

Why Do Customers Return Products?

Using Customer Reviews to Predict Product Return Behaviors

Hao-Fei Cheng
hchenga@amazon.com
Amazon
USA

Eyal Krikon
krikon@amazon.com
Amazon
USA

Vanessa Murdock
vmurdock@amazon.com
Amazon
USA

ABSTRACT

Product returns are an increasing environmental problem, as an estimated 25% of returned products end up as landfill [10]. Returns are expensive for retailers as well, and it is estimated that 15-40% of all online purchases are returned [34]. The problem could be mitigated by identifying issues with a product that are likely to lead to its return, before many have sold. Understanding and predicting return reasons can help identify manufacturing defects, misleading information in the product description or reviews, issues with a seller or shipping company, and customers who are habitual returners. While there has been much work to identify and predict return volume, little attention has been given to the reasons for the return. In this paper we explore how customer reviews could be used as signals to identify return reasons. We developed a multi-class classifier to predict return reasons, with a fine-tuned BERT-based model to encode customer review text as features. The classifier with customer review text yields an increase of more than 20% average precision over the baseline classifier with no reviews text. We also showed that we can use aggregated review information to predict product return in case the customer returning the product did not write a review. Lastly we show that reviews can be used to identify nuanced return reasons beyond what the customer indicated.

ACM Reference Format:

Hao-Fei Cheng, Eyal Krikon, and Vanessa Murdock. 2024. Why Do Customers Return Products? Using Customer Reviews to Predict Product Return Behaviors. In *Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '24)*, March 10–14, 2024, Sheffield, United Kingdom. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3627508.3638326>

1 INTRODUCTION

People are ordering more online, a trend that has been in place for 20+ years [24, 36]. During the 2020 SARS-CoV-2 pandemic, the trend accelerated as people began shopping online instead of in physical stores, with a 30% average traffic growth to e-commerce retailers in 2020 [43]. With this rise comes an increase in product returns because people cannot see or try a product prior to buying it, and instead purchase it speculatively and return it when they are not satisfied. By some estimates, during the winter holiday season in 2020, \$70 Billion in online purchases were returned, a

73% increase over the average of the previous five years [10]. While customers understandably appreciate the convenience of liberal return policies, these policies are problematic for companies and for the environment. In addition to lost revenue from the sale of the product, returns are expensive to process, and lead to waste [1, 9, 37]. While some of the returned products are resold, an estimated 25% are sent to landfills [10].

If sellers and e-commerce retailers could predict returns in advance, they could mitigate the situation by insisting on a higher manufacturing standard, contract with more reliable shipping companies, or adjust the online representation of the product to be more informative or representative of the physical product. To minimize returns, sellers would need to identify return reasons as early as possible after a product is offered. Unfortunately, this information normally is available only after a significant number of products have been sent back. There has been a lot of interest in predicting return volume [5, 15, 42] and return rate of specific categories of products [7, 19], but these studies do not give insight into the root causes of the returns.

Early product reviews may be a bellwether to identify problems with a product prior to an uptick in returns. This rich source of information has been largely overlooked in previous work predicting product returns. Reviews can be noisy, and often mention a mixture of good and bad qualities of a product. They may not always be reliable assessments of product quality. They are also not always readily available—writing reviews is optional for the customer. While many customers use reviews to voice their experience, many others choose not to do so. In spite of these drawbacks, we found that the review text itself is informative for predicting return reasons.

In this paper we answer the question of how to utilize customer reviews in predicting the customer return reasons and identifying product issues. Our contributions are two-fold. First, we show that reviews can be used as features to predict the reasons why customers are returning products they purchase. We show that this can be done on a individual purchase-level before a customer returns a product, or at a macro-level to identify the common issues of a product that lead to returns. Second, we built aspect extraction models which identify common attributes of products mentioned in reviews. We show that these extracted aspects reflect more specific details about the customer dissatisfaction with their purchase, beyond the categorical selections solicited by the e-commerce retailer.

The rest of the paper is organized as follows. We first review the literature on product returns and return prediction and describe our research questions. We then describe the details of the experiments using customer reviews to predict the customer return reasons,



This work is licensed under a Creative Commons Attribution International 4.0 License.

CHIIR '24, March 10–14, 2024, Sheffield, United Kingdom
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0434-5/24/03
<https://doi.org/10.1145/3627508.3638326>

followed by the experimental results. Finally, we discuss the implications of our findings, conclude our work and highlight future directions.

2 RELATED WORK

Product returns have a high cost both to the seller and to the environment [1, 9, 37]. Much of the research on product returns aims to predict return volume and focuses on a single product category (c.f. [15, 16, 20, 23, 32]), under the premise that predicting return volume can help retailers identify problematic products and remove them from the shelves, thus reducing returns. We seek to understand customer return behaviors, to predict the reasons for the returns, with the aim of mitigating problems such as manufacturing quality and shipping reliability. We discuss related literature that provides insight into e-commerce with respect to returns, and insight into customer behaviors to provide background and context.

Cui et al. [5] collaborated with a large commercial auto accessories manufacturer in the U.S. sold through online retailers. In their work, each item was made to order, and the retailers (rather than the manufacturer) determined the return policies, although returns were handled by the manufacturer. They predicted return volume, and made several recommendations to the industry to better manage inventory. Their work differs from ours in that the items to be returned were made-to-order and customized, reducing the ability of the retailer or the manufacturer to re-sell the items. Further, their work covers returns that are the result of many different return policies. Moreover, customer satisfaction with the purchase of a made-to-order item may be driven by different expectations than off-the-shelf purchases. Finally the return behaviors for a single specialized family of products (automotive accessories) may not reflect the return behaviors in general across a wide range of product categories.

Asdecker [1] modeled the cost of returns, including the relationship between return rate and cost to the company, the effect of depreciation of the product due to damage either prior to shipping or as part of the shipping process, and the implications for business in terms of managing returns, for the purpose of recommending improvements to inventory management processes. In a follow-on study [2] they found that ordering multiple sizes or styles of clothing is a predictor of returns; higher priced items are more likely to be returned than lower priced items; customers who returned more items in the past are more likely to return items in the future; and products that were returned more often in the past are more likely to be returned in the future.

Dzyabura et al. [7] used the product image to predict its return rate. Their data consisted of 1.5 million online and offline transactions from a large EU fashion retailer with 39 retail stores in Germany. For the retailer in question, online transactions accounted for 22% of sales with an average return rate of 56%, compared to a 3% average return rate for items purchased offline. They reported that an estimated 1% reduction in return rate yields a 30% increase in profit. They used a set of baseline features that included the month of purchase, the product category and the price. They additionally found that incorporating image features in a Gradient-Boosted Regression Tree to predict return rate yielded an improvement of 10.4% over the baseline.

