В 1960-х годах получила распространение модель Ф. Котлера [Kotler, 1967], согласно которой в любом товаре можно выделить три уровня:

1. Товар по замыслу – основная функциональная характеристика товара, основная выгода для потребителя.

2. Товар в реальном исполнении – набор полезных для потребителя характеристик, его внешний вид, эргономические, эстетические свойства.

3. Товар с подкреплением – дополнительные признаки, включая послепродажное обслуживание, гарантию, сервис, доставку.

Затем стали исследовать атрибуты товаров, появляются работы о потребительском выборе как иерархическом процессе [Howard, 1977], о континууме атрибутов от конкретного к абстрактному и об отношениях между атрибутами [Johnson, 1984].

В плановой советской экономике осмысление процессов потребительского поведения носило рационально-прикладной характер. Это привело к возникновению различных ГОСТов, методических рекомендаций. Так, в 1979 году был разработан документ «Товары народного потребления. Выбор номенклатуры потребительских свойств и показателей качества. Основные положения» (РД 50-165-79). В нем выделяется шесть групп потребительских свойств, оказывающих влияние на восприятие качества товара: показатели социального назначения, функциональные свойства, надежность в потреблении (эксплуатации), эргономические, эстетические и экологические показатели, показатели безопасности.

Однако позднее модели рационального выбора подверглись критике со стороны экономистов-эмпириков, декларировавших иррациональность потребительского поведения при принятии решений [Tversky, Kahneman, 1974; Thaler, 1980]. По их мнению, люди зачастую принимают решения под воздействием информации, не связанной напрямую с потребительскими свойствами продуктов.

Результаты исследований в области потребительского поведения компании активно применяют для продвижения своих товаров, что отражается и в языке маркетинговых коммуникаций. В настоящее время появилось много работ, посвященных изучению языка рекламы и других «продающих текстов». Е. Г. Борисова предлагает выделять такие исследования в отдельную область лингвистики – маркетинговую лингвистику [Борисова, 2016]. Действительно, вопрос о влиянии «продающих текстов» на языковые оценки товаров в отзывах интересен. Однако эта задача требует подготовить специальный набор данных, включающий в себя помимо отзывов также рекламные материалы, и может стать темой отдельного исследования.

Предлагаемый обзор развития науки о потребительском поведении, конечно, не исчерпывающий. Однако он дает понять, что некоторые достижения в этой области могут быть использованы и в анализе языкового поведения потребителей. Так, идея мультиатрибутивных моделей товара перекликается с аспектным анализом тональности в том, что потребитель формирует свое отношение к продукту по набору его атрибутов. Поэтому результаты исследований в области коммуникативного поведения потребителей могут быть использованы в качестве теоретической базы для анализа тональности.

На основе этих теоретических положений и опыта эмпирического ручного аннотирования более 1,4 тыс. отзывов мы выделили шесть тематических классов аспектов, а именно:

1) *функциональные характеристики,*
2) *эксплуатационные характеристики,*
3) *конструкция,*
4) *внешний вид,*
5) *цена,*
6) *общие характеристики.*

Функциональные характеристики выделяются как В. Ш. Рубашкиным [2012], так и в модели Ф. Котлера [Kotler, 1967] и в методических указаниях о номенклатуре потребительских свойств РД 50-165-79, поэтому их выделение в отдельный класс аспектов полностью обоснованно.

Далее мы выделяем класс эксплуатационных характеристик, объединяющий специфические для техники параметры (по В. Ш. Рубашкину [2012]), признаки товара с подкреплением и часть признаков товара в реальном исполнении в модели Ф. Котлера [Kotler, 1967], а также ряд показателей качества из методических указаний о номенклатуре потребительских свойств РД 50-165-79: надежность в эксплуатации, эргономические показатели, показатели безопасности.

Следующий класс – конструкция, который мы выделяем на основе структурных характеристик по В. Ш. Рубашкину [2012]. Внешний вид мы выделяем по аналогии с эстетическими характеристиками в РД 50-165-79 и в модели Ф. Котлера [Kotler, 1967].

Потребительские отзывы как специфический вид текста связаны с покупкой, и это значит, что, помимо свойств объекта, они очень часто содержат информацию о цене. По этой причине, аспекты, связанные с ценой, мы выделяем в отдельный класс. К общим оценкам мы относим оценки товара в целом, а также аспекты, которые не вошли в другие классы.

