Electronic Commerce Research
https://doi.org/10.1007/s10660-024-09901-x

# Shaping the causes of product returns: topic modeling on online customer reviews

**Andrea Mor**[1] · **Carlotta Orsenigo**[1] · **Mauricio Soto Gomez**[2] · **Carlo Vercellis**[1]

Accepted: 11 September 2024


**Abstract**
Product return is a common phenomenon in the online retailing industry and entails several inconveniences for both the seller, who incurs in high costs for restocking the returned goods, and the customer, who has to deal with product re-shipping. In this paper, we outline a data-driven approach, based on Natural Language Processing, in which a broad corpus of customer reviews of an online retailer is exploited with the aim of shaping the main causes of product returns. In particular, a variety of topic modeling techniques represented both by classic methods, given by LDA and variants, and more recent algorithms, i.e., BERTopic, were applied to identify the main return reasons across multiple product categories, and their outcomes were compared to select the best approach. The category-dependent sets of return causes inferred through topic modeling largely enrich the product-agnostic list of return reasons currently used on the e-commerce platform, and provide valuable information to the retailer who can devise ad-hoc strategies to mitigate the returns and, hence, the costs of the related logistic network.

**Keywords** Natural language processing · Topic modeling · Latent Dirichlet allocation · Product return · Customer reviews

## 1 Introduction

In recent years, most retailers progressively shifted from the traditional offline paradigm of 'customers going to goods', through a network of physical stores to finalize sales, to the online paradigm of 'goods going to customers', with sales being typically performed on e-commerce platforms. This phenomenon has a great impact on the dynamics of a purchase. In the offline case, customers can directly see, handle, and test the product (test → buy). Conversely, when the

Andrea Mor, Carlotta Orsenigo, Mauricio Soto Gomez, Carlo Vercellis have contributed equally to this work.

Extended author information available on the last page of the article

Published online: 22 September 2024

purchase is made online, the evaluation of the product takes place only at a later stage (buy → test).

To overcome the risk perceived by customers in online shopping due to the missing touch and feel experience, many online retailers implement incentives in the form of lenient return policies that allow the refund of the full price and the inexpensive, often free, return of the product in case of dissatisfaction [1]. As a consequence, the average return rate observed in online sales is markedly higher than that of brick-and-mortar retailers. For example, the apparel industry reports a return rate of 9% and 20% for the offline and the online channels, respectively [2].

While lenient return policies can sometimes increase the volume of items purchased more than the volume of items returned [3], the cost of the related *reverse* logistic network is, nevertheless, considerably high. In 2007, the cost of product returns, which accounts, among others, for inventory holding and product handling costs, was estimated at $100 billion per year, or as an average loss per company of about 3.8% in profit [4]; this loss increased to $428 billion for US retailers in 2020 [5]. The electronics industry alone is reported to spend approximately $14 billion every year on product returns through repacking, restocking, and reselling, with only about 5% of returns due to defective items [6]. Furthermore, a large amount of the value of a returned product is eroded in the return process itself [7], and although companies try to mitigate this issue through alternative strategies (i.e., re-distribution of products in mint or good conditions, product re-manufacturing or scraping of parts), the final loss can be as high as 45% of the overall asset value [8]. Environmental impact is also a factor, with online returns producing 14% more landfill waste compared to in-store returns [9], to the extent that environmental sustainability frameworks have been developed to solve the issue [10]. Therefore, it is crucial to correctly identify the motivations behind product returns, to address the causes of discontent beforehand, prevent the misalignment between product features and buyer expectations, and keep the return rate at a minimum while increasing customer satisfaction.

To collect the return reasons on e-commerce platforms, retailers usually resort to a list of predefined causes, and buyers are asked to select one of these options. This list is often generic and equally applied to all products, especially in multi-vendor marketplaces, despite the intrinsic differences among goods belonging to distinct categories. Footwear and apparel items, for example, are more prone to fit or size-related problems [11], than to damages or defects. As a consequence, the return reasons made available to customers may be inadequate to punctually reflect the causes of their dissatisfaction.

Unlike offline sellers, online retailers can leverage a prominent asset represented by online customer reviews. Various studies analyzed the importance of product reviews, showing how customer feedback can improve several aspects of the product life-cycle and helps buyers make better-informed decisions. At the same time, customer reviews embed a valuable source of information that has been, for example, used to retrieve product features and opinions [12], as well as customer concerns [13]. We believe that customer reviews can be effectively exploited by retailers also to identify, classify, and address the reasons for product returns.

Although most studies agree on the negative effect of unmet customer expectations [14], the interpretation on how customer feedback impacts their behavior appears highly dependent on the product category. For instance, some authors report that the valence (score) of reviews and sales are positively correlated in the case of movies [15, 16]. Other studies, focused on consumer electronics and video games, highlight the difference between search products (i.e., goods that consumers can evaluate by specific attributes before purchase) and experience products (which are more difficult to describe using specific attributes), with the former being more affected by the valence of the reviews, and the latter by their amount [17]. These findings emphasize the peculiarities of different goods and call for the development of tools able to identify the specific reasons for return pertaining to products in distinct categories.

The present paper aims to advance the existing research by proposing a data-driven approach, based on Natural Language Processing (NLP), that outlines product return causes from the reviews collected on e-commerce platforms. Our study contributes to the literature by widening previous works in terms of scope and applied methodologies. From one side, other authors applied language processing tools to textual customer feedback. However, their goal was to predict customer satisfaction for specific sectors or products, as illustrated in Sect. 2. From the other, topic modeling was applied to explore the reasons for product returns [18], similarly to our work. However, the authors focused on one single category (i.e., luxury items) and employed one modeling technique. Furthermore, they based their analysis on texts extracted from return forms instead of customer reviews. To the best of our knowledge, this paper presents the first systematic analysis of return motivations across several product categories generated from customer reviews by applying different topic modeling methods. Therefore, by accounting for multiple product categories it enables the detection of cross-category insights that can be leveraged to design common return policies. At the same time, it outlines reasons for return that are not dependent on a specific technique, and whose reliability is verified by applying different topic modeling methods.

The data-driven strategy proposed in the paper is composed by three main phases, given by data selection, return reasons inference, and return causes analysis. In particular, from the Amazon Review Dataset [19], which contains about 233.1 million reviews of products divided into 29 categories, we extracted a subset of reviews discussing product returns and applied topic modeling to gain insights and shape the main reasons for the return. These reasons were validated by comparing the results achieved through four topic modeling techniques, represented by classical Latent Dirichlet Allocation (LDA) [20], conventional variants specifically designed for short texts, namely Biterm Topic Model [21] and Self-Aggregation based Topic Model [22], and BERTopic [23], a topic generation method using a transformer-based word embedding. The same causes were also compared to a detailed list of motivations used by Amazon with third-party sellers for handling prepaid returns. This comparison showed that the product-agnostic strategy currently used on the e-commerce platform to collect the return causes is unsuited to capture the existing relationship between these causes and the product features, which is instead reflected in the set of reasons inferred, for each category, through our approach. To

achieve further category-level evidences, we also identified some trends occurred in each product category over time, and performed sentiment analysis on the selected corpus of reviews.

The manuscript is organized as follows. Section 2 presents a review of the main achievements in the research fields related to our work. Section 3 briefly illustrates the topic generation approaches, together with the customer reviews dataset and the proposed experimental settings. Results and managerial implications are discussed in Sect. 4, while conclusions and future research developments are drawn in Sect. 5.

## 2 Literature review

Our paper spans three distinct but related areas: the problem of product returns for online retailers, the analysis of online reviews to gain insights into the user purchasing behavior, and the application of NLP techniques to extract knowledge from customer feedback. This section is, therefore, devoted to summarizing the main studies on the aforementioned subjects useful to frame our work in the respective research fields. Reviews focused on the current and perspective state of the literature discussing product return are presented in [24, 25].

***Product returns and customer reviews*** Product returns in the online retailing industry received great attention in the literature, as they play a prominent role from both the economic and the reputation perspective [7]. Since the flow of returned products is strictly connected to the return policies in place, several authors analyzed how less or more lenient policies can influence customer satisfaction [26, 27], sales [28], and profit [6, 29–31]. Other works, instead, investigated the mechanisms to decrease the uncertainty affecting the purchasing process, such as online forums [32] or the use of online product inspection technologies [33], given that a key driver to reduce the amount of returned products is the matching between the actual product features and the customer expectations [34].

Customer reviews were also extensively examined, primarily for the implications they bring to the customer decisions, from both a theoretical [35–39] and an empirical viewpoint. Several works, for instance, analyzed the empirical relation between reviews and customer purchase behaviour [38, 40–42], sales [16, 43, 44], and pricing [34]. Another relevant stream focused on the impact exerted on sales by different reviews-based quantitative dimensions such as valence, volume, and variance, measured on the rating assigned by customers, the number of evaluations, and the dispersion of the ratings, respectively. A positive correlation between these factors and sales was shown for online book retailers [43] and in the industry of movies and TV shows [45–47]. Similar results were presented in [17], where the valence of the reviews was proven to be positively correlated with sales, especially for search products, whereas their volume has a greater effect on experience products. The studies reported above showed that the impact of these factors, when taken individually, cannot be generalized to a larger set of product categories. To tackle the discrepancies in the effects of these factors, several studies [15, 42, 48] considered them by taking into account their mutual interactions.

While product returns and customer reviews were investigated separately with respect to the purchasing process, fewer attempts were made to explore their potential interplay. In this regard, studies were proposed to analyze the extent to which the valence of customer reviews may influence the return rate [39], sometimes by focusing on specific product categories like fashion, as in [49]. Other studies were, instead, devoted to evaluating the effects of overly positive rating [50], or the relationship between product returns and the willingness to leave a review [51].

***NLP methods on customer reviews*** Another research area of interest is focused on the use of textual processing tools to leverage information from customer feedback. In this context, the primary purpose is to retrieve and/or summarize the product features and the consumer opinion by analyzing customer comments.

Summarization and extraction of relevant features by means of semantic analysis rules are, for instance, discussed in [52–54], whereas in [12] these features are retrieved from a collection of customer reviews by a mixed approach made of a kernel-based model and a statistical mapping between words. Classical textual representation models, such as TF-IDF and word2vec, were employed in [55] to classify customer comments into predefined concern codes, while a combination of NLP and deep learning techniques is exploited in [56] to understand the reasons behind customer perception. Customer reviews are used in [57] to extract topics and, subsequently, the emotions felt by customers while interacting with robotic servants in hotels.

In the NLP domain, a natural technique for the analysis of textual information is topic modeling. Its aim is to discover abstract topics within a collection of documents by identifying recurring themes or patterns in the text data. Among these techniques, LDA-based methods has been widely explored to predict customer satisfaction in different sectors, such as restaurants, hotels [58–60] and banks [61], and for specific products, like baby pacifiers [62], light food [63], and MP3 players [59]. Similar approaches have been proposed to identify opinion features from online reviews [64].