Urbanke et al. [39] presented a model for predicting customer returns using a novel technique for dimensionality reduction. Their experiments were performed on 1,149,262 products purchased from a major German online fashion retailer from July 2014 to November 2014, during which the return rate was 57.3%. Their feature set was comprised of product level features (e.g., brand, color), basket-level features (e.g., time of purchase) and customer-level features (e.g., historical return rate). The authors showed that the model was able to accurately predict cases wherein the likelihood of return is extremely high, before the purchase occurs.

Kedia et al. [19] used a deep neural network to predict return probabilities for fashion items sold by a large online fashion retail store in India. Their results showed an increase in precision from 65% to 74% using a fully connected deep neural network trained with aggregated embeddings of products (using Matrix Factorization [12], Bayesian Personalized Ranking [35]), sizing (using skip-grams [27] and Word2Vec [26]) along with multiple heavily engineered feature sets. As key features contributing to the model performance, the authors point to historical product return rate, cart size, payment type, customer historical return rate, sizing embeddings, customer purchase frequency, product price and product embeddings. Reviews, which are the focus of this paper, were not investigated.

While the research cited above often includes features of the customer transaction, there is a body of work focusing specifically on the customers. Some customers are more inclined to return products than others, and it can be helpful to understand customer behavior and preferences to better predict and manage returns. For example, while we generally would like to reduce product returns for environmental and financial reasons, Griffis et al. [11] showed that customers who return more also buy more, and that faster refunds increase future purchases by the customer.

Zhu et al. [42] built a graph of customers and products using data from online fashion retailers to predict the probability an individual customer (or group of similar customers) will return a specific product (or a set of similar products). Similarity between customers was based on historical purchase, return behavior and personal demographic attributes such as gender, age and income. They showed that leveraging the customer graph as well as product similarity improves accuracy. In our work, we do not focus on any specific category of products, but rather use that information in our feature set. In addition, since the focus of our work was to evaluate the impact of reviews on the prediction of customer return reasons, we did not include customer characteristics and historical behavior in our model which we leave for future work.

Pei and Paswan [29] focus on the different behaviors that lead to product returns, separating them into legitimate reasons (needs unfulfilled, buyer's remorse and external market reasons) versus "unethical reasons" (e.g., returning a Halloween costume after wearing it). The authors also divide the return drivers into internal (e.g., variety seeking or level of morality) versus external (e.g., product compatibility and return cost). Using crowd-sourcing, the authors collected data from 400 people to test multiple hypotheses on the relation between different personal/product traits and return behaviors. They found that certain personal characteristics such as impulsiveness lead to a higher likelihood of return for legitimate reasons. One of the main limitations of this study lies in the fact

that it relies on customer responses rather than real purchase and return data. Our work also groups the return reasons into several categories, but since our primary goal is to provide actionable explanations to the sellers, we focus less on the psychological drivers and more on the practical aspects of the returns (e.g., was the product defective versus did it fit the customer need).

Powers and Jack [31] described product and emotional dissonance as triggers for returns. Product dissonance is a state in which the customer believes that the purchase they have just made was either unnecessary or that the most appropriate product was not selected. Emotional dissonance is described as a state of psychological discomfort. The authors proposed that product dissonance can lead to emotional dissonance, and to alleviate the latter, customers may seek to return the purchased product. Using data collected from 308 customers of physical retailers (Walmart and Target) they showed a positive correlation between product dissonance and emotional dissonance. In addition, they looked into two direct reasons for returns: “expectation not met” and “found better product or price”. They found that emotional dissonance is positively correlated with both return reasons, and that the two reasons positively correlate with each other. Both reasons were also found to positively correlate with return frequency.

Our work differs from this work in several respects. There are key differences between online shopping and in-store shopping that make the return behaviors distinct. The lack of ability to touch and try the product prior to purchase may lead to many of the returns, especially for clothing which must be tried on. Our work does not consider the psychological state of the customer. Instead we take the customer reviews as the only representation of the customer rationale. Because many customers leave reviews, in this work we investigate how this rich source of information can be leveraged to improve customer return reason prediction.

3 USING REVIEWS TO PREDICT RETURN REASONS

Product reviews are explicit, often detailed, feedback about the customer experience with the product. A review typically includes a rating from one to five stars, and a textual description of the customer’s experience. Returns are also a form of customer feedback. When customers are not completely satisfied with their purchase, many retailers and e-commerce websites offer the option to return the product, to receive refunds or a product replacement, within a time window. When returning a product, customers are often prompted by the e-commerce platform to provide a reason for the return by selecting from a list of pre-defined reason codes (see table 2). In the work to follow we investigated the relationship between reviews and reason codes.

In the first part of our research, we explored the cases in which the customers who returned the product left a review about their experience. Our aim was to establish whether reviews from customers returning the product are important signals in predicting their return reasons. If no such connection exists, we would not expect reviews to be predictive of return reasons. We limited our first investigation to the subset of product returns where the returning customer also left a product review. Our first question was:

- RQ1: Can we predict the return codes from the returned product reviews?

When customers provide direct explicit feedback about their purchase experience, it is reasonable to expect that the review and the return reason would reflect the same source of dissatisfaction with the product. However, only a small subset of purchases are reviewed¹ and we cannot rely on the returning customer providing details and context in a product review. For the majority of returned products, where the customer did not provide a detailed explanation, reviews written by all other customers may help determine the pattern of return reasons and potential shortcomings of the product. In these cases, we investigated whether reviews of a product in aggregate, represented as text embeddings are effective features to predict the distribution of product return codes. Thus the second research question is:

- RQ2: Can we infer the return reason codes from reviews, even when the customer returning did not provide a review?

Often the pre-determined reason codes do not tell the whole story, because they lack detail and context. As shown in table 2, there could still be multiple, more specific reasons represented by each broader reason code (for example, a mis-advertised item and missing item accessories both belong to merchant-related reason code). We posit that one way to identify more detailed return reasons is through the reviews written by the customers. Therefore the third research question:

- RQ3: Can we infer detailed return reasons from reviews, to expand on the customer-indicated reason code and to summarize the reviews?

4 EXPERIMENTATION

In this section we describe experiments designed to address each of the research questions. For each research question, we detail the experimental methodology and the results, with additional analysis or discussion as needed.

