INVESTIGATING CUSTOMER PERCEPTIONS OF SUSTAINABLE DESIGN FEATURES TO DRIVE PURCHASING DECISIONS FOR SUSTAINABLE PRODUCTS

A DISSERTATION

SUBMITTED TO THE DEPARTMENT OF MECHANICAL ENGINEERING

AND THE COMMITTEE ON GRADUATE STUDIES

OF STANFORD UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

NASREDDINE EL DEHAIBI

AUGUST 2021

This dissertation is online at: https://purl.stanford.edu/sk676zf3930

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Erin MacDonald, Primary Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Noah Goodman

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Conrad Tucker

Approved for the Stanford University Committee on Graduate Studies.

Stacey F. Bent, Vice Provost for Graduate Education

This signature page was generated electronically upon submission of this dissertation in electronic format. An original signed hard copy of the signature page is on file in University Archives.

ABSTRACT

Fierce competition on e-commerce platforms challenges designers to create products that appeal to customers. In particular, this occurs with sustainable products where an apparent demand for sustainable products fails to translate into real purchasing decisions. When creating sustainable products, designers tend to prioritize engineered sustainability features while neglecting perceived sustainability features. Engineered sustainable features are often hidden, for example energy usage or manufacturing methods of a product. Customers therefore rely on visual and descriptive features that align with what they perceive is sustainable, although these features may not contribute to real engineered sustainability. For a sustainable product to be successful, it therefore needs to meet both engineered requirements and perceived requirements.

To study the role of perceived sustainability on driving purchasing decisions, the work presented in this dissertation takes a multidisciplinary approach borrowing techniques from computer science, design, and marketing. First, a data-driven approach was used to extract features perceived as sustainable from online reviews using crowdsourced annotations and a natural language processing machine learning algorithm. Second, a novel collage design approach was developed to test the extracted features from online reviews in terms of how users identify the features as sustainable. Third, a shopping simulation was developed to validate how features perceived as sustainable can influence purchasing decisions of products when compared to dummy features.

Chapter 2 presents a method for designers to extract features perceived as sustainable from online reviews. Annotators identified phrases in product reviews that were relevant to one of the three sustainability pillars – social, environmental, and economic – and rated the positive and negative sentiment in the phrases. A logistic classifier was then used to extract salient features perceived as sustainable from the annotations. The method was tested on 1500 reviews of French presses and the extracted features were compared to a life cycle analysis output. The findings demonstrated that a gap exists between perceived and engineered sustainability, highlighting the importance of understanding features perceived as sustainable and the value of the proposed method.

Chapter 3 investigates validity metrics of highly qualitative text annotations. While external validity metrics, for example, precision, recall, and F1, are commonly used in computer science, internal validity metrics such as inter-rater reliability, are commonly used in design. The study tested four variations of Krippendorff's U-alpha using the annotations from Chapter 2 to compare internal validity metrics with external validity metrics. The results found that external validity metrics are more robust in the case of highly qualitative text annotations, providing insight for designers on best practices for assessing validity of highly qualitative annotations.

Chapter 4 presents a novel design method using a collage to test extracted features perceived as sustainable. The collage consisted of two axes,

sustainability, and likeability, where participants placed products and selected features from a dropdown menu according to how they perceived the products. Participants evaluated six French press on the three sustainability pillars – social, environmental, and economic – and on how much they like the products. In the dropdown menu, participants selected between features perceived as sustainable and features perceived as not sustainable. The results suggested that participants more often selected features perceived as sustainable for products they placed higher on the sustainability axis, validating that they identified those features as sustainable. Moreover, a significant but low correlation was measured between the placement of products on the sustainability and likeability axis, demonstrating the value of the collage tool to measure both dimensions separately. The findings confirm that the collage is an effective method for testing features perceived as sustainable with users.

Chapter 5 investigates the generalizability of the proposed methods by recreating them using electric scooters and baby glass bottles. Features perceived as sustainable were extracted for both products using the method outlined in Chapter 2. External validity metrics for electric scooters were comparable to that of the French presses, while they were much lower for baby glass bottles. It was identified that an imbalance of positive and negative reviews for baby glass bottles led to the weak performance in the machine learning models. The remainder of the study focused on electric scooters, testing the features using the collage approach outlined in Chapter 4. The findings were

comparable to those in Chapter 4 with French presses. The study demonstrated that the proposed methods generalize with limitations, mainly that the selected products should have a balanced set of positive and negative reviews.

Finally, Chapter 6 presents a shopping simulation to test how extracted features perceived as sustainable can influence purchasing decisions. A variety of features, images and descriptions were tested using a within-subject fractional factorial experiment. Participants navigated mockups of Amazon shopping pages and selected a product to purchase. They also rated products in terms of willingness to pay and sustainability. The results showed that participants were more likely to select to purchase products with features perceived as sustainable than dummy features. Moreover, participants were willing to pay more for products with perceived sustainability features and rated them as more sustainable, despite none of the features contributing to engineered sustainability. The findings validated that features perceived as sustainable can drive purchasing decisions and highlighted the importance of including both engineered and perceived features in sustainable design.

This dissertation demonstrates the value of features perceived as sustainable in sustainable design. While they may not contribute to engineered sustainability, they align with what the customer expects is sustainable. The results underscore the importance of designing for both perceived and engineered sustainability requirements to drive growth for sustainable products.

ACKNOWLEDGEMENTS

I would like to thank my principal advisor, Professor Erin MacDonald, for pushing me to do my best work possible. I am deeply grateful for the transformative experience she has provided me over the last five years. I have grown considerably since my first day as a PhD student.