3.2. Классификация предложений на шесть классов

Baseline-модель

На следующем этапе решалась задача сортировки предложений согласно шести описанным классам аспектов. Для этого мы разработали базовый подход (baseline), в котором в качестве признаков использовались векторы TF-IDF. Векторизации предшествовала предобработка текстов, включающая приведение к нижнему регистру, удаление небуквенных символов и удаление стоп-слов. Мы протестировали несколько популярных алгоритмов машинного обучения: Random Forest, Gradient Boosting, Naive Bayes, k-NN, SVM, Decision Tree. Для оценки качества классификации использовалась F-мера, представляющая собой гармоничное среднее точности и полноты. Она вычислялась с макро-усреднением, поскольку в наших данных наблюдается дисбаланс классов, при этом все классы одинаково важны. Результаты классификации с применением базового подхода представлены в таблице 2.

Таблица 2

Размер тренировочной и тестовой коллекции данных для каждой модели товара

Table 2

The Training and Test Datasets Size for Each Product

Продукт	**Подход**	**Random Forest**	**Gradient Boosting**	**Naive Bayes**	**k-NN**	**SVM**	**Decision Tree**
Кофеварка рожковая VT-1518	TF-IDF (baseline)	0,74	0,71	0,67	0,76	0,8	0,61
Робот-пылесос RV-R304	TF-IDF (baseline)	0,69	0,71	0,68	0,73	0,77	0,59
Стиральная машина EW6S4R06W	TF-IDF (baseline)	0,77	0,77	0,71	0,78	0,81	0,66
Телевизор 42PFL5405H	TF-IDF (baseline)	0,65	0,62	0,64	0,71	0,71	0,62
Электронная книга PRS-T2 2 ГБ	TF-IDF (baseline)	0,55	0,43	0,56	0,54	0,61	0,54

Для дальнейшей обработки были отобраны алгоритмы SVM, k-NN, Random Forest, показавшие лучшие результаты.

Предлагаемая модель

В качестве альтернативы базовому подходу мы предлагаем подход, который использует модель векторного представления слов FastText [Joulin et al., 2016; Bojanowski et al., 2017], обученную на собственной коллекции из ~1,1 млн отзывов. Модель обучалась со следующими параметрами: *size* = 50, *window* = 8, *min_count* = 5, *workers* = 2, sg = 1. Для получения векторных представлений предложений использовался метод центроидов [Brokos et al., 2016], который вычисляет вектор предложения как сумму векторов отдельных слов в предложении, взвешенных по TF-IDF, и деленных на сумму этих весов TF-IDF. Полученные векторные представления предложений далее используются в качестве данных для моделей машинного обучения.

В рамках предложенного подхода мы провели два эксперимента: первый только на размеченных данных; во втором эксперименте в полуавтоматическом режиме была расширена обучающая коллекция. Поскольку размер нашей размеченной коллекции отзывов не очень большой, мы хотели проверить идею улучшения качества работы классификатора за счет расширения обучающих данных.

Остановимся подробней на втором эксперименте. Проблема расширения обучающего датасета возникает достаточно часто из-за отсутствия достаточного количества размеченных данных для задач классификации с применением моделей машинного обучения [d'Sa et al., 2020]. Однако мы предлагаем свой подход к расширению тренировочного набора данных с использованием метода ключевых слов (КС). Для этого эксперимента мы сформировали дополнительный корпус отзывов на интересующие нас пять категорий товаров. При этом из дополнительного корпуса были исключены отзывы на модели, участвующие в первоначальном эксперименте. Затем были выделены две категории ключевых слов (КС): *общие* (которые в целом характеризуют данный класс аспектов вне зависимости от категории товара) и *специальные* (характеризующие данный класс аспектов применительно к конкретной категории товара).

Для первичного отбора кандидатов мы рассчитали меру TF-IDF для отдельных слов, биграмм и триграмм, но немного модифицировав ее: считая вместо частоты слова в документе частоту слова в категории. В качестве обратной частоты использовалась частота по всей коллекции отзывов. Затем списки слов сортировались по убыванию меры TF-IDF и выделялись первые 100 слов для каждой из пяти категорий товара. Далее эти списки были оценены экспертом. В итоге сформировались списки тех слов, которые, по мнению эксперта, наиболее точно соответствуют тематическому классу аспектов и используются активнее других в его описании. Окончательные списки ключевых слов представлены в таблице 3.