To the best of our knowledge, the use of topic modeling to investigate product returns is largely unexplored. The study closest to the present work is given by [18], where the authors perform a topic generation task on customers' feedback, collected through return forms, of a luxury-home goods company. Unlike this work, our study resorts to online, free-text customer reviews as the documents for knowledge extraction, and copes with the return causes for a wide range of product categories, thus accounting for the specificity of each group. Moreover, compared to the previous work in which a single topic modeling method (i.e., LDA) is used, different techniques are here applied and the robustness of the outlined return causes is evaluated across all of them.

## 3 Methods

### 3.1 Topic modeling for return causes extraction

In the following section, the methodological steps used to extract return caused from customer reviews though topic modeling are reported. Note that, for the sake of

readability, a short definition of the most common terminology of topic modeling is reported in Table 1.

Four topic modeling methods were considered for deriving the main reasons for product returns. The first and most classical is Latent Dirichlet Allocation (LDA) [20]. In this model, each document $d$ within a given corpus $D$ is assumed to be the result of a stochastic generation process comprising words that belong to a random mixture of $K$ latent topics. In LDA the distribution $\theta_d$ of topics in a document is supposed to be the realization of a Dirichlet distribution, $\theta_d \sim \text{Dir}(\alpha)$. Furthermore, each word $w$ in the corpus is assumed to belong to a topic, with each topic having its own distribution $\phi_k \sim \text{Dir}(\beta)$ over the words forming the dictionary.

Despite their well-proven effectiveness on large and medium documents, conventional topic modeling methods, such as LDA, generally perform poorly when applied to short texts like the user-generated content on online platforms and social media. The performance degradation over short-texts corpora is due to their severe lack of word co-occurrence information at a document-level, which is implicitly captured by classical models for extracting the topics. To overcome this data sparsity issue several approaches have been developed [65]. Among these, we focused on the Biterm Topic Model (BTM) [21], which is among the earliest works proposed for short texts, and the Self-Aggregation based Topic Model (SATM) [22].

The distinctive trait of BTM lies in the idea of explicitly modeling the co-occurrence patterns in the whole corpus instead of the patterns at a document-level, which are much harder to trace in documents containing a small number of words. In BTM, co-occurrence patterns are represented by biterms, defined as unordered pairs of words co-occurred in a short context. Biterms are first extracted from the corpus, and the inference is then performed on them. Compared to LDA, BTM was shown to reveal more coherent topics [21] and generally requires a greater computational time, since the set of biterms can be significantly larger than the document length. However, the amount of memory required by BTM is shown to increase less rapidly than LDA with regards to the number of documents and the number of topics [66].

**Table 1** A brief definition of the topic modeling-related terms most commonly used in the manuscript

| Term | Definition |
|---|---|
| Word | The basic unit of discrete data, defined to be an item from a vocabulary [20] |
| Document | A sequence of words [20]. In the dataset at hand, a customer review |
| Corpus | A collection of documents [20] |
| Topic | Collection of words that frequently occur together within a set of documents |
| Lemmatization | The process of reducing a word to its base or root form, known as the lemma, which is the canonical form of the word. For example, {running, ran, runs} → {run, run, run} |
| Part-of-Speech (POS) tagging | The process of labeling each word in a sentence with its corresponding part of speech (e.g., noun, verb) |
| Stopword | Commonly used word considered to carry very little meaningful information and are frequently removed to reduce noise and improve the efficiency of text analysis. For example, “and”, “the” |
| Coherence score | Metric used to evaluate the quality and interpretability of the topics generated by a topic model |

A different approach is implemented in SATM, in which the problem of data sparsity is tackled by aggregating short texts with similar topics into longer texts before inferring the ultimate set of latent topics. In particular, given the assumption that each short document in the corpus, $d \in D$, is sampled from an unobserved long pseudo-document $l \in P$ in the text collection, with $|D| \gg |P|$, SATM employs a two-step procedure, in which a standard topic modeling (e.g., LDA) is first run on the dataset of short texts and the derived topics are then used to generate longer documents, called pseudo-texts, to increase the word co-occurrence. It is worth noticing that finding the optimal number of pseudo-documents that best suits a dataset, and which likely depends on the optimal topic count [22], requires a large amount of computation. On the other hand, the pseudo-document framework was shown to be extremely effective in case of data sparsity, outperforming both LDA and BTM in some application domains.

Finally, BERTopic was also considered. As the name implies, this method uses a pre-trained Bidirectional Encoder Representations from Transformers (BERT) language model. This model considers both the preceding and the succeeding context of each word to generate a document-level vectorial representation that takes into account semantic relationships between words and phrases (text embedding), via a deep learning architecture. A dimensionality reduction technique is then applied, followed by a clustering method on this reduced representation to obtain the topics (clusters). Lastly, the semantic description of the topics is derived from a modified version of the TF-IDF method, which quantifies the importance of terms in a document according to their relative frequency. In this modified version, documents are composed of the set of comments belonging to the same cluster.

### 3.2 Customer reviews dataset

The purpose of the study is to apply topic modeling to identify the main factors driving product returns across diverse product categories. To the best of our knowledge, no (publicly available) dataset exists that specifically collects the feedback given by customers while returning their purchase. For this reason, the Amazon Review Dataset [19], called ARD hereafter, was chosen as the source of information.

ARD contains 233.1 million reviews for products divided into 29 categories. The reviews span from May 1996 to October 2018, with about 95% being from 2011 or later. Since the object of investigation are the causes of returns, from ARD we extracted a subset of data suitable to the desired goal. In particular, a two-step filtering procedure was applied to the available reviews. In the first step, comments in each product category were processed to keep only the reviews containing either one of the following words: *return*, *returned*, *returning*. To increase the robustness and representativeness of the results, in the second step we retained only the categories satisfying at least one of the following three conditions: 1) they contained at least 150k comments, 2) their filtered size was larger than 2.5% of their original size (i.e., before filtering), 3) the ratio between the number of reviews and the number of products was above a given threshold (here fixed to 0.6). These conditions were empirically found to generate a set of categories that

is sufficiently large to be a representative and heterogeneous sample and, at the same time, for which the number of comments associated to returns covers an adequate portion of the products within each category.

It is worth noticing that, filtering the data using the aforementioned keywords might not guarantee to keep only the comments discussing product returns. Indeed, there are categories that naturally contain a high rate of false positives: for instance, books whose title contains *return*-related words, such as "The Return of the King", or software to file tax returns. For this reason, to base our analysis on a more reliable set of data, from the list of categories obtained after the filtering process we also dropped "Books" and "Software".

The final subset used consists of the 12 categories listed in Table 2. The table also contains a detailed description of each group of product reviews in terms of original size (i.e., original number of reviews), filtered size, final size (i.e., number of reviews after filtering and texts preprocessing, see Sect. 3.3), and the ratio of final size to original size (in percentage). As can be inferred from the table, the final set of reviews under analysis formed a corpus composed by more than 3.4 million documents.

### 3.3 Texts preprocessing

Prior to the return causes extraction, some preprocessing was performed to make the collection of texts suitable for topic modeling.

**Table 2** Statistics of the corpus of customer reviews in each product category, before filtering (Column 2), after reviews/categories filtering (Column 3), and after texts preprocessing (Columns 4-7)

| Product category | Original size | Filtered size | Final size | Final/Original size ratio (%) | Avg. text length (words) | Dictionary size (words) |
|---|---|---|---|---|---|---|
| Amazon fashion | 883,636 | 28,939 | 25,592 | 2.90 | 6.5 | 8,740 |
| automotive | 7,990,166 | 156,234 | 143,999 | 1.80 | 9.6 | 28,761 |
| Cell Phones and accessories | 10,063,255 | 228,548 | 210,203 | 2.09 | 9.8 | 32,116 |
| Clothing shoes and jewelry | 32,292,099 | 1,189,586 | 1,061,935 | 3.29 | 7.1 | 57,874 |
| Electronics | 20,994,353 | 759,960 | 717,535 | 3.42 | 15.2 | 82,707 |
| Home and kitchen | 21,928,568 | 595,192 | 563,884 | 2.57 | 10.7 | 54,978 |
| Luxury beauty | 574,628 | 9,864 | 9,298 | 1.62 | 9.6 | 6,909 |
| Musical instruments | 1,512,530 | 38,493 | 36,631 | 2.42 | 14.8 | 18,544 |
| Pet supplies | 6,542,483 | 133,108 | 124,786 | 1.91 | 11.0 | 27,533 |
| Sports and outdoors | 12,980,837 | 305,626 | 285,945 | 2.20 | 11.1 | 44,805 |
| Tools and home improvement | 9,015,203 | 223,974 | 210,477 | 2.34 | 11.6 | 35,872 |
| Video games | 2,565,349 | 57,834 | 55,270 | 2.15 | 29.2 | 39,605 |

As depicted in Fig. 1, the raw dataset achieved after filtering was cleaned by removing from each document (i.e., review) undesired characters or terms, such as punctuation marks, emails, hashtags, html tags, web links, and by converting all words to lower case. In a second step, lemmatization was applied through the `WordNetLemmatizer` available in the NLTK library [67], and all the stopwords included in the library's list of English stopwords were removed. Moreover, from the entire collection of lemmas we retained only those classified as either verbs or adjectives by the Part-of-speech (POS) tagging technique implemented in NLTK. This step turned out to be extremely useful in the identification and the distinction of topics, describing the cause of the return, from the products being returned, especially in all cases when nouns, representing the name of the products, were kept in the text. Finally, an evaluation phase was carried out by estimating an LDA model and analyzing the resulting topics. All the words shared by multiple topics were inspected and added to the stopwords list in case they were recognized as stopwords. This last phase, therefore, further improved the quality of the texts at hand by inducing a retro-action based on the identification of new stopwords to remove, as depicted in Fig. 1. The entire process was iterated until no novel stopwords were added. It is worth mentioning that, in this phase, LDA was chosen as the model guiding the stopwords list enrichment mainly for its computational time, which is lower compared to BTM, SATM, and BERTopic, at least for their currently available implementation.

The average length of the reviews, defined as the average number of terms contained in the documents, and the size of the dictionary, given by the number of distinct words, obtained after preprocessing are provided in Table 2 for each product category.

### 3.4 Selection of the number of topics

The number $K$ of topics to be inferred is the main hyperparameter provided in input to any topic modeling and plays a fundamental role. Indeed, an insufficient number of topics may generate coarse models, in which the meaning behind each topic is too broad. On the other hand, if $K$ is too large, the model may lose effectiveness by including useless topics or topics sharing too much similarity.

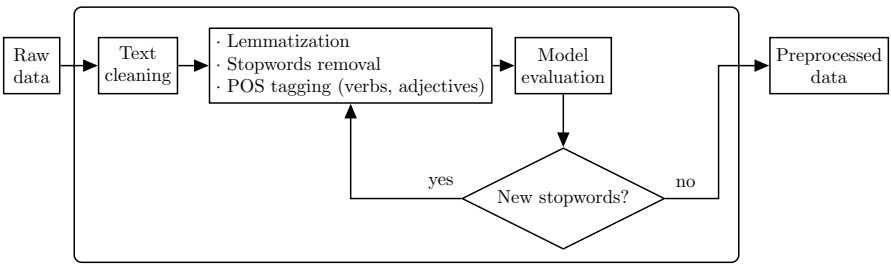



**Fig. 1** Flowchart summarizing the texts preprocessing scheme

Finding the number of topics that enables the best topic model is still an open question in the literature and is not trivial when dealing with unstructured document sets, such as the collection of customer feedback. Indeed, compared to corpora for which the number of topics is known a priori, such as manually curated collections, for unstructured document sets the number of relevant themes is unknown. The common strategy relies on an iterative approach, in which different models are obtained by varying the number of topics and the best-quality model is then selected, where the quality can be computed through various measurements based, for instance, on the notion of perplexity, as in [20], or on more comprehensive indices [68].