Experimental Data. The data is English language data, from North American customers of Amazon.com.² We randomly sampled one day per month for a year from October 2019 to September 2020, and collected data for all products that were returned by customers on those specific days. We analyzed the span of one year to cover seasonal effects in e-commerce. We defined a return as a product sent back to the seller within thirty days of a confirmed purchase. Thirty days is a common window for free returns allowed by retailers and e-commerce companies. The data includes basic information about the original purchase, all valid customer reviews of the product, the return reason code, and textual reason description entered by the customer upon returning the product, if it exists (see Table 1).

4.1 Predicting Return Reason Codes from the Customer Review

We start by examining whether the review left by the customer who returned a product is predictive of their selected return reason code (**RQ1: Can we predict the return codes from the returned**

¹<https://www.usatoday.com/story/tech/news/2017/03/20/review-you-wrote-amazon-priceless/99332602/>, accessed January 2024

²<https://www.amazon.com>, accessed January 2024

Time period	2019/10 - 2020/09
Number of returns	90,962
Number of products	73,965
Number of product categories	34
Mean number of reviews	612
Mean length of reviews	35 words

Table 1: Descriptive statistics of the data analyzed in the study.

product reviews?). In this set of experiments, we limit the data to examples where the customer returning a product also left a review, and we did not include the text of the reviews from other customers.

4.1.1 Methods. As mentioned above, the data was sampled over one year, and consisted of products that were both returned and reviewed by the customer. For each of the returns, we collected basic information about the original purchase, as well as the review of the product written by the returning customer, if they wrote one. Each return record includes the reason code for the product return selected by the customer, shown in Table 2. The customer can also leave a free-text comment at the point of return, explaining why they are returning the item. The resulting data set has 90,962 total returns, for 73,965 products across 34 product categories (which covers all major product categories on Amazon.com). Among these returns; 63,828 (70.2%) are returned for product-related reasons; 17,682 (19.5%) for merchant related-reasons; 6,769 (7.4%) for customer-related reasons; 2,212 (2.4%) for carrier-related reasons and 471 (0.5%) for other reasons.

To avoid a situation where a given product in the test set might have many returns in a future time period as part of the training set (giving the model an oracle), we split the train and test data by date. We set a cut-off date such that the train set contains events (purchase, returns and reviews) that happened before the cut-off, while the test set contains events that happened after the cut-off. Examples in which purchases happened before the cut-off and the return or the review happened after the cut-off were discarded. No example in the test set has a corresponding product in the training set returned at a future date. This results in 79,773 training examples and 11,189 test examples. The five predicted classes were the five grouped customer return codes.

We first built a baseline model to predict return reason codes using features of the product such as its brand or category, information from the customer such as their membership status, and features of the seller, such as whether the seller is a vendor or a third-party seller, shown in Table 3. To investigate if customer reviews are predictive of the customer return reason codes, we created three additional models that use features extracted from customer reviews. We included the average star-rating and the number of reviews as features in all three models. To compare the importance of context in review text, we represented the review

Reason Code	Explanation
Item-related	Defective/Does not work properly
	Defect emerged after use
Merchant-related	Different from what was ordered
	Missing item in the set
Carrier-related	Damaged during shipping
	Item never arrived
Customer-related	No longer needed/wanted
	Better price available
Others	Other reasons provided by customers

Table 2: The set of pre-defined reasons customers are prompted to select when returning products, with representative examples.

Name	Type	Example Value
Product Features		
brand	text	<i>Nike</i>
product category1	categorical	<i>housewares</i>
product category2	categorical	<i>home goods</i>
all-time average rating	decimal	<i>4.2 stars</i>
recent average rating	decimal	<i>3.9 stars</i>
Customer Features		
active membership	boolean	<i>is an active member</i>
membership level	categorical	<i>frequent flyer</i>
Merchant and Offer Features		
seller type	categorical	<i>third party</i>
condition	categorical	<i>used</i>
resolution type	categorical	<i>refunded</i>
is return free	boolean	<i>free return</i>
is B2B Order	boolean	<i>not B2B</i>
purchase method	categorical	<i>1-click-purchase</i>
supplier type	categorical	<i>vendor</i>
Time Features		
days from purchase to return	decimal	<i>7 days</i>
days from delivery to return	decimal	<i>3 days</i>

Table 3: Baseline features for predicting product return codes. Product category1 and product category2 are two different types of product categorizations.

text using non-contextual word embeddings for one model, and contextual word embeddings for the other two.

For the non-contextual model, to capture the general opinion of each review, we used VADER (Valence Aware Dictionary and sEntiment Reasoner) [14], which is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. We obtained the positive, neutral, negative and compound score of each review text. VADER is also known to perform well on text that contains emojis, slang, and acronyms, which are common in product reviews. We then used the global vectors for word representation (GloVe)[30] to represent the review text. We used pre-trained word vectors based on Twitter data³, which map each word into a vector of 100 dimensions. Specifically, we mapped each term in the review into the GloVe vectors and took the average

³acquired from <https://nlp.stanford.edu/projects/glove/>, accessed January 2024

of the word embeddings. Both VADER and GloVe vectors were then appended to the features of the baseline model.

For the two contextual models, we used BERT (Bidirectional Encoder Representations from Transformers)[6] to extract information encoded in the review text. We create a pre-trained model and a fine-tuned model for the task. For both models, we tokenized and padded the text so that all review texts are the same size for pre-processing. For the pre-trained model, we encode the review text using uncased BERT-base⁴ and take the [CLS] in the last layer as the output. For the fine-tuned BERT model, we created an Adam optimizer with batch size of 16, learning rate of 2e-5, with 3 epochs. After training the model we use the fine-tuned model to extract last four hidden states and mean-pooled them as the final output. In both cases, the review text was transformed to a 1D vector of length 768. We appended the output vector from BERT to features of the baseline models. Since BERT is context-aware, we did not include features of sentiment analysis.

All models were trained using a multi-class classification with a LightGBM gradient-boosted decision tree [18]. The number of bins for feature buckets *max_bin* was set to 255, which was recommended by the official documentation and as shown in [41]. The number of iterations was set to 3500 with the learning rate of 0.005. Increasing the number of iterations did not increase the model performance. Early stopping was also enabled with number of rounds set to 1500.

4.1.2 Results. As shown in Table 4, the baseline model is able to predict the return reason codes with a micro average precision of 0.66 and a macro average precision of 0.552. In general the micro average is higher than the macro average across models and metrics, indicating the models are doing a better job at prediction for the larger classes.