Следующий шаг – извлечение из дополнительного корпуса предложений, содержащих КС. Для этого мы применили подход, использующий регулярные выражения. Извлеченным предложениям автоматически присваивалась метка тематического класса аспектов, соответствующая классу КС. Если предложение содержало слова из нескольких классов, то они не включались в расширенный датасет. Мы провели три эксперимента с разным размером расширенного датасета: в 10, в 30 и в 50 раз больше исходной тренировочной коллекции. Результаты всех экспериментов в сопоставлении с базовым подходом представлены в таблице 4.

Таблица 3

Ключевые слова для предустановленных классов аспектов

Table 3

Manually Defined Aspects and Their Most Frequent Keywords

Аспект	**Тип ключевых слов**	**Ключевые слова**
1	**2**	**3**
Внешний вид	общие	красив*, внешний вид, стильн*, симпатичн*, дизайн, эстет*, изящн*
Конструкция	общие	сборк*, материал, пластик, пластмасса, желез*, металл, покрытие, хлипк*, поверхност*, габарит*, занимает мало, занимает много, компактн*, узк*, зазор, легк*, тяжел*, собран, размер
Эксплуатационные характеристики	общие	надежн*, удобн*, инструкция, управление, быстр*, легко, эсплуатац*, обслуживан*, мощн*, простота, понятн*, шум, тих*, запах, интуитивн*, использован, нужно, расход, чистить, хранить, чистк*, уход, сломал, ремонт, долговечн*, потребл*, комплект, запасн*
	Специальные (кофемашины)	мыть, нагрев, наливать, проливат*, капает
	Специальные (роботы-пылесосы)	большой пылесборник, вместительный контейнер, вреза, застрев*, заряжается, заряд, батарея, аккумулятор, станц*, базу
	Специальные (стиральные машины)	вибрир*, прыгает, скачет, окончании стирки, отсек для порошка, класс энергопотребл*
	Специальные (телевизоры)	долгое переключение, переключать, hdmi, излучение, меню
	Специальные (электронные книги)	сенсорный, навигация, подсветка, держать, аккумулятор, заряд
Функциональные характеристики	общие	программ*, функц*, режим
	Специальные (кофемашины)	получается, кофе, варит, готовит, взбивает, вкус, пенка, вспенивает, аромат, эспрессо, латте
	Специальные (роботы-пылесосы)	автоматическ*, помощник, пылесосит, уборка, убирает, засасывает, всасывает, моющий, собирает, пыль, влажн*, моет, шерсть, залезает, график, угл*
	Специальные (стиральные машины)	одеял*, отстирыв*, стир*, отжим, глажка, полоск*, сушк*, вместител*, загрузка

Окончание табл. 3

1	2	3
	Специальные (телевизоры)	скайп, плеер, тюнер, эфир, full hd, цвет, изображение, картинка, углы обзора, прием, флешка, герц, фильм, гц, засвет, звук, приём, usb, читает, формат
	Специальные (электронные книги)	ink, pdf, djvu, читает, формат, контраст, epub, pdf, шрифт, поддерживает
Общие характеристики	общие	советую, рекомендую, доволен, супер
Цена	общие	цена, ценник, дорог*, дешев*, стоимость, деньг*, цена-качество, цена/качество

Таблица 4

Оценка результатов разных моделей классификации предложений на шести предустановленных классах аспектов, выполненная при помощи F-меры

Table 4

Comparison of Different Models F-Scores in Sentences' Classification for Six Predefined Aspect Categories

Продукт	Подход	Random Forest	k-NN	SVM
	2	3	4	5
Кофеварка рожковая VT-1518	TF-IDF	0,74	0,76	0,8
	fasttext_small	0,8	0,84	**0,86**[4]
	fasttext_expanded_10	0,83	**0,85**	0,86
	fasttext_expanded_30	**0,84**	0,82	0,84
	fasttext_expanded_50	0,81	0,81	0,81
Робот-пылесос RV-R304	TF-IDF	0,69	0,73	0,77
	fasttext_small	0,79	0,82	0,82
	fasttext_expanded_10	**0,82**	**0,84**	**0,86**
	fasttext_expanded_30	0,81	0,83	0,82
	fasttext_expanded_50	0,82	0,83	0,82
Стиральная машина EW6S4R06W	TF-IDF	0,77	0,78	0,81
	fasttext_small	0,8	0,81	0,85
	fasttext_expanded_10	0,83	0,8	**0,86**
	fasttext_expanded_30	**0,86**	0,82	0,8
	fasttext_expanded_50	0,84	**0,83**	0,82
Телевизор 42PFL5405H	TF-IDF	0,65	0,71	0,71
	fasttext_small	0,71	0,74	**0,81**
	fasttext_expanded_10	0,78	0,81	0,81
	fasttext_expanded_30	**0,87**	**0,84**	0,79
	fasttext_expanded_50	0,85	0,84	0,8

[4] Жирным выделены лучшие результаты.