Among the quality criteria proposed so far, our attention was drawn to metrics grasping the coherence of the topics, where topics, here given by return causes, are said to be coherent if they support each other [69]. Coherence measures were investigated in scientific philosophy as a way to quantify how well pieces of information hang and fit together [70]. In particular, a set made of coherent topics can be interpreted in a context covering all or most of them. The goal of getting interpretable return reasons was, precisely, the main motivation for our choice.

To quantify the coherence of the topics, four different metrics, denoted as $C_{UCI}$, $U_{MASS}$, $C_{NPMI}$ and $C_V$, were employed. $C_{UCI}$ uses the point-wise mutual information (PMI) between two words and the word co-occurrence counts in a sliding window that moves over an external corpus, such as Wikipedia. $U_{MASS}$ is based on the document co-occurrence and counts the number of documents containing two words over the same corpus used to train the topic model. $C_{NPMI}$ resorts to the normalized PMI (NPMI), whereas $C_V$, proposed more recently, is a combination of the cosine measure, the NPMI and the boolean sliding window. For a detailed description of these criteria the reader may refer to [69] and the references therein. All of these metrics were proven to generate topic ratings, according to their quality, which are close to human coherence judgments. However, since they perform differently on distinct datasets, and a dominant criterion does not exist, the final decision was taken by comparing the scores of the four metrics valued as a whole.

Examples of score curves computed for each coherence metric and topic modeling are shown in Fig. 2, where higher scores are supposedly assigned to models revealing more coherent topics. These curves report the normalized coherence values and were built by varying the number of topics $K$ for each topic modeling in the interval (1, 20). For readability reasons, Fig. 2 collects the results for one product category, i.e., “Cell Phones and Accessories”. The curves for all categories are reported in the supplementary materials.

Different behaviors can be observed for the four topic models. In particular, LDA shows consistent similarities in the trends of all metrics with the exception of $U_{MASS}$. These metrics show a similar dampened upward trend, whereas $U_{MASS}$ exhibits an opposite behavior. This is observed across all categories with the exception of “Luxury Beauty”. A similar interpretation holds also for SATM and BTM, although the two models, especially BTM, present higher fluctuations in the curves and differ in their trend for a higher number of product categories (e.g., “Amazon Fashion” for both SATM and BTM, “Electronics” and “Musical Instruments” for BTM). All metrics show an entirely different behavior for BERTopic where, across most

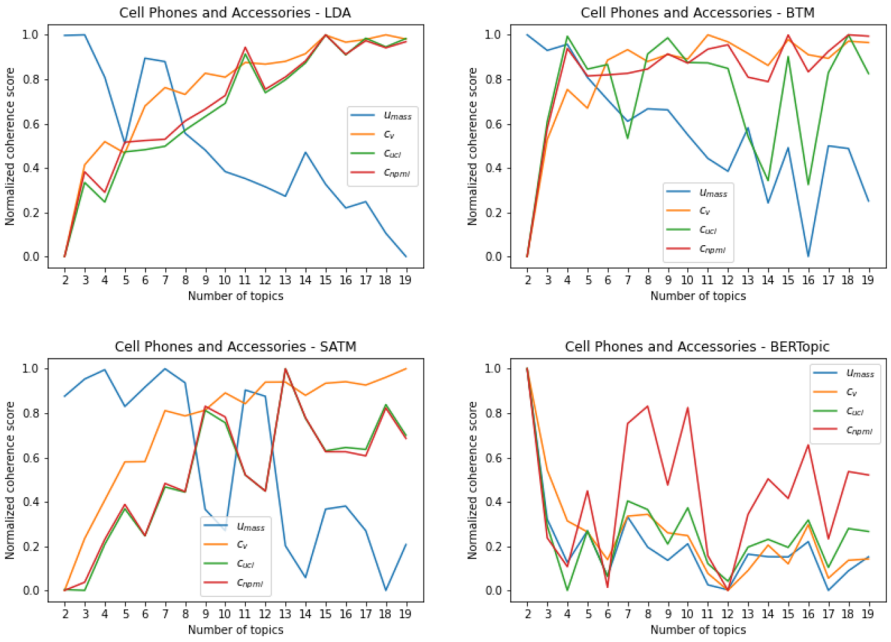



**Fig. 2** Curves of the min-max normalized coherence scores for each topic modeling applied on the reviews in the category "Cell Phones and Accessories". The curves were obtained by varying the number $K$ of topics in the range (1, 20), with step size one

categories, all the curves typically have a downward trend dampening for low values in the number of topics.

The coherence scores were used to identify the regions, i.e., intervals for $K$, where to focus the selection of the most promising number of topics. The final decision was then taken by comparing the quality of the models generated in these intervals in terms of interpretability of the resulting topics. The number of topics ultimately set for each topic modeling and product category is reported in Table 3. Notice that, on average, the value of $K$ deemed appropriate for LDA, based on the coherence measures and the models assessment, was slightly larger than BTM and SATM and on par with that of BERTopic.

To confirm ex-post the appropriateness of our choice, we finally built for each model and category the intertopic distance map, which enables the visualization of the topics in the form of bubbles projected in a two-dimensional space via multidimensional scaling. The size of each bubble is proportional to the prevalence of the topic within the collection of reviews, while the distance between bubbles indicates how similar are the topics they represent, based on the their word distribution. By looking at the maps as the one in Fig. 3, obtained for "Cell Phones and Accessories" by using LDA, we observed that the models based on the selected number of topics gave rise to uniformly large, fairly scattered, and moderately overlapping bubbles, as it is commonly required by topic models of good quality. The plots for all the categories are reported in the supplementary materials.

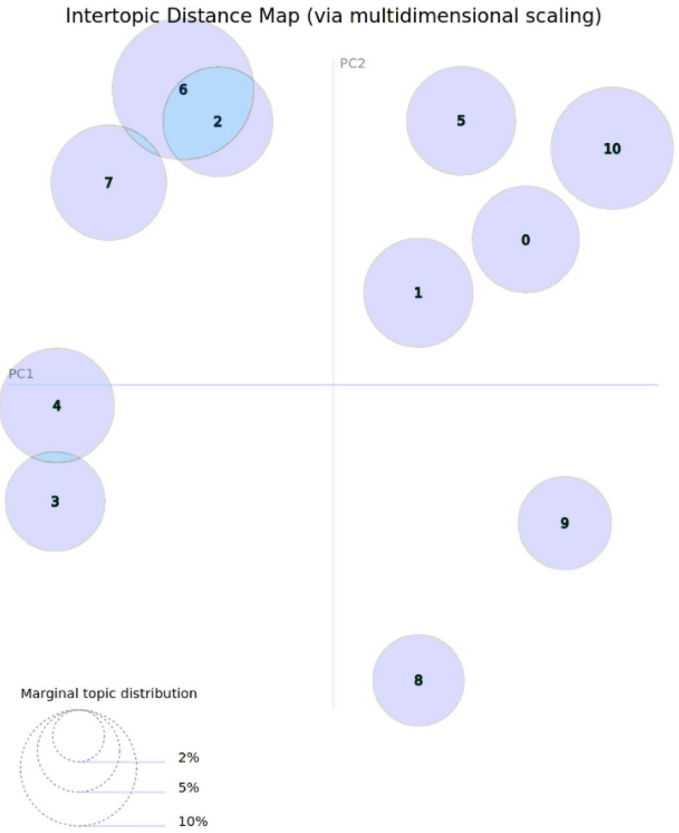



**Fig. 3** Intertopic distance map for LDA applied to the reviews in the category "Cell Phones and Accessories", by setting $K = 11$

**Table 3** Number $K$ of topics set for each topic modeling method in each product category

| Product category | LDA | BTM | SATM | BERTopic |
|---|---|---|---|---|
| Amazon fashion | 5 | 7 | 8 | 7 |
| Automotive | 8 | 4 | 8 | 7 |
| Cell phones and accessories | 11 | 4 | 9 | 7 |
| Clothing shoes and jewelry | 7 | 7 | 9 | 6 |
| Electronics | 8 | 4 | 4 | 9 |
| Home and kitchen | 8 | 8 | 9 | 8 |
| Luxury beauty | 7 | 9 | 5 | 8 |
| Musical instruments | 7 | 4 | 7 | 7 |
| Pet supplies | 9 | 6 | 5 | 5 |
| Sports and outdoors | 10 | 5 | 10 | 12 |
| Tools and home improvement | 8 | 7 | 7 | 12 |
| Video games | 5 | 5 | 2 | 6 |
| Average | 7.8 | 5.8 | 6.9 | 7.8 |

**Table 4** Values of the topic modeling hyperparameters used in LDA, BTM, and SATM

| Hyperparameter | LDA | BTM | SATM |
|---|---|---|---|
| $\alpha$ | 0.1 | 0.1 | 0.1 |
| $\beta$ | 0.01 | 0.01 | 0.01 |
| N. of topic words | 20 | 20 | 20 |
| Max n. of iterations | 1000 | 1000 | 1000 |
| N. of pseudo-documents | - | - | 300 |
| Threshold on $\eta$ | - | - | 0.001 |

**Table 5** Values of the topic modeling hyperparameters used in BERTopic. [a]https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

| Module | Hyperparameter | Value |
|---|---|---|
| Embedding (pre-trained) model | all-MiniLM-L6-v2[a] | |
| | Vector Dimensions | 384 |
| Dimensionality reduction method | UMAP [71] | |
| | Number of neighbors | 15 |
| | Similarity metric | Cosine |
| Clustering method | HDBSCAN [72] | |
| | Minimum cluster size | 15 |
| | Similarity metric | Euclidean |
| | Cluster selection method | Expectation of mass |
| Topic word-representation model | c-TF-IDF [23] | |

### 3.5 Other topic modeling hyperparameters

While the number of topics was tuned for each method, other main hyperparameters were set at their default values. Table 4 collects the parameters set for LDA, BTM, and SATM, given by $\alpha$ and $\beta$, which regulate the document-topic and topic-word density, respectively, the number of words to extract for each topic, the maximum number of iterations and, for SATM, the number of pseudo-documents and a threshold on the probability $\eta$ of the occurrence of a pseudo-document conditioned on a piece of short text snippet, as suggested in [22]. Table 5 shows, instead, the modules used in BERTopic and their hyperparameters. Specifically, the embedding was obtained through a pre-trained model, the dimensionality reduction relied on a manifold learning technique, reduced vectors were clustered according to a hierarchical density-based method, and the topic description was based on a modified version of TF-IDF. Notice that, BERTopic was applied on the same corpus of processed texts used for the other methods, since the semantic descriptions of the topics extracted were found to be more informative compared to those inferred from the unfiltered texts. Moreover, since the default clustering method used in BERTopic is not complete, the comments not associated to any cluster were grouped into the same topic referred to as Topic 0.