Features derived from the customer reviews consistently improve performance over the baseline model, showing a statistically significant relative improvement in average precision of 2% (0.674 vs. 0.660) in the case of the GloVe embeddings, and 4% in the case of pre-trained BERT (0.687 vs. 0.660), and a large 23% gain in average precision (0.894 vs. 0.660) for fine-tuned BERT with similar improvements in recall and F1 score. When comparing the GloVe representation of review text to BERT, BERT performs better. To test if the improvements are statistically significant, we use McNemar's tests to compare the performances between the each experimental model with the baseline model. The tests show that the improvements are statistically significant after Bonferroni correction.

4.1.3 Feature Importance Analysis. To analyze the importance of customer reviews in determining the return reasons, we conducted an ablation test (see Table 5). We grouped the baseline features according to Table 3, together with customer reviews. We then trained multiple LightGBM models, while using one set of features only (first two columns), or all of the other features but excluding that single feature set (last two columns). When using only one set of features, we see that customer reviews is the most important feature as it outperforms all other feature sets in both precision and F1-score. This agrees with the results shown in Table 4, where the effect of including customer reviews is statistically significant.

When using all features but one, we again see that reviews have the biggest impact on performance (as seen by the reduction in precision and F1-score compared to All). Curiously, the rest of the features have similar importance with merchant features having the largest drop in precision and F1-score.

4.1.4 How customer reviews help predicting return reasons. After showing how features derived from customer reviews consistently improve performance over the baseline model, we now provide qualitative analysis on how the reviews assist models in predicting customer return reason codes over the baseline. By comparing predictions the baseline models got wrong but that the models trained with reviews got right, we summarized the following scenarios where customer reviews prove to be critical in determining the return reason codes (and see Table 6 for examples).

- *Baseline model makes predictions based on common return reasons of the product category.*
In situations where the baseline model lacks evidence on the return reasons, it predicts the most common return reason codes for the product category. As one example, cellphone accessories are commonly returned when customers do not like their purchases after receiving and seeing them in person. The customer reviews however show more nuanced reasons beyond the most common return reason.
- *Baseline model makes predictions based on merchant features.*
In other cases the baseline model makes predictions based on the merchant features. For example, for a return sold from a third party seller with mediocre reviews (3.5 stars), the baseline model predicts the return is due to merchant-related issues (e.g., false advertising). The specific customer review again is able to reveal additional information behind the return reasons, which shows defects with the product (see Table 6).
- *Baseline model defaults to the majority class.* When the baseline model is unable to determine the return reasons from the feature sets (e.g., item is sold from a reputed seller and is not a commonly returned item or item category), the model may default to predicting the majority class (i.e., the item is defective). The customer reviews are however able to provide the real reasons in certain cases, such as false advertising, product damaged due to shipment, or customers ordering the wrong size.

4.1.5 Error Analysis. It is important to understand the limitations of the model, including situations where customer reviews are not adequate in predicting the customer return reason codes. Here we summarize the types of error of the BERT model, where customer reviews do not contain helpful information regarding the return reasons.

- *Inaccurate return reason code*
In some cases, while the customer reviews explain the return reason (e.g., "They broke as soon as I put them up"), they selected an inaccurate return reason code for the returns (*Merchant-related issues: missing parts*). In other words, the ground truth of these returns does not match the label.

⁴<https://huggingface.co/bert-base-uncased>, accessed January 2024

Model	Avg. Precision	Avg. Recall	Avg. F1-Score
Base	0.660 (0.552)	0.713 (0.406)	0.644 (0.410)
Base+reviews (GloVe+VADER)	0.674* (0.579)	0.721* (0.437)	0.663*(0.450)
Base+ reviews (pre-trained BERT)	0.687* (0.576)	0.730* (0.446)	0.677* (0.460)
Base+ reviews (finetuned BERT)	0.894* (0.862)	0.897* (0.740)	0.894* (0.788)

Table 4: Review text significantly improves the prediction of customer return reasons, and are best represented with BERT. The best result is in bold, and statistically significant results are indicated with an asterisk. The macro average is in parentheses. The fine-tuned BERT result is statistically significant with respect to the GloVe result, as well as with respect to the baseline.

Feature	This feature only		All but one	
	Precision	F1	Precision	F1
All	0.894 (0.862)	0.894 (0.788)	0.894 (0.862)	0.894 (0.788)
Product features	0.569 (0.226)	0.585 (0.179)	0.894 (0.868)	0.893 (0.799)
Merchant features	0.600 (0.417)	0.589 (0.346)	0.887 (0.803)	0.886 (0.637)
Customer features	0.484 (0.139)	0.571 (0.164)	0.895 (0.862)	0.894 (0.784)
Time features	0.536 (0.256)	0.572 (0.167)	0.895 (0.871)	0.894 (0.801)
Customer reviews	0.888 (0.869)	0.886 (0.637)	0.662 (0.543)	0.651 (0.428)

Table 5: Ablation test analyzing the importance of product, merchant, customer, time and customer reviews. Both ‘This feature only’ and ‘All but one’ tests shows customer reviews are the most important feature (Macro precision & F1-score are shown in brackets).

Examples	Baseline pred.	Return reasons	Customer reviews
Prediction based on common return reasons of the product category.	Customer-related	Item-related	Doesn't fit my phone.... I have a Samsung s20+ the package sticker says s20+ but doesn't fit at all.
Prediction based on less reputed seller.	Merchant-related	Item-related	This car is garbage. The lights don't work. Pictures are misleading too. Is eight inches long.
Baseline model predicting majority class	Item-related	Merchant-related	Not as advertised in my opinion. It still has the same lines looking out the window as others. I felt false advertising took place there.
	Item-related	Carrier-related	I waited a week for this and my heart just could not believe I got a damage product and the screen was coming away from the trim and scratches all over it like it has been used , I'm so dissatisfied
	Item-related	Customer-related	Very good but I had to returned because I needed a bigger one.

Table 6: Examples of cases where customer reviews helped predicting the return reasons where the baseline model got wrong.

- *Users return for a reason not explained in reviews*

In some situations, the review was written before the customer had the intent to return a product, and the product was returned for a reason not mentioned in the review (e.g., a customer wrote a positive review before returning the product later after finding the product at a lower price from a different seller). This is confusing for the model especially if the review is positive, as the model interprets it as the customers liking the product, and then returning it due to customer-related reasons. In some other cases, information about the return is mentioned in the review but is insufficient for the model (e.g., *“Thank you for replacing the missing one. And I really like it.”*).