Окончание табл. 4

1	2	3	4	5
Электронная книга PRS-T2 2 ГБ	TF-IDF	0,55	0,54	0,61
	fasttext_small	0,64	0,71	0,72
	fasttext_expanded_10	0,67	0,74	**0,77**
	fasttext_expanded_30	0,74	**0,78**	0,77
	fasttext_expanded_50	**0,75**	0,78	0,76

Как видно из таблицы выше, расширение тренировочной коллекции при помощи КС положительно влияет на качество классификации. В 13 из 15 случаев расширение приводит к улучшению метрики качества классификации.

Анализ тональности

На следующем шаге проводился анализ тональности на уровне предложений внутри каждого класса аспектов. Предложения были поделены на два класса: *положительные* и *отрицательные*. В тренировочную коллекцию также были добавлены предложения из расширенного датасета, полученного на предыдущем этапе. Для автоматического присваивания метки тональности мы использовали метаинформацию из той части отзыва, в которой находилось предложение (достоинства, недостатки, комментарий). Метка «–» автоматически ставилась предложениям из группы «Недостатки», а «+» – из группы «Достоинства». Предложения из «комментария» для расширения тренировочной коллекции не использовались. В качестве признаков для моделей классификации использовались те же векторы предложений, что и на предыдущем шаге. Оценка результатов классификации представлена в таблице 5.

Таблица 5

Оценка результатов разных моделей классификации предложений по тональности, выполненная при помощи F-меры

Table 5

Comparison of Different Models F-Scores in Sentiment Analysis Task

Продукт	Подход	Random Forest	k-NN	SVM
Кофеварка рожковая VT-1518	fasttext_expanded_30	0,86	0,85	0,87
Робот-пылесос RV-R304	fasttext_expanded_30	0,9	0,91	0,9
Стиральная машина EW6S4R06W	fasttext_expanded_30	0,91	0,93	0,92
Телевизор 42PFL5405H	fasttext_expanded_30	0,92	0,91	0,91
Электронная книга PRS-T2 2 ГБ	fasttext_expanded_30	0,91	0,9	0,91

3.4. Ранжирование предложений

После классификации по тональности встала задача выбрать самые информативные предложения для итогового резюме. Для решения был выбран алгоритм TextRank [Mihalcea et al., 2004], в основе которого лежит представление текста в виде неориентированного графа. Вершинами графа являются векторы предложений, а ребра между ними имеют вес, равный похожести вершин, вычисляемый при помощи косинусного расстояния. Для итогового резюме мы отбирали по три предложения в каждом классе с наибольшим весом.