## 4 Results and discussion

### 4.1 Evaluation and comparison of the topic models

In the absence of a gold-standard list of topics to be used as benchmark, the evaluation of a topic model is usually performed by "eye-balling" the topics word-list, to analyze the most representative terms associated to each topic. Indeed, the ease of labeling a topic based on the top words in the list can be seen as a proxy of its coherence and degree of interpretability and, hence, as an expression of its quality.

Figure 4 shows the ten most relevant words for each of the topics inferred from the reviews of "Cell Phones and Accessories", by using LDA, BTM, SATM and BERTopic, respectively. By analyzing these terms we were able to devise, for each topic, a textual description, called *return cause label*, which summarizes the return reason expressed by the topic. These labels were used to investigate the potential existence of common return causes across the categories and to analyze the reasons extracted by topic modeling towards the set of options made available on the e-commerce platform, as illustrated in Sect. 4.2.

To set up a comparison across the topic modeling approaches, each review in a product category was assigned to a single topic, namely the topic with the highest estimated probability of occurring in that document. This unique allocation was performed for all the models, in order to evaluate the similarity of the review-to-topic assignment and the consistency of the topic labeling. To this aim, for selected pairs of models we built a contingency table based on the distribution of the reviews over the topics, and graphically displayed the results through heatmaps, as those reported in Fig. 5 for "Cell Phones and Accessories".

Various conclusions can be drawn by analyzing the contingency tables and the related heatmaps. The high incidence of some review-to-topic assignments, combined with the congruence of the most relevant terms describing the topics in different models, validate the interpretation of the topic word lists and support the adequacy of the topic labeling. For instance, the majority of the reviews in "Cell Phones and Accessories" associated to topics discussing quality problems in LDA (i.e., Topic 0 and Topic 8), were assigned to their closest equivalent in SATM (i.e., Topic 7 and Topic 4, respectively), with the mutual pairs of topics sharing top representative terms. Similarly, a notable proportion of reviews were linked in both models to install (LDA-Topic 10 and SATM-Topic 2), battery charging (LDA-Topic 7 and SATM-Topic 5), and audio quality (LDA-Topic 2 and SATM-Topic 0) issues.

The assignments based on BTM were intrinsically affected by the choice of $K$, which turned out to be smaller than the other models. While a certain affinity can be established between BTM and SATM, for which a large part of the customer comments were traced back to battery charging problems (BTM-Topic 1 and SATM-Topic 5), the same is less evident for BTM and LDA. It is worth noticing that BTM assigned the vast majority of the documents to a small subset of return causes out of the entire set of inferred reasons, namely Topic 1 and

**LDA**
Topic 0: **Accessory/replacement quality and compatibility**
`hold open strong remove fell belt close tight magnetic break`
Topic 1: **Installation issue**
`bubble remove instal easy clean clear first apply sure touch`
Topic 2: **Audio quality**
`hear ear sound headset call listen turn loud clear comfortable`
Topic 3: **Accessory/replacement quality and compatibility**
`last stop second first replace defective disappointed send turn start`
Topic 4: **Item defective or doesn't work**
`send defective original tell wrong email replace arrive ask contact`
Topic 5: **Camera image quality**
`easy free recommend honest unbiased perfect lens best review happy`
Topic 6: **Phone signal issue**
`signal mobile android old turn call apps touch sim easy`
Topic 7: **Battery charging**
`charge full charger plug charging different turn second fine stop`
Topic 8: **Product quality issue**
`cheap worth disappointed open arrive wrong poor shipping expect horrible`
Topic 9: **Color issue**
`black white wrong different clear red blue disappointed pink cheap`
Topic 10: **Installation issue**
`hard protect easy thin top difficult soft remove plastic hold`

**BTM**
Topic 0: **Android smartphones**
`android nokia mobile best apps easy setting available include high`
Topic 1: **Battery charging**
`send first last second full sure charge turn different old`
Topic 2: **Installation issue**
`easy hard remove hold top protect open first black clear`
Topic 3: **Audio quality**
`ear sound headset hear listen comfortable different earbuds wear best`

**SATM**
Topic 0: **Audio quality**
`hear ear sound headset call different easy fine poor sure`
Topic 1: **Phone signal issue**
`sim compatible unlocked old turn read mobile signal full first`
Topic 2: **Installation issue**
`hard protect open difficult remove easy top bubble thin cover`
Topic 3: **Color issue**
`black white blue red different pink expect lens clear green`
Topic 4: **Product quality issue**
`cheap disappointed worth arrive poor horrible terrible expect shipping throw`
Topic 5: **Battery charging**
`stop defective last charge second replace win first full fine`
Topic 6: **Review related**
`easy recommend free honest perfect sure unbiased happy review read`
Topic 7: **Accessory/replacement quality and compatibility**
`hold break fell first stick stay close open loose strong`
Topic 8: **Wrong item was sent**
`wrong send refund correct right sure happy galaxy first different`

**Fig. 4** Top ten words (out of the set of 20 words extracted) for each topic inferred by LDA, BTM, SATM, and BERTopic applied on the reviews of the "Cell Phones and Accessories" category

**BERTopic**
Topic 0: *Difficult interpretation (BERTopics' outliers topic)*
`send first hold easy last second charge different hear sure`
Topic 1: **Product quality issue**
`wrong cheap disappointed hard hold open worth different otterbox easy`
Topic 2: **Connection issue - Battery charging**
`signal connect unlocked charge ear iphone android headset sim kindle`
Topic 3: **Accessory/replacement quality and compatibility**
`compatible remote spanish english international straight europe website european venezuela`
Topic 4: **Security issue**
`label hazardous dispose classify flammable local federal send shipping deffective`
Topic 5: **Performance issue**
`turbo adaptive lag laggy fast prouduct turbocharger iceland sail disatisfied`
Topic 6: **Description issue**
`ned nedded folk described describe mistake right wrong`

**Fig. 4** (continued)

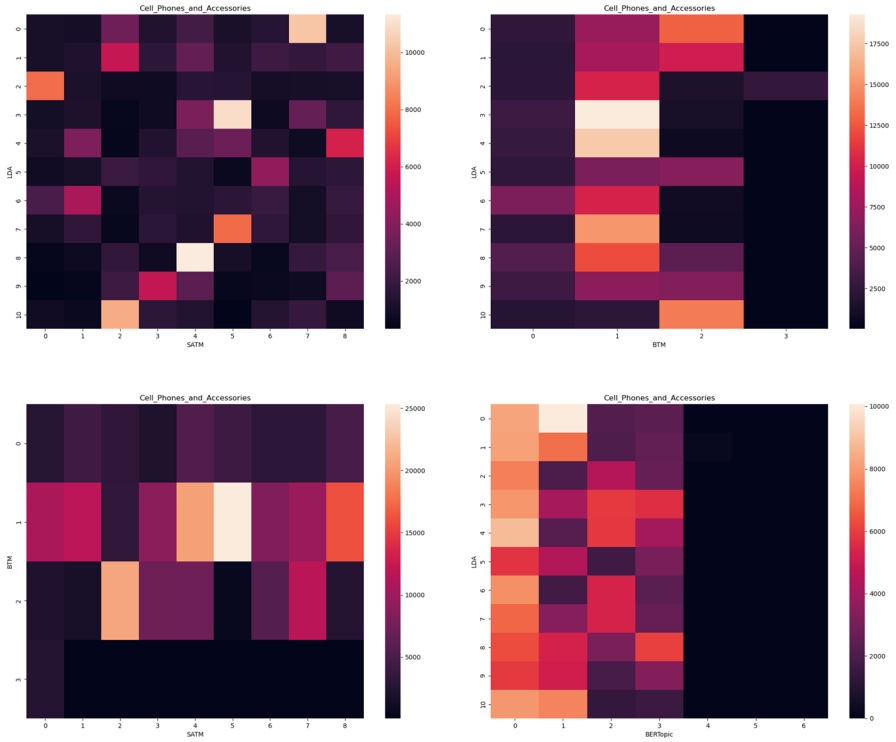



**Fig. 5** Heatmaps representation of the contingency tables used to compare the assignment of the reviews in "Cell Phones and Accessories" across the topics, for selected pairs of topic models. The color scale ranges according to the maximum number of reviews assigned to a pair of topics. Lighter colors denote larger proportions and, therefore, a higher agreement of the models in the reviews allocation to their own topic

Topic 2 for "Cell Phones and Accessories". The same behavior was also shown for BTM models with a larger number of topics, with the majority of the reviews being typically distributed over two or three topics, and is consistent across all the categories. This rules out the case of distinct return causes being forced to fit in a single topic as a result of a wrong choice of $K$. It also denotes the lower capability of BTM, compared to LDA and SATM, to distinguish the topics of the corpus at hand, despite its higher coherence scores.

Various considerations can be made by analyzing the results of BERTopic. First, a considerable portion of the documents were assigned to Topic 0, the one used to label outlying documents. Second, BERTopic-Topic 1 and 3 are observed to discuss similar subjects, i.e., the quality of products and accessories/replacements. A comparison with LDA yields interesting results. Specifically, BERTopic-Topic 1 appears to incorporate also the installation issue, which is present in LDA-Topic 1 and 10. BERTopic-Topic 2 complements BERTopic-Topic 1 and contains documents mostly labeled with LDA-Topic 2, 3, 4, 6, and 7, which discuss issues in the product use, such as poor connection and phone charging. Overall, while BERTopic results appear to be informative, the fact that documents are concentrated in a reduced number of topics makes the interpretation of the topics themselves harder.

As a final remark, we observed that the topics outlined for some product categories sometimes reflect return causes connected to sub-categories. This is, for instance, the case of "Pet Supplies", where some topics seem to be specifically focused on litter boxes, pet food, toys, and leashes. These sub-categories, moreover, can be addressed differently by the distinct models. For example, in "Clothing Shoes and Jewelry" (see Fig. 6), the topics in LDA discussing size-related issues (Topics 0 and 1) seem to refer to footwear and clothes, whereas SATM makes a distinction between footwear comfort and size (Topics 0 and 5) and jeans and dresses size (Topics 2 and 3). These differences have a natural impact on the shapes of the heatmaps, and justify the topic assignment of some reviews. In light of this finding, it is indeed reasonable that comments associated to LDA-Topic 0 (*Sizing issue (footwear)*) were mostly assigned to SATM-Topic 0 and SATM-Topic 5, or that reviews attributed to LDA-Topic 1 (*Sizing issue (clothes)*) were, for the large part, linked to SATM-Topic 2 and SATM-Topic 3.