- *Non-English reviews*

We used the BERT base uncased model, which is pre-trained on English corpus only. A small portion of the customer

reviews in our dataset contain non-English tokens, which the model is unable to interpret. Using the BERT multilingual model or pre-training it on a larger corpus of non-English reviews could potentially help in avoiding these errors.

4.2 Inferring Return Reason Codes from Aggregated Reviews

While some customers provide direct feedback about their purchasing experience, that is not the case with all purchases. In many situations a customer may purchase a product, not be fully satisfied with the purchase, and subsequently return the product without leaving any explicit feedback. There are also situations where the customers only provide direct feedback some time after they return the product.

In these situations, e-commerce retailers do not have access to the review written by the customers returning the product (the focus of RQ1). To investigate and predict product return codes, e-commerce retailers have to rely on other indirect information. Reviews written about the returned product by other customers is one source of such information. Therefore the second research question we ask is **RQ2: Can we infer the return reason codes from reviews, even when the customer returning did not provide a review?**

As opposed to RQ1, we are not predicting a single instance of a return. Instead, we are investigating if we can use the *aggregated* reviews of a product to predict the distribution of return codes of a product. Identifying the trends in return reason codes for a product can help e-commerce retailers address specific issues and can provide feedback to the seller or manufacturer, or in extreme cases alert the customer to the issues of the product.

4.2.1 Methods. We expanded the dataset collected in the previous experiment. For the same 73,965 products (across 34 product categories), we collected all reviews associated with the products. For each product, we also collected all return records within the same one year period (October 2019 to September 2020). We aggregated these return records to obtain the distributions of return reasons codes of each product (described in Table 2).

To distill the reviews, we used BERT to extract word and sentence embedding vectors from the review text. For each product review, we do the following: (1) tokenize the review text, (2) use pre-trained BERT model to convert the tokenized text into layers of embeddings, (3) take the last layer of the output and apply mean pooling which yields a vector of length 768 of the *review embedding*. To maintain a reasonable computation time, we set an upper limit of the number of reviews extracted per product. The median number of reviews is 611. For products with more than 100 total reviews, we randomly sampled 100 and converted them to *review embeddings*.

For each product, we predicted the percentage distributions of five different return reason codes, using the *review embeddings* of the product. We modeled the problem as a multi-target regression with one regressor per targeted return reason code. The 73,965 products were split into 80% training set, 10% validation set and 10% test set. We chose XGBoost multi-output regression for the task.⁵

During our error analysis we noticed that many customers write positive reviews for products that they return (20,221 of the returns, which constitutes 22% of the data). This finding is somewhat surprising as we would expect customers returning a product to leave a negative review, because they are not satisfied with some aspect of it. One possible explanation is that these reviews were written prior to identifying issues with the product and returning it. This was not supported by the data as we found that many of the reviews were written *after* the return. To investigate the impact of the sentiment of the review on the reason code prediction, we trained four different models using different configurations of *review embeddings*.

- (1) **All reviews:** The mean *review embedding* of all reviews of the product.

Model	MSE	KL Divergence
Baseline	0.02276	0.28471
All reviews	0.01279	0.14829
Positive reviews only	0.01404	0.18094
Negative reviews only	0.01641	0.19959
Positive & Negative concatenated	0.01326	0.16076

Table 7: Results of using aggregated product reviews to predict the distribution of return reasons.

- (2) **Positive reviews only:** The mean *review embedding* of all 4 or 5 stars reviews of the product.
- (3) **Negative reviews only:** The mean *review embedding* of all 1 or 2 stars reviews of the product.
- (4) **Positive & Negative concatenated:** The mean *review embedding* of positive & negative reviews as separate embeddings, concatenated together.

We also set-up a baseline model, which predicted the average distribution of return reason codes aggregated over all products in the dataset.

4.2.2 Results. In order to compare the four experimental models we evaluated on two different metrics: (1) Mean Squared Error and (2) KL Divergence. Since we model this as a multi-target regression problem, Mean Squared Error computes the uniform average regression loss between the predicted percentage of each return reason code with the actual percentage (ground truth). As the target (return reason codes) is a probability distribution, KL Divergence measures the similarity between the predicted distribution and the actual distribution (ground truth).

Table 7 shows the results of the four models compared to the baseline. All four models outperform the baseline model considerably, showing aggregated reviews are helpful in predicting the distribution of return reason codes. Among all four experimental models, *All reviews* performed the best, which has almost half the mean squared error (0.01279 vs. 0.02276) and KL divergence compared to the baseline (0.14829 vs. 0.28471). *Positive reviews only* and *Negative reviews only* performed similarly, with mean squared errors of 0.01404 and 0.01641; and KL Divergence of 0.18094 and 0.19959 respectively. This result suggest that positive reviews contain information about the return reasons despite being positive. Lastly using *Positive & Negative only* reviews concatenated together as distinct sets of features achieved comparable results as using all reviews (which included 3-star reviews), but does not result in a better performance.

4.3 Extracting Detailed Return Reasons from Customer Reviews

In RQ1 and RQ2 we showed that customer reviews can be used as features to predict the return reason(s), both for individual purchases and for aggregated scenarios. Until now these reviews are only used to predict the fixed reason code selected by the customer. However, there could be multiple more nuanced reasons that lead a customer to return a product beyond the indicated reason code, and this information could be embedded in the reviews. Hence we now turn to investigate **RQ3: Can we infer detailed return reasons from reviews?**

⁵<https://xgboost.readthedocs.io/>, accessed January 2024

4.3.1 Methods. We propose that customer reviews refer to product attributes, such as short battery life, or complex installation, that provide insight into sources of customer dissatisfaction with a product. We extract descriptive phrases (keyphrases) from the customer reviews, and consolidate them into product aspects that describe attributes of the product. For example, a review might say that a digital thermometer has a “poor battery life”, another might say the “battery ran out fast”, another might say the “battery didn’t last”. These keyphrases all refer to the same product attribute (battery life), and together with the sentiment of the phrases, can be aggregated into a product aspect (poor battery life). To study if we can infer detailed return reasons from product aspects extracted from customer reviews, we first trained a model to extract product aspects from review text, and then conducted a between-subject experiment with human annotators to determine if the extracted aspects corresponded to the return reason codes.