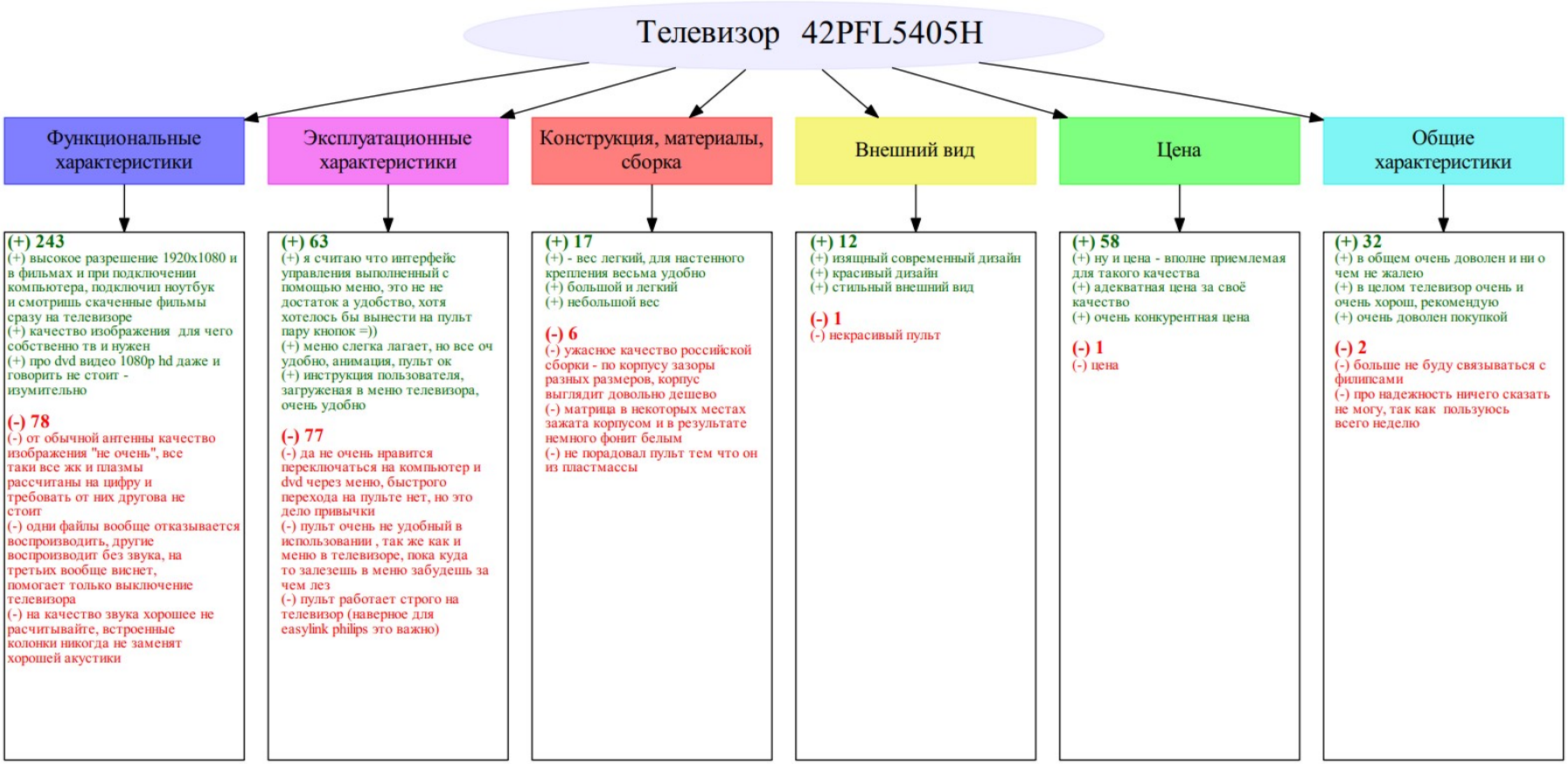

Рис. 3. Пример визуализации суммаризации отзывов на телевизор 42PFL5405H
Fig. 3. Visualization of reviews' summarization for the TV 42PFL5405H

3.5. Визуализация

Результат работы нашего алгоритма суммаризации мы визуализировали с помощью библиотеки graphviz [Ellson et al., 2004]. Пример суммаризации отзывов на Телевизор 42PFL5405H представлен на рисунке 3.

Заключение

Исследование показало, что предложенный и реализованный в статье алгоритм эффективен в решении задачи суммаризации большого количества отзывов для разных моделей товара.

Предложенный подход имеет следующие преимущества:

1) экспертное установление универсальных для цифровой и бытовой техники тематических классов аспектов позволяет применять его к широкому классу изделий без адаптации под конкретную модель;

2) визуализация показывает структуру резюме в соответствии с тематической иерархией аспектов;

3) сгенерированное резюме дает число предложений с положительной и отрицательной тональностью, относящихся к каждому классу аспектов, что позволяет легко оценить картину потребительских предпочтений.

Недостаток текущей версии подхода заключается в отсутствии гибкости при определении тематических классов аспектов для конкретной группы товаров.

В ходе эксперимента было показано, что выбор подхода к векторизации предложений – это важный фактор, влияющий на качество работы алгоритмов. Также была продемонстрирована зависимость качества работы алгоритмов от размера обучающей коллекции и предложен новый подход к расширению обучающих данных.

Качество работы алгоритма оценивалось на разных этапах при помощи F-меры на пяти моделях товара из различных категорий. Для этапа классификации предложений по заданным классам аспектов F-мера составила от 77 до 86 %, а по классам тональности (положительные и отрицательные) – от 86 до 92 %. Это свидетельствует о достаточно высокой эффективности и универсальности предложенного подхода.