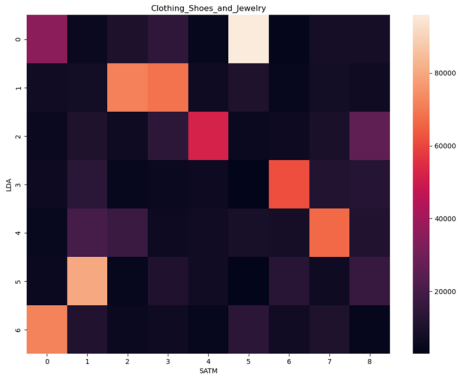



**Fig. 6** Heatmap representation of the review-to-topic assignment for LDA and SATM in the "Clothing Shoes and Jewelry" category, highlighting the similarity in the assignment of the two methodologies

### 4.2 Comparison between currently available and inferred return causes

Given that none of the topic models consistently outperformed the others in terms of model coherence and topic interpretability, and that methods designed for short-texts corpora didn't bring significant benefits, despite their usefulness in validating the results, in this section we focus on the outcomes of classical LDA. This final analysis aims to show how the generic return causes displayed by Amazon to the customers during the return process do not fully capture the actual causes, expressed by the users in the reviews. In particular, the proven dependency of the return reasons on the product category implies the need of devising specific causes for each group of products, to identify the major source of dissatisfaction. However, this dependency is not reflected in the product-agnostic list of options made available to the customers on the e-commerce platform.

The results are shown in Table 6, where the rows report the return causes available at the time of writing (rows 21-32) and those outlined by LDA (rows 1-22), indicated through their own label. Check marks are used in the table to indicate the presence of LDA-inferred causes in the list of available options, as well as their association to the different product categories. Finally, the last column shows the inclusion of each return reason in a list of causes provided by Amazon to third-party sellers to help them in the return process [73]. This list, reported in the supplementary materials, consists of 66 unique return codes which can be grouped, based on their meaning, according to the return reasons reported in Table 6. Although it seems to be intended for back-end management purposes, the comparison of this list with the set of outlined return causes can further shed light on the usefulness of the proposed study.

From the analysis of Table 6 various comments can be made. First, the most frequent return causes revealed from the reviews across all categories refer to malfunctioning, defective or poor-quality products. These causes are properly covered both by the list of available options (label *Item defective or doesn't work*) and the set of extracted causes (labels *Item defective or doesn't work* and *Product quality issue*).

Second, some of the reasons proposed to the customer, and applied equally to all the categories, are sometimes too generic. For example, the option *Inaccurate website description* is unable to grasp the real motivation for a product return in the categories including clothes and footwear, for which issues related to wearability (*Sizing issue (clothes & footwear)*), comfort (*Comfort*) and color (*Color issue*) are instead reported in the customer feedback. For certain categories, therefore, it would be worthwhile to fine-tune the set of available options, to directly find out the product features giving rise to the return.

Third, the table demonstrates how different categories are actually linked to distinct causes, also highlighting the presence of return reasons that are confined to one single category and, as such, that might deserve attention. This is the case, for example, of noisy products within the category "Home and Kitchen" or of smelly goods in "Luxury Beauty".

Fourth, the table shows that a significant number of novel return reasons play indeed a relevant role, being included in the set of causes used by the company with third-party sellers. This achievement confirms the need of making use of a more

**Table 6** Return causes extracted with LDA-based topic modelling (rows 1-22), currently available to customers (rows 21-32), and used with third-party sellers for prepaid returns (checked rows in the last column)

| | Return cause | Available to customers | Amazon fashion | Automotive | Cell phones and accessories | Clothing shoes and jewelry | Electronics | Home and kitchen | Luxury beauty | Musical instruments | Pet supplies | Sports and outdoors | Tools and home improvement | Video games | Available to third-party sellers |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Accessory/Replacement quality and compatibility | | | ✓ | ✓ | | ✓ | | | | ✓ | | | ✓ | ✓ |
| 2 | Audio quality | | | | ✓ | | ✓ | | | ✓ | | | | ✓ | |
| 3 | Battery charging | | | | ✓ | | | | | | | | | | ✓ |
| 4 | Broken product | | | | | | | ✓ | | | | | ✓ | | ✓ |
| 5 | Camera image quality | | | | ✓ | | ✓ | | | | | | | | |
| 6 | Cleaning related | | | ✓ | | | | ✓ | | | ✓ | ✓ | | | |
| 7 | Color issue | | ✓ | | ✓ | ✓ | | ✓ | | | ✓ | ✓ | | | ✓ |
| 8 | Comfort | | ✓ | | | ✓ | | ✓ | | | | ✓ | | | ✓ |
| 9 | Game plot | | | | | | | | | | | | | ✓ | |
| 10 | Graphics quality | | | | | | | | | | | | | ✓ | |

**Table 6** (continued)

| | Return cause | Available to customers | Amazon fashion | Automotive | Cell phones and accessories | Clothing shoes and jewelry | Electronics | Home and kitchen | Luxury beauty | Musical instruments | Pet supplies | Sports and outdoors | Tools and home improvement | Video games | Available to third-party sellers |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 11 | Installation issue | | | | ✓ | | | ✓ | | ✓ | ✓ | ✓ | ✓ | | ✓ |
| 12 | Leaky product | | | | | | | ✓ | | | | ✓ | ✓ | | |
| 13 | Light quality | | | ✓ | | | | | | | | | ✓ | | |
| 14 | Noisy product | | | | | | | ✓ | | | | | | | |
| 15 | Phone signal issue | | | | ✓ | | | | | | | | | | ✓ |
| 16 | Possible counterfeit | | | | | | | | ✓ | | | | | | |
| 17 | Product quality issue | | ✓ | ✓ | ✓ | ✓ | | | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ |
| 18 | Sizing issue (clothes & footwear) | | ✓ | | | ✓ | | | | | | ✓ | | | ✓ |
| 19 | Smelly product | | | | | | | | ✓ | | | | | | |

**Table 6** (continued)

| | Return cause | Available to customers | Amazon fashion | Automotive | Cell phones and accessories | Clothing shoes and jewelry | Electronics | Home and kitchen | Luxury beauty | Musical instruments | Pet supplies | Sports and outdoors | Tools and home improvement | Video games | Available to third-party sellers |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 20 | Software update | | | ✓ | | | ✓ | | | | | ✓ | | | ✓ |
| 21 | Item defective or doesn't work | ✓ | | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ |
| 22 | Wrong item was sent | ✓ | ✓ | ✓ | ✓ | ✓ | | | | | | | | | ✓ |
| 23 | Better price available | ✓ | | | | | | | | | | | | | ✓ |
| 24 | Bought by mistake | ✓ | | | | | | | | | | | | | ✓ |
| 25 | Didn't approve purchase | ✓ | | | | | | | | | | | | | ✓ |
| 26 | Inaccurate website description | ✓ | | | | | | | | | | | | | ✓ |
| 27 | Item arrived too late | ✓ | | | | | | | | | | | | | ✓ |

**Table 6** (continued)

| | Return cause | Available to customers | Amazon fashion | Automotive | Cell phones and accessories | Clothing shoes and jewelry | Electronics | Home and kitchen | Luxury beauty | Musical instruments | Pet supplies | Sports and outdoors | Tools and home improvement | Video games | Available to third-party sellers |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 28 | Missing or broken parts | ✓ | | | | | | | | | | | | | ✓ |
| 29 | No longer needed | ✓ | | | | | | | | | | | | | ✓ |
| 30 | Product and shipping box both damaged | ✓ | | | | | | | | | | | | | |
| 31 | Product damaged, but shipping box OK | ✓ | | | | | | | | | | | | | |
| 32 | Received extra item I didn't buy (no refund needed) | ✓ | | | | | | | | | | | | | ✓ |
| 33 | Phone/carrier related | | | | | | | | | | | | | | ✓ |

**Table 6** (continued)

| | Return cause | Available to customers | Amazon fashion | Automotive | Cell phones and accessories | Clothing shoes and jewelry | Electronics | Home and kitchen | Luxury beauty | Musical instruments | Pet supplies | Sports and outdoors | Tools and home improvement | Video games | Available to third-party sellers |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 34 | Other | | | | | | | | | | | | | | ✓ |

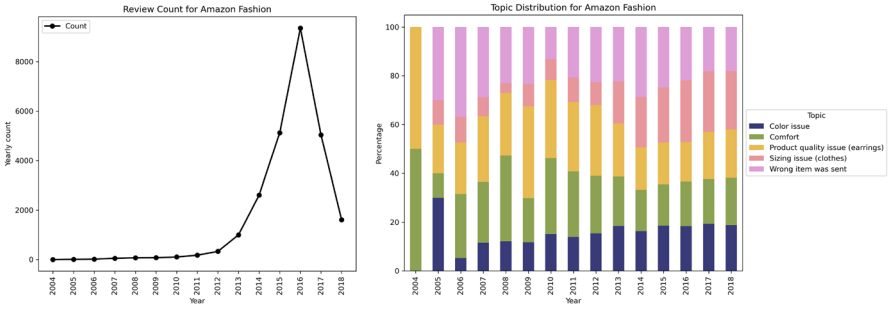
detailed list of reasons to accurately investigate the return causes, and supports the purpose of this study which paves the way for a more comprehensive understanding and management of product returns.

Finally, it is useful to mention that the set of inferred causes is not meant to replace the currently available list of options, since the latter contains reasons that can be hardly detected by topic modeling. In general, all the causes which are not directly connected to the product's features or behavior, such as damaged shipping boxes or delays in the delivery, aren't usually reported in the customer reviews and, therefore, cannot be extracted by language processing tools. Nevertheless, the set of outlined reasons can enrich and complement the currently available options, to mitigate product returns through adequate improvement measures and enhance the customer experience in the return process.

### 4.3 Return causes insights and managerial implications

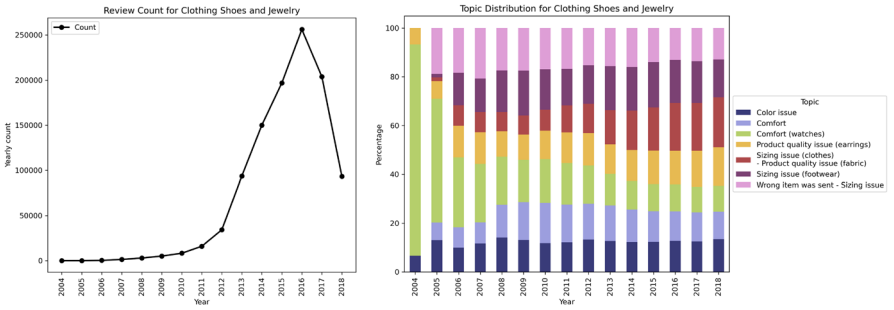
To gather further insights on the return reasons and provide category-level evidences useful from the managerial perspective, we identified some trends occurred in each product category over time, and performed additional analysis by evaluating the sentiment expressed in the selected corpus of reviews.