Aspect extraction Inspired by [33], we extracted keyphrases that describe and summarize the customers experience described in the review text. We trained a span classification model adapted from [40], which is shown to be state-of-the-art in named entity recognition on multiple corpora. Instead of predicting a named entity, we adapted the model to recognize salient keyphrases. We collected a set of 750 review texts, and human-labeled all keyphrases in each review. We encoded the review text and obtained word embeddings using RoBERTa model [21]. We fed the word embeddings as input to a BiLSTM and finally to a bi-affine classifier. In the training set each span is labeled as either a keyphrase or a non-keyphrase. The bi-affine classifier ranked the spans and classified as keyphrases those whose score was above a threshold.

To standardize the aspects extracted by the model, we applied agglomerative hierarchical clustering on the extracted keyphrases, and labeled the resulting clusters as one of 60 pre-defined aspects of the customer experience (e.g., short battery life, hazardous product, missing advertised product features). We then trained a BERT-based keyphrase categorizer that classified each extracted keyphrase into one of the 60 classes of aspect. The result is an end-to-end system that extracts aspects from the review text, similar to [28].

Large Language Model-based aspect extraction The rapid development of large language models has been shown to match the performance of state-of-the-art fully-fine-tuned models (e.g., [3]). We developed a large language model-based aspect extraction by fine-tuning one of the state-of-the-art language models with customer review data. We started with 160 human annotated reviews in the format of instruction-input-response (see Table 8) as initial input. We then prompted LLaMA-13B model [38] to generate 8000 additional input-response pairs. We used Low-rank adaptation (LoRA) [13] to fine-tune the LLaMA-7B model with these input-response pairs, which was done using 4 * NVIDIA V100 GPU with a total batch size of 128, a learning rate of 3×10^{-4} , and a total of 3 epochs.

4.3.2 Results. To evaluate if customer reviews contain product aspects that could be extracted by the models, we conducted a between-subject experiment with human annotators to evaluate the models’ performance. We collected 612 reviews from customers who were unsatisfied with their purchase and returned the products for various reasons. Two sets of aspects were generated, one with

Instruction:

Identify aspects mentioned in the product review and then list their sentiments.

Input:

poor quality. I returned this set, very poor quality for the price.

Response:

The customer mentions Quality, Value for Money.

Quality is Negative

Value for Money is Negative

Table 8: Example response from the LLaMA model. In this case the customer indicated item-related issue as the return reason. The model identified the negative aspects to be quality and value for money.

statistical keyphrase extraction and another with an LLM-based model.

An in-house professional annotation team not connected to the research project labeled the data for the task. For each annotation, the annotators were provided with the review text written by the customer, the reason code selected by the customer upon returning the product, and the set of aspect(s) extracted by one of the models. The annotators were instructed to indicate (1) if the aspect is an explanation for the reason code selected by the customer and provides additional information on why the product was returned, and (2) if the aspect summarizes the customer experience mentioned in the review text.

The results are shown in table 9. The aspects extracted from the statistical keyphrase model attained 0.811 accuracy in explaining the return reason code selected by the customer. Breaking down the model accuracy in explaining different return codes, we see that it is 0.877 for item-related return codes, 0.675 for merchant-related return codes, 0.800 for carrier-related return codes and 0.204 for customer-related return codes.

The LLM-based aspect extraction performed very similarly—0.877 for item-related return codes, 0.735 for merchant-related return codes, 0.733 for carrier-related return codes and 0.333 for customer-related return codes, which is within a margin of a few percent in most categories. In both cases the extracted aspects were not adequate to explain customer-related return codes (0.204 and 0.333 respectively). Similar to the error analysis in the first study, the models failed on examples where the customer returned a purchase for reasons not described in the review. While most of the returns are due to the customer changing their mind, the review texts focus on attributes of the product unrelated to the return reason code.

In terms of the model’s ability to summarize review text, the statistical keyphrase model has an overall accuracy of 0.658. Here the performance is relatively even across different return codes: 0.658 for item-related returns, 0.547 for merchant-related returns, 0.533 for carrier-related returns and 0.750 for customer-related returns. Aspects extracted by the LLaMA model performed slightly worse in summarizing the reviews, but with similar declines in each category: 0.551 for item-related returns, 0.427 for merchant-related returns, 0.400 for carrier-related returns and 0.625 for customer-related returns.

It is important to note that the goal of this experiment is not to provide a definitive comparison between keyphrase extraction-based models and large language models. We chose a mainstream

language model for the task. A larger language model may perform the task better. Human annotations are also subjective in nature. In this case the annotation focused on identifying if the aspects extracted by each method explained the return reasons, rather than choosing which model the annotators preferred. The goal is to verify if we can infer more detail from customer reviews, to inform relevant stakeholders. We see that in more than 80% of examples, the aspects extracted from review text contained more detailed information about why customers returned a product than the reason codes indicated. This is a new insight for the e-commerce (and broader online) communities. Key information can be mined from customer reviews to identify reasons why users are (dis)satisfied with their experiences.

Task	Return Reasons	Accuracy	
		Keyphrase-based	LLM-based
Explains return reason	Overall	0.811	0.825
	Item-related	0.877	0.877
	Merchant-related	0.675	0.735
	Carrier-related	0.800	0.733
	Customer-related	0.208	0.333
Summarizes review text	Overall	0.658	0.527
	Item-related	0.684	0.551
	Merchant-related	0.547	0.427
	Carrier-related	0.533	0.400
	Customer-related	0.750	0.625

Table 9: Results of keyphrase extraction model and LLaMA model in identifying aspects that explain return reasons from reviews.

5 DISCUSSION

Prior research on e-commerce return behavior has primarily focused on predicting the return volume and return rate of products. While predicting the volume of product returns ahead of time is helpful for merchants and e-commerce retailers, this leaves a gap in understanding the origin of the issues, which limits the steps a company can take to prevent product returns. To the best of our knowledge, this is the first work to identifying the reasons for a product return.

Product reviews have been employed to predict retail sales [8]. Owing to the fact that reviews are explicit feedback provided by the customers, they capture (sometimes negative) experiences customers have with their purchase. We investigated how they could be employed to predict customer dissatisfaction, leading to a product return. We found that reviews are very effective at predicting the reasons customers return products. Signals used in prior work to predict return volume and rate were less effective than customer reviews at predicting return reason codes.