Список литературы

Борисова Е. Г. Маркетинговая лингвистика: направления и перспективы // Верхневолжский филологический вестник. 2016. № 4. С. 140–143.

РД 50-165-79 Товары народного потребления. Выбор номенклатуры потребительских свойств и показателей качества. Основные положения [Электронный ресурс]. URL: https://files.stroyinf.ru/Index2/1/4293762/4293762287.htm (дата обращения: 28.05.2022).

Рубашкин В. Ш. Онтологическая семантика. Знания. Онтологии. Онтологически ориентированные методы информационного анализа текстов. М.: Физматлит, 2012. 348 с.

Akhtar N. et al. Aspect based sentiment oriented summarization of hotel reviews // Procedia computer science. 2017. T. 115. Pp. 563–571.

Bojanowski P. et al. Enriching word vectors with subword information // Transactions of the association for computational linguistics. 2017. Vol. 5. Pp. 135–146.

Brokos G. I., Malakasiotis P., Androutsopoulos I. Using centroids of word embeddings and word mover's distance for biomedical document retrieval in question answering // Proceedings of the 15th Workshop on Biomedical Natural Language Processing, 2016. Pp. 114–118.

Condori R. E. L., Pardo T. A. S. Opinion summarization methods: Comparing and extending extractive and abstractive approaches // Expert Systems with Applications. 2017. Vol. 78. Pp. 124–134.

d'Sa A. G. et al. Label propagation-based semi-supervised learning for hate speech classification // Insights from Negative Results Workshop, EMNLP, 2020.

Ellson J. et al. Graphviz and dynagraph—static and dynamic graph drawing tools // Graph drawing software. Springer, Berlin, Heidelberg, 2004. Pp. 127–148.

Fishbein M. An investigation of the relationships between beliefs about an object and the attitude toward that object // Human relations. 1963. Vol. 16, № 3. Pp. 233–239.

Howard J. A. Consumer behavior: Application of theory. McGraw-Hill Companies. 1977.

Johnson E. J., Tversky A. Representations of Perceptions of Risk // Journal of Experimental Psychology: General, 1984. Pp. 55–70.

Hu M., Liu B. Mining and summarizing customer reviews // Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004. Pp. 168–177.

Kim S. et al. A hierarchical aspect-sentiment model for online reviews // Proceedings of the AAAI Conference on Artificial Intelligence. 2013. Vol. 27. Pp. 526–533.

Joulin A., Grave E., Bojanowski P., Mikolov T. Bag of Tricks for Efficient Text Classification // Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 2017. Pp. 427–431.

Kotler F. Marketing Management: Analysis, Planning, and Control. Prentice-Hall. 1967. 628 p.

Lancaster K. J. A new approach to consumer theory // Journal of political economy. 1966. Vol. 74, № 2. Pp. 132–157

Mihalcea R., Tarau P. Textrank: Bringing order into text // Proceedings of the 2004 conference on empirical methods in natural language processing. 2004. Pp. 404–411.

Mithun S., Kosseim L. Summarizing blog entries versus news texts // Proceedings of the Workshop on Events in Emerging Text Types. 2009. Pp. 1–8

Mukherjee A., Liu B. Aspect extraction through semi-supervised modeling // Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2012. Pp. 339–348.

Rosenberg M. J. Cognitive structure and attitudinal affect // The Journal of abnormal and social psychology. 1956. Vol. 53, № 3. P. 367.

Thaler R. Toward a positive theory of consumer choice // Journal of economic behavior & organization. 1980. Vol. 1, № 1. Pp. 39–60.

Tversky A., Kahneman D. Judgment under uncertainty: Heuristics and biases // Science. 1974. Vol. 185, № 4157. Pp. 1124–1131.

Zhai Z. et al. Clustering product features for opinion mining // Proceedings of the fourth ACM international conference on Web search and data mining. 2011. Pp. 347–354

Yu J. et al. Domain-assisted product aspect hierarchy generation: towards hierarchical organization of unstructured consumer reviews // Proceedings of the 2011 conference on empirical methods in natural language processing. 2011. Pp. 140–150.

References

Akhtar, N. et al. Aspect based sentiment oriented summarization of hotel reviews. Procedia computer science, 2017, vol. 115, pp. 563–571.