The distribution of the return causes, together with the overall amount of reviews collected, was investigated for the years 2004-2018 (accounting for 99.8% of the selected reviews), to check for differences among product categories. Figures 7, 8, 9, 10 show the results for categories standing out in this regard (see the supplementary materials for the figures of all categories). Interestingly, for "Amazon Fashion" the main causes of product return progressively shifted from logistic and quality-related aspects to sizing issues (Fig. 7). A similar phenomenon was observed for "Clothing Shoes and Jewelry", for which an increasing incidence of reviews dealing with the items size was recorded, especially by 2013 (Fig. 8). On one hand, these insights show a general improvement of the service offered by the platform since 2004 in the form, for example, of a reduced number of delivered wrong items. On the other hand, they shed light on possible users' behavioural changes due to the



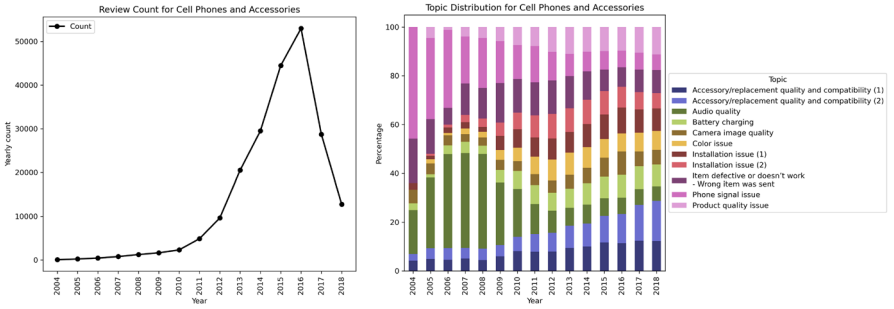

(a) (b)

**Fig. 7** Number of reviews (**a**) and topic distribution (**b**) over time for the "Amazon Fashion" category, highlighting the evolution of review incidence and that of return reasons



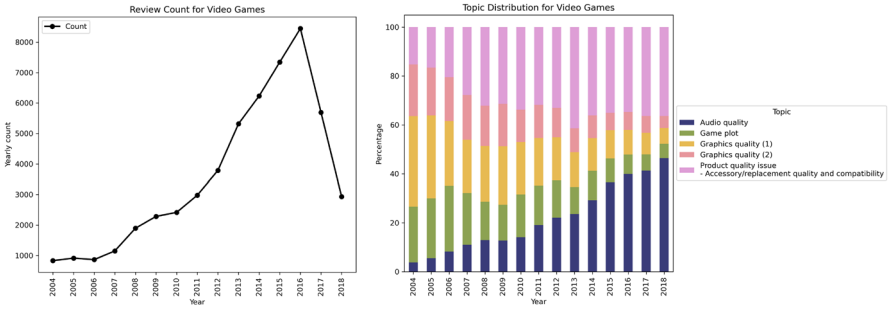

(a) (b)

**Fig. 8** Number of reviews (**a**) and topic distribution (**b**) over time for the "Clothing Shoes and Jewelry" category, highlighting the evolution of review incidence and that of return reasons




(a) (b)

**Fig. 9** Number of reviews (**a**) and topic distribution (**b**) over time for the "Cell Phones and Accessories" category, highlighting the evolution of review incidence and that of return reasons




(a) (b)

**Fig. 10** Number of reviews (**a**) and topic distribution (**b**) over time for the "Video Games" category, highlighting the evolution of review incidence and that of return reasons

lenient return policies, which encourage customers to avoid buying products without trying them first, at the cost of a potential return. Technological advancements might, instead, explain the changing distribution of different return causes for "Cell Phones and Accessories" in the given time horizon. With the spread of smartphones and their rapid increase in performance, audio quality and phone signal gradually became less of an issue, and other return reasons played a more relevant role, such as those related to accessories and replacements, installation (possibly for mobile applications), and battery charging. Even more evident trends were revealed for "Video Games", which might reflect both the technological improvement and the evolution experienced in the market (Fig. 10). Indeed, the decrease in prominence of graphics quality issues was steadily offset by the growing prevalence of audio quality problems, which can be traced back to the advent of online gaming.

The outcome of topic modeling was also combined with sentiment analysis, to evaluate whether some return causes gave origin to a more explicitly manifested dissatisfaction than others. To this aim, we applied VADER (Valence Aware Dictionary for sEntiment Reasoning) [74], a rule-based model able to assess both the polarity (positive, negative, or neutral) and the intensity of the sentiment in a given text. The results of this analysis for the same categories discussed above are shown in part (a) of Figs. 11, 12, 13, 14, whereas the output for all categories are reported in the supplementary materials. The plot referred to a given category displays, for each return reason, the so called compound score, which is a measure of the underlying sentiment of a text comprised between -1 and +1, and computed on the positivity, neutrality, and negativity scores generated by the model. It is worth to mention that, positive and negative sentiments are associated to compound scores $\geq 0.05$ and $\leq -0.05$, respectively, while neutrality is expressed by the remaining values. An interesting, and quite surprising, evidence was the overall tendency of the reviews of unveiling neutral or mildly positive sentiment. At a certain extent, this result may be due to the reviewers' indulgence promoted by the easy and low-expensive return policy currently in place at Amazon, which alleviates customers' irritation toward the purchasing of products that do not match their expectations. On the other hand, a negative average sentiment was observed for three topics referred, respectively, to product quality, defective items, and wrong items delivered. This outcome, consistent for all the categories, entails a strong indication on the customers' most

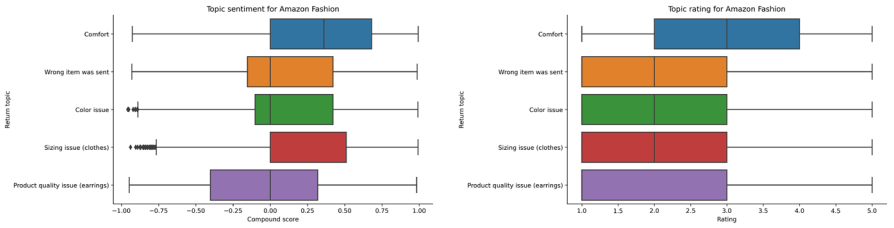
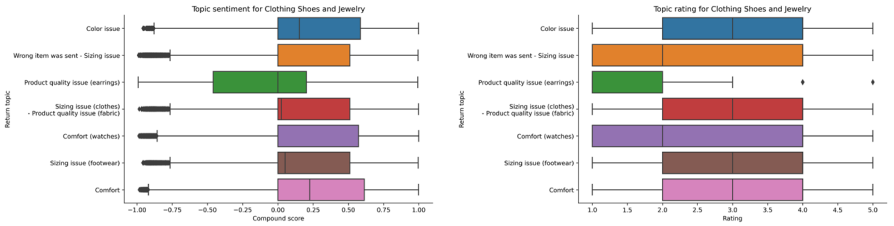
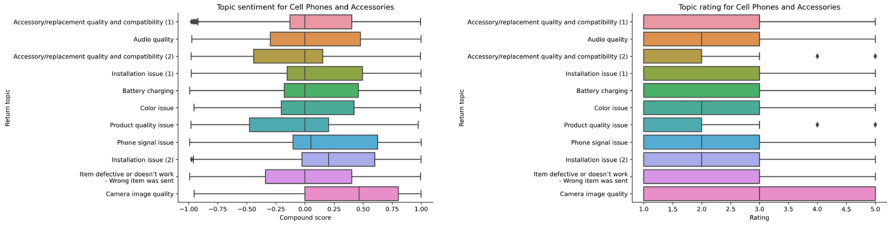



**Fig. 11** Topic sentiment as measured by the VADER compound score (**a**), and rating expressed by the customers (**b**), for the "Amazon Fashion" category



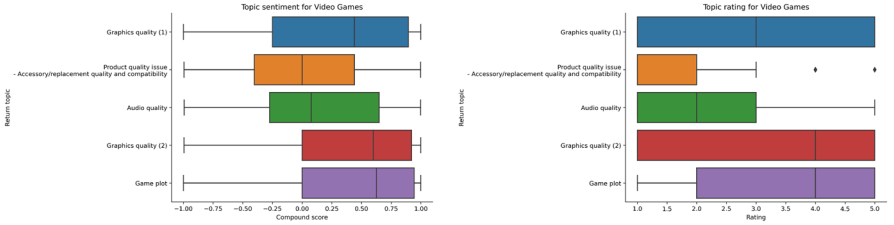

(a) (b)

**Fig. 12** Topic sentiment as measured by the VADER compound score (**a**), and rating expressed by the customers (**b**), for the "Clothing Shoes and Jewelry" category




(a) (b)

**Fig. 13** Topic sentiment as measured by the VADER compound score (**a**), and rating expressed by the customers (**b**), for the "Cell Phones and Accessories" category




(a) (b)

**Fig. 14** Topic sentiment as measured by the VADER compound score (**a**), and rating expressed by the customers (**b**), for the "Video Games" category

displeasing factors and can be leveraged when defining the priority of intervention and drafting the proper corrective actions.

Additional insights were gathered by analysing, for each topic, the distribution of the ratings of the associated reviews (part (b) of Figs. 11, 12, 13, 14). In particular, it was observed that the lowest average rating was assigned to reviews dealing with product quality issues. This result is in line with the sentiment analysis findings and further emphasises the prominent role played by low quality-related return reasons compared to the others.

The distribution of the return causes and the sentiment analysis supported the evidence that, for certain categories, defective or poor-quality products are the focal point of return-related reviews. Going more into detail, the first analysis allowed also to grasp the relative incidence of different product quality dimensions, and how it changed over time. Taken together, these empirical findings have relevant managerial implications, since they help e-commerce platforms in prioritizing the actions useful to contrast product returns.

Despite the complete prevention of quality-related issues is hard in practice, some strategies can be put in place. For example, it would be recommended to define and apply clear and stringent product standards, eventually providing the sellers with training programs and resources to meet the aforementioned standards and offer high-quality products. Applying stricter quality controls to check the integrity of the products before the order fulfillment would be likewise beneficial. This recommended action might also guide the investment in AI-based technologies aimed, for example, at automatically detecting damaged packages, defective parts or counterfeit products. Another useful intervention could be enhancing the product display. In its simplest form, this means providing a more comprehensive and adherent product description, or using high-resolution images which accurately depict the product from multiple angles, to mitigate the misalignment between product key features and customer expectations. More expensive strategies could enable the usage of interactive images or of virtual try-on or Augmented Reality technologies, which allow customers to visualize how the product fits or looks in the target environment. Besides a better examination of the product quality, this action would offer a more immersive shopping experience and would be helpful also to address size-related issues.

Another aspect worthy of attention is the growing number of reviews consistently observed for all product categories until 2016, when any form of "incentivized reviews" tied, for example, to free or discounted products was prohibited by Amazon, except for those within its own invitation-only review program. The constant decrease recorded thereafter doesn't bode well, since customer feedbacks are nowadays recognized as one of the most valuable tools that e-commerce websites offer to foster more informed purchase decisions. In light of this, further actions to be taken could be devoted to design and implement mechanisms to encourage genuine user reviews. For example, by leveraging gamification elements, to make the review process more fun and rewarding, or by showing how past customer feedbacks influenced the buying behaviour or contributed to product improvements.