These findings extend the applications of behavior modeling from user reviews. For example, the field of recommender system research, there is a wealth of research using reviews in the form of ratings to provide recommendations, but using review text as a source of signals is more recent (e.g. [4, 25]). In our study we found that reviews helped most in outlier cases where the baseline models failed, for example by predicting the majority class or because the data was skewed by a specific feature. Our work shows that review

text as a source of features significantly improves over the baseline. We also show that reviews are helpful beyond predicting individual customer returns, but also to predict a reason code distribution for a given product or product line. This is critical to sellers and e-commerce retailers to address negative customer experiences. User reviews are public information on most e-commerce retailers and customers rely on reviews to make purchase decisions. Identifying customer complaints about products gives the retailer the opportunity to alert customers when they are unlikely to be satisfied with a purchase, by identifying products with a high return rate.

This work also brought a new perspective on the relationship between customer reviews and customer satisfaction. Previous research has shown that a customer’s product review should reflect their satisfaction [22]. One might expect that when it comes to predicting returns, only negative reviews would be helpful. We found, somewhat surprisingly, that customers sometimes leave positive reviews for products they return (about 25% of customers who return a product). Furthermore our results show that only using negative reviews to predict return reasons actually decreases performance when compared to using all reviews. In some cases the review rating may reflect the customer’s satisfaction with the products, however, customers may become dissatisfied with a purchase for reasons not related to the actual product, which are not captured in the review text. Therefore important information would be lost by only using negative reviews or review ratings to infer the return reason codes. Many reviews (often 3-star reviews) mention both positive and negative attributes of a product, which is valuable information in predicting return reason codes.

We showed product aspects extracted from review text provide explanations for the categorical reason codes. Human annotations showed that the extracted aspects provided explanations in over 80% of examples. Comparing a statistical keyphrase aspect extraction and LLM-based aspect extraction resulted in similar performance on aspect extraction, with the LLM-based model performing marginally better in explaining the reason codes. There are different advantages between the statistical keyphrase approach and the LLM-based approach for aspect extraction. Few-shot learning can be used with the large language model, requiring much less annotated data for training. On the other hand they are vulnerable to malicious injection attacks[17], the extracted aspects may contain inappropriate or unprofessional language, or the LLM might hallucinate. This is not a problem for statistical keyphrase extraction, as the extracted keyphrases are classified into a pre-defined taxonomy of aspects, eliminating the possibility of inappropriate language. These trade-offs should be considered when designing a system to inform sellers or customers of reasons for product returns.

5.1 Limitations

Due to the sensitive nature of the data in this paper that joins customer purchases, reviews and return information it cannot be released as public data. We acknowledge this is a limitation of the work. The main contribution of the paper is providing insight into customer return behavior, and demonstrating that reviews are helpful for predicting return reasons. As there is not an established baseline, the work addresses an important problem, provides insights into the underlying causes, and proposes a solution.

On the other hand, our experiments also highlighted a challenge of using reviews for the inference. In cases where there is a disagreement between the customer indicated return reason, and the review text, or where the review text does not contain useful information about the customer experience, the reviews are not helpful. From the error analysis, we observed that this happened when customers do not select an accurate reason code (and hence do not agree with the issues described in the review text) or when customers return products for reasons not described in the review (e.g., reviewing a product for its positive attributes, but returning it for other reasons). In these situations, using reviews alone is not sufficient to determine the customer's (dis)satisfaction. Future work is needed to investigate how to leverage other signals in these situations (such as analyzing the customer's purchase and return patterns).

Most often customers return products without writing reviews. We showed that the reviews of other customers, those who didn't return the product, including those who wrote positive reviews, is still helpful in predicting return reasons. In many online purchasing systems, customers are required to provide a brief description of the reasons at the point of return. We propose that our results also apply to the text written by the customers.

There are also other confounding factors that we could not capture in this experiment. The demographics of the customers (such as income) could influence one's decision to return a product. Due to privacy concerns, we did not control for any demographic variations between individuals in the experiment.

5.2 Broader Perspectives

The main data used in this work, customer reviews, are publicly accessible data. While records of customers returning products are not public information, we did not collect or use any personally identifiable information in the paper. This work demonstrated the potential for customer reviews to help merchants, sellers and e-commerce retailers identify the source of customer dissatisfaction. Ideally, this will help improve the customer experience, and reduce future returns, ultimately reducing the environmental impact of product returns.

6 CONCLUSION

In this paper, we showed that we can reliably predict return reason codes, and using review data significantly improves return reason prediction performance over a baseline model that does not include review text. We showed that using review text is not only helpful for predicting individual returns, but in aggregate review text improves the prediction of the distribution of return reason codes for a given product. Finally, we showed that we could extract fine-grained reasons from reviews that provide additional information not captured by reason codes.

The contributions of our work include using reviews to predict return reason codes *given* that a return has taken place, and identifying more detailed reasons why a product was returned. Combining our prediction models with a model that predicts whether a product is returned in the first place could prevent unnecessary returns. Detailed information about the sources of customer dissatisfaction could allow a better triage of product returns, and could be used to identify problems when they are indicated by early reviews, before

the product is returned and ends up in a landfill. Finally, understanding product aspects and sources of dissatisfaction, combined with a rich personalization system, could help prevent customers from purchasing otherwise good products that don't meet their specific needs.