Bojanowski, P. et al. Enriching word vectors with subword information. Transactions of the association for computational linguistics, 2017, vol. 5, pp. 135–146.

Borisova, E. G. Marketing linguistics: prospects and trends. Verhnevolzhski philological bulletin, 2016, no. 4, pp. 140–143.

Brokos, G. I., Malakasiotis, P., Androutsopoulos, I. Using centroids of word embeddings and word mover's distance for biomedical document retrieval in question answering. Proceedings of the 15th Workshop on Biomedical Natural Language Processing, 2016, pp. 114–118.

Condori, R. E. L., Pardo, T. A. S. Opinion summarization methods: Comparing and extending extractive and abstractive approaches. Expert Systems with Applications, 2017, vol. 78, pp. 124–134.

d'Sa, A. G. et al. Label propagation-based semi-supervised learning for hate speech classification. Insights from Negative Results Workshop, EMNLP, 2020.

Ellson, J. et al. Graphviz and dynagraph—static and dynamic graph drawing tools. Graph drawing software. Springer, Berlin, Heidelberg, 2004, pp. 127–148.

Fishbein, M. An investigation of the relationships between beliefs about an object and the attitude toward that object. Human relations, 1963, vol. 16, no. 3, pp. 233–239.

Howard, J. A. Consumer behavior: Application of theory. McGraw-Hill Companies. 1977.

Johnson, E. J., Tversky, A. Representations of Perceptions of Risk. Journal of Experimental Psychology: General, 1984, pp. 55–70.

Hu, M., Liu, B. Mining and summarizing customer reviews. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004, pp. 168–177.

Joulin, A., Grave, E., Bojanowski, P., Mikolov, T. Bag of Tricks for Efficient Text Classification. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 2017, pp. 427–431.

Kim, S. et al. A hierarchical aspect-sentiment model for online reviews. Proceedings of the AAAI Conference on Artificial Intelligence, 2013, vol. 27, pp. 526–533.

Kotler, F. Marketing Management: Analysis, Planning, and Control. Prentice-Hall, 1967, 628 p.

Lancaster, K. J. A new approach to consumer theory. Journal of political economy, 1966, vol. 74, no. 2, pp. 132–157.

Mihalcea, R., Tarau, P. Textrank: Bringing order into text. Proceedings of the 2004 conference on empirical methods in natural language processing. 2004, pp. 404–411.

Mithun, S., Kosseim, L. Summarizing blog entries versus news texts. Proceedings of the Workshop on Events in Emerging Text Types. 2009, pp. 1–8

Mukherjee, A., Liu, B. Aspect extraction through semi-supervised modeling. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2012, pp. 339–348.

RD 50-165-79 Consumer goods. Selection of the nomenclature of consumer properties and quality indicators. The main features [Online]. URL: https://files.stroyinf.ru/Index2/1/4293762/4293762287.htm (accesed on: 28.05.2022) (in Russ.)

Rosenberg, M. J. Cognitive structure and attitudinal affect. The Journal of abnormal and social psychology, 1956, vol. 53, no. 3, p. 367.

Rubashkin, V. Ontological semantics. Knowledge. Ontologies. Ontology-based approach to information analysis of text. Moscow, Fizmatlit, 2012, 348 p. (in Russ.)

Thaler, R. Toward a positive theory of consumer choice. Journal of economic behavior & organization, 1980, vol. 1, no. 1, pp. 39–60.

Tversky, A., Kahneman, D. Judgment under uncertainty: Heuristics and biases. Science, 1974, vol. 185, no. 4157, pp. 1124–1131.

Zhai, Z. et al. Clustering product features for opinion mining. Proceedings of the fourth ACM international conference on Web search and data mining, 2011, pp. 347–354

Yu, J. et al. Domain-assisted product aspect hierarchy generation: towards hierarchical organization of unstructured consumer reviews. Proceedings of the 2011 conference on empirical methods in natural language processing, 2011, pp. 140–150.

Информация об авторе

Чечнева Надежда Сергеевна, аспирант, Санкт-Петербургский государственный университет, Санкт-Петербург, Россия

Information about the Author

Nadezhda S. Chechneva, PhD student, Saint Petersburg State University, Saint Petersburg, Russia

Статья поступила в редакцию 01.07.2022;
одобрена после рецензирования 10.10.2022; принята к публикации 28.10.2022
The article was submitted 01.07.2022; approved after reviewing 10.10.2022;
accepted for publication 28.10.2022