The present study is focused on a specific online retailer. However, the variety and heterogeneity of the categories of products analyzed make the results potentially useful for other retailers, both online and offline. For example, based on the findings that the overall wearability of a garment plays a prominent role, retailers in the clothing sector may refine their offering to include products promoting good fit design and comfort, or may inform marketing campaigns highlighting these features. As another example, while preserving the quality of the products offered at all times, retailers of consumer electronics may enhance customer service by providing more accurate assistance, based on a deeper knowledge on products accessories, replacements and installation. It is worthwhile to notice that also the methodological

framework here presented is not a prerogative of e-commerce retailers directly collecting customers feedback in their portal, but can be proficiently adapted to different retail environments and purposes. Consider, for example, the possibility of analysing the corpus of customer reviews gathered by third parties, such as search engines or social networks on the Web, with the aim of identifying the main topics therein addressed to understand what customers appreciate about the products/services offered, where there may be areas for improvement, or to compare the content of the reviews with those of competitors to identify potential strengths and weaknes ses.

## 5 Conclusions and future works

In this paper topic modeling was applied to a corpus of customer reviews with the aim of revealing the main reasons for product returns for an online retailer. In particular, different topic modeling were used and compared on the publicly available Amazon Review Dataset, from which a subset of textual documents (i.e., reviews), specifically focused on returned products, was extracted.

The proposed methodology enabled the discovery of rich category-dependent sets of return reasons that can be potentially exploited by e-commerce companies to reduce the rate of product returns and the cost of the reverse logistic network. Indeed, a deeper knowledge on the return root causes can help online retailers to act proactively; for example, by improving the information made available for each product through more adherent product descriptions, by better supporting the customer return activities in the post-purchase phase, and by implementing more informed returns management strategies.

The findings of the current study must be evaluated in light of two major limitations. From one side, results were achieved and discussed for a unique set of customer reviews. Although the original Amazon Review Dataset is rather large (233.1 million reviews) and comprehends a wide variety of product categories, collecting further outcomes on additional data sources would corroborate the robustness and generalizability of the methodology at hand. To the best of our knowledge, however, no dataset of comparable size, and informative keywords, made up by textual customers' feedback is publicly available at the time of writing. From the other, the work is focused on the usage of topic modeling techniques and, therefore, naturally inherits their limitations and drawbacks. Among these, the sensitivity to the text preprocessing steps (i.e., tokenization, stop-word removal, stemming or lemmatization), and the difficulty of selecting the number of topics and providing an interpretation for all of them. These elements prevent the process of return causes extraction from being completely automatized and, by requiring human judgment in some of its phases, still implies a human-in-the-loop analytical approach.

The present study can be extended in several directions, from both a methodological and managerial perspective. As formerly described, the analysis relied on a subset of texts obtained from an entire collection of customer reviews and deemed related to returned products. As the first development, the selection of this sub-corpus could be refined; for example, by leveraging the feedback ratings or their sentiment, to isolate

the reviews expressing negative sentiments, or by training text classification models able to distinguish between reviews referred or not to product returns. Moreover, for some product categories the topics extracted through topic modeling seemed to be connected to specific products or product sub-categories. A more granular analytical method addressing these sub-categories may be useful to manage this aspect. As a further research extension, the use of hierarchical topic modeling could be explored, to model the potential hierarchical nature of the topics, in order to analyze more in depth the inter-topic relationships and identify ancestor topics in the hierarchy that represent conceptually broader versions of their descendants. Additionally, the effectiveness of alternative word-embedding strategies to be used in conjunction with topic modeling could be investigated, in the attempt to produce better-quality topics and, therefore, more accurate return causes. Metadata, when captured and made available, could also be exploited for topics extraction by resorting to alternative techniques, such as Structural Topic Models (STM, see [75–77]), able to account for such information, with the aim of achieving better separated topics. In addition, metadata such as the ID of the reviewers, together with the number of helpful votes their feedback received, could be analysed to identify, within each product category, "influential reviewers". By focusing on the content of their comments, it would be possible to highlight further positive, or negative, product features attracting the interest of the customers. Likewise, the images accompanying the textual comments could be exploited to extract additional information that can enrich both topic modeling and sentiment analysis.

From the managerial perspective, it would be useful to deepen this study by collecting information such as the economic value of the returned products, the product life-cycle, and the location of the returning user. By combining these data with the topic modeling outcome, it would be then possible to perform cost-benefit analysis and fine-tune the priority of the return mitigation strategies. Finally, extending the scope of the study beyond topic modeling to embrace a multidisciplinary approach, it would be worthwhile to incorporate behavioural economics theories with the aim of investigating, among others, whether customer feedbacks on returned products, and therefore the return reasons therein expressed, are somehow subject to systematic biases or dependent on customers emotions.

**Acknowledgements** The authors wish to express their gratitude to the anonymous referees who provided very useful and detailed comments on a previous version of the manuscript.

The present study has been developed within the HumanTech Project, which is financed by the Italian Ministry of University and Research (MUR) for the 2023-2027 period as part of the ministerial initiative "Departments of Excellence" (L. 232/2016).

**Funding** Open access funding provided by Politecnico di Milano within the CRUI-CARE Agreement.

## Declarations

**Conflict of interest** On behalf of all authors, the corresponding author states that there is no Conflict of interest.

## References

1. Rokonuzzaman, M., Iyer, P., & Harun, A. (2021). Return policy, No joke: An investigation into the impact of a retailer's return policy on consumers' decision making. *Journal of Retailing and Consumer Services, 59*, 102346. https://doi.org/10.1016/j.jretconser.2020.102346
2. RetailDive. (2021). The Right Fit: How AI is changing ecommerce apparel returns. Retrieved from https://www.retaildive.com/library/rakuten-whitepaper-the-right-fit/. Accessed 31/08/2023.
3. Janakiraman, N., Syrdal, H., & Freling, R. E. (2016). How to design a return policy. *Harvard Business Review, 2*, 2–5.
4. Blanchard, D. (2007). Supply chains also work in reverse. Industry Week **1**, 48–49. Retrieved from https://www.industryweek.com/supply-chain/planning-forecasting/article/21954433/supply-chains-also-work-in-reverse. Accessed 31/08/2023.
5. National Retail Federation. (2021). \$428 billion in merchandise returned in 2020. Retrieved from https://nrf.com/media-center/press-releases/428-billion-merchandise-returned-2020. Accessed 31/08/2023.
6. Petersen, J. A., & Kumar, V. (2010). Can product returns make you money? *MIT Sloan Management Review, 51*, 85.
7. Guide, V. D. R., Souza, G. C., Van Wassenhove, L. N., & Blackburn, J. D. (2006). Time value of commercial product returns. *Management Science, 52*, 1200–1214. https://doi.org/10.1287/mnsc.1060.0522
8. Blackburn, J. D., Guide, V. D. R., Souza, G. C., & Van Wassenhove, L. N. (2004). Reverse supply chains for commercial returns. *California Management Review, 46*, 6–22. https://doi.org/10.2307/41166207
9. Guerinot, M. (2021). Grow a sustainable business through returns . https://www.optoro.com/returns-blog/grow-a-sustainable-business-through-returns/. [Online; accessed 16-February-2024].
10. Zhang, D., Frei, R., Wills, G., Gerding, E., Bayer, S., & Senyo, P. K. (2023). Strategies and practices to reduce the ecological impact of product returns: An environmental sustainability framework for multichannel retail. *Business Strategy and the Environment*. https://doi.org/10.1002/bse.3385
11. Gustafsson, E., Jonsson, P., & Holmström, J. (2021). Reducing retail supply chain costs of product returns using digital product fitting. *International Journal of Physical Distribution & Logistics Management, 51*, 877–896. https://doi.org/10.1108/ijpdlm-10-2020-0334
12. García-Moya, L., Anaya-Sánchez, H., & Berlanga-Llavori, R. (2013). Retrieving product features and opinions from customer reviews. *IEEE Intelligent Systems, 28*, 19–27. https://doi.org/10.1109/MIS.2013.37
13. Zhan, J., Loh, H. T., & Liu, Y. (2009). Gather customer concerns from online product reviews - A text summarization approach. *Expert Systems with Applications, 36*, 2107–2115. https://doi.org/10.1016/j.eswa.2007.12.039
14. Chen, B., & Chen, J. (2017). When to introduce an online channel, and offer money back guarantees and personalized pricing? *European Journal of Operational Research, 257*, 614–624. https://doi.org/10.1016/j.ejor.2016.07.031
15. Chintagunta, P. K., Gopinath, S., & Venkataraman, S. (2010). The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets. *Marketing Science, 29*, 944–957. https://doi.org/10.1287/mksc.1100.0572
16. Duan, W., Gu, B., & Whinston, A. B. (2008). The dynamics of online word-of-mouth and product sales-An empirical investigation of the movie industry. *Journal of Retailing, 84*, 233–242. https://doi.org/10.1016/j.jretai.2008.04.005
17. Cui, G., Lui, H.-K., & Guo, X. (2012). The effect of online consumer reviews on new product sales. *International Journal of Electronic Commerce, 17*, 39–57. https://doi.org/10.2753/JEC1086-4415170102

18. Cuffie, H.G., Najar, R.I., & Khasawneh, M.T. (2020). Topic modeling for customer returns retail data. In: Proceedings of the 2020 IISE Annual Conference, New Orleans, Louisiana, USA. Retrieved from https://www.proquest.com/scholarly-journals/topic-modeling-customer-returns-retail-data/docview/2522430747/se-2
19. Ni, J., Li, J., & McAuley, J. (2019). Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, China. https://doi.org/10.18653/v1/D19-1018
20. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.
21. Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013). A biterm topic model for short texts. In: Proceedings of the 22nd International Conference on World Wide Web, Rio de Janeiro, Brazil. https://doi.org/10.1145/2488388.2488514
22. Quan, X., Kit, C., Ge, Y., & Pan, S.J. (2015). Short and sparse text topic modeling via self-aggregation. In: Proceedings of the 24th International Conference on Artificial Intelligence, Buenos Aires, Argentina. https://doi.org/10.5555/2832415.2832564
23. Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. https://doi.org/10.48550/arXiv.2203.05794
24. Ambilkar, P., Dohale, V., Gunasekaran, A., & Bilolikar, V. (2022). Product returns management: A comprehensive review and future research agenda. *International Journal of Production Research, 60*, 3920–3944. https://doi.org/10.1080/00207543.2021.1933645
25. Duong, Q. H., Zhou, L., Meng, M., Van Nguyen, T., Ieromonachou, P., & Nguyen, D. T. (2022). Understanding product returns: A systematic literature review using machine learning and bibliometric analysis. *International Journal of Production Economics, 243*, 108340. https://doi.org/10.1016/j.ijpe.2021.108340
26. Sun, M., Chen, J., Tian, Y., & Yan, Y. (2021). The impact of online reviews in the presence of customer returns. *International Journal of Production Economics, 232*, 107929. https://doi.org/10.1016/j.ijpe.2020.107929
27. Pei, Z., Paswan, A., & Yan, R. (2014). E-tailer's return policy, consumer's perception of return policy fairness and purchase intention. *Journal of Retailing and Consumer Services, 21*, 249–257. https://doi.org/10.1016/j.jretconser.2014.01.004
28. Wood, S. L. (2001). Remote purchase environments: The influence of return policy leniency on two-stage decision processes. *Journal of Marketing Research, 38*, 157–169. https://doi.org/10.1509/jmkr.38.2.157.18847
29. Su, X. (2009). Consumer returns policies and supply chain performance. *Manufacturing & Service Operations Management, 11*, 595–612. https://doi.org/10.1287/msom.1080.0240
30. Choi, T.-M., Liu, N., Ren, S., & Hui, C.-L. (2013). No refund or full refund: When should a fashion brand offer full refund consumer return service for mass customization products? *Mathematical Problems in Engineering*. https://doi.org/10.1155/2013/561846
31. Hsiao, L., & Chen, Y.-J. (2015). Retailer's rationale to refuse consumer returns in supply chains. *Naval Research Logistics (NRL), 62*, 686–701. https://doi.org/10.1002/nav.21673
32. Hong, Y. K., & Pavlou, P. A. (2014). Product fit uncertainty in online markets: Nature, effects, and antecedents. *Information Systems Research, 25*, 328–344. https://doi.org/10.1287/isre.2014.0520
33. De, P., Hu, Y. J., & Rahman, M. S. (2013). Product-oriented web technologies and product returns: An exploratory study. *Information Systems Research, 24*, 998–1010. https://doi.org/10.1287/isre.2013.0487
34. Chen, B., & Chen, J. (2017). Compete in price or service?-A study of personalized pricing and money back guarantees. *Journal of Retailing, 93*, 154–171. https://doi.org/10.1016/j.jretai.2016.12.005
35. Ketzenberg, M. E., Abbey, J. D., Heim, G. R., & Kumar, S. (2020). Assessing customer return behaviors through data analytics. *Journal of Operations Management, 66*, 622–645. https://doi.org/10.1002/joom.1086
36. Dellarocas, C. (2003). The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management Science, 49*, 1407–1424. https://doi.org/10.1287/mnsc.49.10.1407.17308