REFERENCES

- [1] Björn Asdecker. [n. d.]. Returning mail-order goods: analyzing the relationship between the rate of returns and the associated costs. *Logistics Research* 8 ([n. d.]).
- [2] Björn Asdecker, David Karl, and Eric Sucky. 2017. Examining Drivers of Consumer Returns in E-Tailing with Real Shop Data. In *Proceedings of the 50th Hawaii International Conference on System Sciences*.
- [3] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023* (2023).
- [4] Li Chen, Guanliang Chen, and Feng Wang. 2015. Recommender systems based on user reviews: the state of the art. *User Modeling and User-Adapted Interaction* 25 (2015), 99–154.
- [5] Hailong Cui, Sampath Rajagopalan, and Amy R Ward. 2020. Predicting product return volume using machine learning methods. *European Journal of Operational Research* 281, 3 (2020), 612–627.
- [6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [7] Daria Dzyabura, Siham El Kihal, and Marat Ibragimov. 2018. Leveraging the power of images in predicting product return rates. *SSRN Electronic Journal* (2018), 1–33.
- [8] Kristopher Floyd, Ryan Freling, Saad Alhoqail, Hyun Young Cho, and Traci Freling. 2014. How online product reviews affect retail sales: A meta-analysis. *Journal of retailing* 90, 2 (2014), 217–232.
- [9] Rolf Gehring. 2017. *U.S. E-Commerce Sales*. <https://www.mytotalretail.com/article/the-cost-of-e-commerce-returns-and-why-you-should-care/>
- [10] John General. 2021. Here's what really happens to the items you return online. *CNN Business* (30 January 2021). <https://www.cnn.com/2021/01/30/business/online-shopping-returns-liquidators/index.html>
- [11] Stanley E. Griffis, Shashank Rao, Thomas J. Goldsby, and Tarikere T. Niranjana. 2012. The customer consequences of returns in online retailing: An empirical analysis. *Journal of Operations Management* 30, 4 (2012), 282–294. <https://doi.org/10.1016/j.jom.2012.02.002>
- [12] Patrik O Hoyer. 2004. Non-negative matrix factorization with sparseness constraints. *Journal of machine learning research* 5, 9 (2004).
- [13] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).
- [14] Clayton Hutto and Eric Gilbert. 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.
- [15] Tanuj Joshi, Animesh Mukherjee, and Girish Ippadi. 2018. One size does not fit all: Predicting product returns in e-commerce platforms. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 926–927.
- [16] Tanuj Joshi, Animesh Mukherjee, and Girish Ippadi. 2018. One Size Does Not Fit All: Predicting Product Returns in E-Commerce Platforms. In *IEEE/ACM 2018 International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018, Barcelona, Spain, August 28-31, 2018*, Ulrik Brandes, Chandan Reddy, and Andrea Tagarelli (Eds.). IEEE Computer Society, 926–927. <https://doi.org/10.1109/ASONAM.2018.8508486>
- [17] Daniel Kang, Xuechen Li, Ion Stoica, Carlos Guestrin, Matei Zaharia, and Tatsunori Hashimoto. 2023. Exploiting Programmatic Behavior of LLMs: Dual-Use Through Standard Security Attacks. *arXiv preprint arXiv:2302.05733* (2023).
- [18] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems*. 3146–3154.
- [19] Sajjan Kedia, Manjit Madan, and Sumit Borar. 2019. Early Bird Catches the Worm: Predicting Returns Even Before Purchase in Fashion E-commerce. *arXiv preprint arXiv:1906.12128* (2019).
- [20] Jianbo Li, Jingrui He, and Yada Zhu. 2018. E-tail product return prediction via hypergraph-based local graph cut. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 519–527.
- [21] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

- [22] Filipe R Lucini, Leandro M Tonetto, Flavio S Fogliatto, and Michel J Anzanello. 2020. Text mining approach to explore dimensions of airline customer satisfaction using online customer reviews. *Journal of Air Transport Management* 83 (2020), 101760.
- [23] Jungmok Ma and Harrison M. Kim. 2016. Predictive model selection for forecasting product returns. *Journal of Mechanical Design - Transactions of the ASME* 138, 5 (May 2016). <https://doi.org/10.1115/1.4033086>
- [24] MarketPlace Pulse. 2021. U.S. Statistics. <https://www.marketplacepulse.com/stats/us-ecommerce>. accessed September 2021.
- [25] Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*. 165–172.
- [26] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. [arXiv:1301.3781 \[cs.CL\]](https://arxiv.org/abs/1301.3781)
- [27] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Eds.), Vol. 26. Curran Associates, Inc., 3111–3119. <https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf>
- [28] Seoyeon Park and Cornelia Caragea. 2020. Scientific keyphrase identification and classification by pre-trained language models intermediate task transfer learning. In *Proceedings of the 28th International Conference on Computational Linguistics*. 5409–5419.
- [29] Zhi Pei and Audhesh Paswan. 2018. Regression Shrinkage and Selection via the Lasso. *Journal of Electronic Commerce Research* 19, 4 (2018), 301–319.
- [30] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.
- [31] T. Powers and E. Jack. 2015. Understanding the causes of retail product returns. *International Journal of Retail & Distribution Management* 43 (2015), 1182–1202.
- [32] Thomas L. Powers and Eric P. Jack. 2013. The Influence of Cognitive Dissonance on Retail Product Returns. *Psychology & Marketing* 30, 8 (2013), 724–735. <https://doi.org/10.1002/mar.20640> [arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/mar.20640](https://onlinelibrary.wiley.com/doi/pdf/10.1002/mar.20640)
- [33] Vahed Qazvinian, Dragomir Radev, and Arzucan Özgür. 2010. Citation summarization through keyphrase extraction. In *Proceedings of the 23rd international conference on computational linguistics (COLING 2010)*. 895–903.
- [34] Courtney Reagan. 2019. That sweater you don't like is a trillion-dollar problem for retailers. These companies want to fix it. *CNBC* (12 January 2019). <https://www.cnbc.com/2019/01/10/growing-online-sales-means-more-returns-and-trash-for-landfills.html>
- [35] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2012. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI2009)*.
- [36] Statista. 2020. Retail e-commerce sales in the United States from 1st quarter 2009 to 3rd quarter 2020. <https://www.statista.com/statistics/187443/quarterly-e-commerce-sales-in-the-us/>
- [37] Statista. 2020. Return deliveries - costs in U.S. 2017-2020. <https://www.statista.com/statistics/871365/reverse-logistics-cost-united-states/>
- [38] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
- [39] Patrick Urbanke, Johann Kranz, and Lutz M. Kolbe. 2015. Predicting Product Returns in E-Commerce: The Contribution of Mahalanobis Feature Extraction. In *Proceedings of the International Conference on Information Systems - Exploring the Information Frontier, ICIS 2015, Fort Worth, Texas, USA, December 13-16, 2015*, Traci A. Carte, Armin Heinzl, and Cathy Urquhart (Eds.). Association for Information Systems. <http://aisel.aisnet.org/icis2015/proceedings/DecisionAnalytics/2>
- [40] Juntao Yu, Bernd Bohnet, and Massimo Poesio. 2020. Named Entity Recognition as Dependency Parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 6470–6476. <https://doi.org/10.18653/v1/2020.acl-main.577>
- [41] Huan Zhang, Si Si, and Cho-Jui Hsieh. 2017. GPU-acceleration for Large-scale Tree Boosting. *arXiv preprint arXiv:1706.08359* (2017).
- [42] Yada Zhu, Jianbo Li, Jingrui He, Brian L. Quanz, and Ajay A Deshpande. 2018. A Local Algorithm for Product Return Prediction in E-Commerce.. In *Proceedings of IJCAI*. 3718–3724.
- [43] Natalia Zhukova. 2020. Dissecting Ecommerce Growth: The Key Traffic Drivers. *SEMrush Blog* (25 November 2020). <https://www.semrush.com/blog/dissecting-ecommerce-growth-the-key-traffic-drivers/>