37. Chen, Y., & Xie, J. (2008). Online consumer review: Word-of-mouth as a new element of marketing communication mix. *Management Science, 54*, 477–491. https://doi.org/10.1287/mnsc.1070.0810
38. Li, X., & Hitt, L. M. (2008). Self-selection and information role of online product reviews. *Information Systems Research, 19*, 456–474. https://doi.org/10.1287/isre.1070.0154
39. Sahoo, N., Dellarocas, C., & Srinivasan, S. (2018). The impact of online product reviews on product returns. *Information Systems Research, 29*, 723–738. https://doi.org/10.1287/isre.2017.0736
40. Zhou, W., & Duan, W. (2016). Do professional reviews affect online user choices through user reviews? An empirical study. *Journal of Management Information Systems, 33*, 202–228. https://doi.org/10.1080/07421222.2016.1172460
41. Markopoulos, P. M., Aron, R., & Ungar, L. H. (2016). Product information websites: Are they good for consumers? *Journal of Management Information Systems, 33*, 624–651. https://doi.org/10.1080/07421222.2016.1243885
42. Sun, M. (2012). How does the variance of product ratings matter? *Management Science, 58*, 696–707. https://doi.org/10.1287/mnsc.1110.1458
43. Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research, 43*, 345–354. https://doi.org/10.1509/jmkr.43.3.345
44. Forman, C., Ghose, A., & Wiesenfeld, B. (2008). Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Information Systems Research, 19*, 291–313. https://doi.org/10.1287/isre.1080.0193
45. Liu, Y. (2006). Word of mouth for movies: Its dynamics and impact on box office revenue. *Journal of Marketing, 70*, 74–89. https://doi.org/10.1509/jmkg.70.3.074
46. Dellarocas, C., Zhang, X., & Awad, N. F. (2007). Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *Journal of Interactive Marketing, 21*, 23–45. https://doi.org/10.1002/dir.20087
47. Godes, D., & Mayzlin, D. (2004). Using online conversations to study word-of-mouth communication. *Marketing Science, 23*, 545–560. https://doi.org/10.1287/mksc.1040.0071
48. Kostyra, D. S., Reiner, J., Natter, M., & Klapper, D. (2016). Decomposing the effects of online customer reviews on brand, price, and product attributes. *International Journal of Research in Marketing, 33*, 11–26. https://doi.org/10.1016/j.ijresmar.2014.12.004
49. Lohse, T., Kemper, J., & Brettel, M. (2017). How online customer reviews affect sales and return behavior - an empirical analysis in fashion e-commerce. In: Proceedings of the 25th European Conference on Information Systems, Guimarães, Portugal. Retrieved from https://aisel.aisnet.org/ecis2017_rip/16
50. Minnema, A., Bijmolt, T. H., Gensler, S., & Wiesel, T. (2016). To Keep or Not to Keep: Effects of Online Customer Reviews on Product Returns. *Journal of Retailing, 92*, 253–267. https://doi.org/10.1016/j.jretai.2016.03.001
51. Dellarocas, C., & Wood, C. A. (2008). The sound of silence in online feedback: Estimating trading risks in the presence of reporting bias. *Management Science, 54*, 460–476. https://doi.org/10.1287/mnsc.1070.0747
52. Hu, M., & Liu, B. (2004). Mining opinion features in customer reviews. In: Proceedings of the 19th National Conference on Artificial Intelligence, San Jose, California. Retrieved from https://dl.acm.org/doi/10.5555/1597148.1597269
53. Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA. Retrieved from https://dl.acm.org/doi/10.1145/1014052.1014073
54. Popescu, A.-M., & Etzioni, O. (2005). Extracting product features and opinions from reviews. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, Vancouver, British Columbia, Canada. https://doi.org/10.3115/1220575.1220618
55. Akella, K., Venkatachalam, N., Gokul, K., Choi, K., & Tyakal, R. (2017). Gain customer insights using NLP techniques. *SAE International Journal of Materials and Manufacturing, 10*, 333–337. https://doi.org/10.4271/2017-01-0245
56. Ramaswamy, S., & DeClerck, N. (2018). Customer perception analysis using deep learning and NLP. *Procedia Computer Science, 140*, 170–178. https://doi.org/10.1016/j.procs.2018.10.326
57. Filieri, R., Lin, Z., Li, Y., Lu, X., & Yang, X. (2022). Customer emotions in service robot encounters: A hybrid machine-human intelligence approach. *Journal of Service Research, 25*, 614–629. https://doi.org/10.1177/10946705221103937

58. Büschken, J., & Allenby, G. M. (2016). Sentence-based text analysis for customer reviews. *Marketing Science, 35*, 953–975. https://doi.org/10.1287/mksc.2016.0993
59. Titov, I., & McDonald, R. (2008). Modeling online reviews with multi-grain topic models. In: Proceedings of the 17th International Conference on World Wide Web, Beijing, China. https://doi.org/10.1145/1367497.1367513
60. Zhao, W.X., Jiang, J., Yan, H., & Li, X. (2010). Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Cambridge, Massachusetts. Retrieved from https://dl.acm.org/doi/10.5555/1870658.1870664
61. Piris, Y., & Gay, A.-C. (2021). Customer satisfaction and natural language processing. *Journal of Business Research, 124*, 264–271. https://doi.org/10.1016/j.jbusres.2020.11.065
62. Duan, X., Li, J., & Chen, Y. (2020). Analysis of amazon market product satisfaction based on LDA theme model. In: 2020 International Conference on Computer Vision, Image and Deep Learning, Chongqing, China. https://doi.org/10.1109/CVIDL51233.2020.00048
63. Huang, M., Wen, S., Jiang, M., & Yao, Y. (2021). LDA topic mining of light food customer reviews on the Meituan platform. In: Tan, Y., Shi, Y., Zomaya, A., Yan, H., Cai, J. (eds.) Data Mining and Big Data, pp. 108–121. Springer, ???.
64. Hai, Z., Chang, K., Kim, J.-J., & Yang, C. C. (2014). Identifying features in opinion mining via intrinsic and extrinsic domain relevance. *IEEE Transactions on Knowledge and Data Engineering, 26*, 623–634. https://doi.org/10.1109/TKDE.2013.26
65. Qiang, J., Qian, Z., Li, Y., Yuan, Y., & Wu, X. (2022). Short text topic modeling techniques, applications, and performance: A survey. *IEEE Transactions on Knowledge and Data Engineering, 34*, 1427–1445. https://doi.org/10.1109/TKDE.2020.2992485
66. Cheng, X., Yan, X., Lan, Y., & Guo, J. (2014). BTM: Topic modeling over short texts. *IEEE Transactions on Knowledge and Data Engineering, 26*, 2928–2941. https://doi.org/10.1109/tkde.2014.2313872
67. Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media, Inc. Available at https://www.nltk.org/book/. Accessed 31/08/2023.
68. Gan, J., & Qi, Y. (2021). Selection of the optimal number of topics for LDA topic model-taking patent policy analysis as an example. *Entropy (Basel), 23*, 1301. https://doi.org/10.3390/e23101301
69. Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, Shanghai, China. https://doi.org/10.1145/2684822.2685324
70. Douven, I., & Meijs, W. (2007). Measuring coherence. *Synthese, 156*, 405–425. https://doi.org/10.1007/s11229-006-9131-z
71. McInnes, L., Healy, J., & Melville, J. (2020). UMAP: Uniform manifold approximation and projection for dimension reduction. https://doi.org/10.48550/arXiv.1802.03426
72. Campello, R.J.G.B., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates, pp. 160–172. Springer. https://doi.org/10.1007/978-3-642-37456-2_14
73. Amazon. (2021). Return reason codes for prepaid returns. Retrieved from https://sellercentral.amazon.com/gp/help/external/202080050?language=en_US&ref=efph_202080050_cont_202072200. Accessed 31/08/2023.
74. Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 8, pp. 216–225.https://doi.org/10.1609/icwsm.v8i1.14550
75. Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). Stm: An r package for structural topic models. *Journal of Statistical Software, 91*, 1–40. https://doi.org/10.18637/jss.v091.i02
76. Hu, N., Zhang, T., Gao, B., & Bose, I. (2019). What do hotel customers complain about? text analysis using structural topic model. *Tourism Management, 72*, 417–426. https://doi.org/10.1016/j.tourman.2019.01.002
77. Biswas, B., Sengupta, P., Kumar, A., Delen, D., & Gupta, S. (2022). A critical assessment of consumer reviews: A hybrid nlp-based methodology. *Decision Support Systems, 159*, 113799. https://doi.org/10.1016/j.dss.2022.113799

## Authors and Affiliations

**Andrea Mor[1] · Carlotta Orsenigo[1] · Mauricio Soto Gomez[2] · Carlo Vercellis[1]**

✉ Andrea Mor
andrea.mor@polimi.it

Carlotta Orsenigo
carlotta.orsenigo@polimi.it

Mauricio Soto Gomez
mauricio.soto@unimi.it

Carlo Vercellis
carlo.vercellis@polimi.it

[1] Department of Management, Economics and Industrial Engineering, Politecnico di Milano, Via Raffaele Lambruschini, 4/B, 20156 Milan, Italy

[2] Department of Computer Science, Universitá degli Studi di Milano, Via Celoria, 18, 20133 Milan, Italy