I would like to thank Qatar National Research Fund (QNRF), for funding my work under the Qatar Research Leadership Program (QRLP). Specifically, I'd like to thank Dr. Ayman Bassil, Dr. Aisha Al-Obaidly, and Susi Estacio for their continued support and enabling me to stay focused on my graduate career.

I would like to thank my defense committee: Dr. Conrad Tucker, Dr. Noah Goodman, Dr. Larry Leifer, and Dr. Emma Brunskill for lending me their expertise and their time. Your input was instrumental in bringing this multidisciplinary work to life.

I would like to thank my IRIS Design lab mates: Dr. Ting Liao, Yiqing Ding, Aiyana Herrera, Disney Rattanakonkham, and Dr. Wan-Lin Hu for their constant source of inspiration, guidance, support, and entertainment over the last five years.

I was lucky enough to work not only with one program administrator but two administrators. I would like to thank Tammy Liaw and Renee Chao for their incredible support and positive energy during my PhD.

Last but not least, I would like to thank my family and friends for always being there for me. I could not have completed my PhD without your encouragement and support. Thank you for taking countless surveys for me. I am deeply grateful for all of you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. CHAPTER 1
INTRODUCTION

Sustainable products today are typically created using engineered sustainability requirements, for example, using a life cycle analysis (LCA) [1] (Fig. 1.1). An LCA quantifies the environmental impact of a product across the product life stages and can help designers guide their decisions. For example, a designer may select a product material to reduce the energy required for manufacturing, or they may select a recyclable material that reduces the environmental impact during the disposal phase.

Figure 1.1: Life cycle example

While it is crucial for sustainable products to meet engineered sustainability requirements, designers tend to neglect how customers perceive sustainability [2]. Customers perceive sustainability based on how they think about and identify sustainability, forming perceptions based on available information, thoughts, and prior experiences [3]. Perceived sustainability may not always align with engineered

sustainability. A sustainable product therefore needs to meet engineered and perceived sustainability requirements to be successful.

Designing for both engineered and perceived requirements is common practice in design. Looking at the automobile industry, designers typically style exterior and interior features based on perceived customer requirements while engineers control technical specifications [4]. Customers associate certain design features with car requirements despite the features not contributing meaningfully. For example, curvy and streamlined exteriors tend to be perceived as sporty and fast while boxlike exteriors tend to be perceived as stronger and more reliable [5] (Fig. 1.2)

Figure 1.2: Exterior comparison of sports car (left) versus off-road car (right)

Within sustainability, Reid et al. demonstrated that customers perceived car silhouettes with shorter vertical and longer horizonal dimensions as more fuel efficient, although there is a trade-off in-reality [6]. The authors quantified perceived environmental friendliness of silhouettes using design of experiments and survey design, and later incorporated the perceptions in a vehicle optimization model to assess trade-offs. These examples demonstrate that perceptions are real and highlight the importance of considering them in sustainable design.

This dissertation integrates perceived sustainability requirements into sustainable design to create sustainable products that are successful with users. Specifically, the dissertation investigates the following research question: how can designers identify perceived sustainable design features to create sustainable products that align with customer needs?

To answer this question, the dissertation takes a multidisciplinary approach. First, the dissertation develops a data-driven approach of identifying customer perceptions using online reviews and a natural language processing algorithm. This contrasts with traditional design approaches that typically include surveys, interviews, focus groups, and observations. While traditional approaches are successful at identifying design requirements, they are also open to biases. For example, questions in a survey might yield different answers based on how a question is phrased. Moreover, the approaches are difficult to scale for many customers due to time and cost constraints. The growing source of online customer preferences in the form of product reviews has created opportunities for data-driven approaches to scale design insights.

Second, this dissertation tests and validates extracted features perceived as sustainable using design and psychology approaches. While data-driven approaches are analytical and measurable, perceptions are subjective and emotion-based. The dissertation tackles this by developing a novel collage approach to bridge analytical data with subjective concepts. In doing so designers can identify actionable insights to creating sustainable products that align with customer needs.

Third, this dissertation investigates how features perceived as sustainable can influence and drive purchasing decisions using marketing techniques. It develops a realistic shopping simulation to assess real-world implications of extracted features perceived-as-sustainable in terms of purchase decisions and sales. This dissertation provides an added validation from a practical perspective for the proposed methods in this work.

The dissertation outline is as follows: Chapters 2 and 3 investigate methods to identify and extract product design features perceived as sustainable, Chapters 4 and 5 investigate methods to test the extracted features and validate them, and Chapter 6 investigates how features perceived as sustainable may influence and drive purchasing decisions.

2. CHAPTER 2
EXTRACTING CUSTOMER PERCEPTIONS OF PRODUCT SUSTAINABILITY FROM ONLINE REVIEWS

Abstract

In order for a sustainable product to be successful in the market, designers must create products that are not only sustainable in reality, but are also sustainable as perceived by the customer—and reality vs. perception of sustainability can be quite different. This paper details a design method to identify perceived sustainable features (PerSFs) by collecting online reviews, manually annotating them using crowd-sourced work, and processing the annotated review fragments with a Natural Language machine learning algorithm. We analyze all three pillars of sustainability—social, environmental, and economic—for positive and negative perceptions of product features of a French press coffee carafe. For social aspects, the results show that positive PerSFs are associated with intangible features, such as giving the product as a gift, while negative PerSFs are associated with tangible features perceived as unsafe, such as sharp corners. For environmental aspects, positive PerSFs are associated with reliable materials such as metal while negative PerSFs are associated with the use of plastic. For economic aspects, PerSFs mainly serve as a price constraint for designers to satisfy other customer perceptions. We also show that some crucial sustainability concerns related to environmental aspects, for example energy and water consumption, did not have a significant impact on customer sentiment, thus demonstrating the anticipated gap in sustainability perceptions and the realities of sustainable design, as noted in previous literature. From these results, online reviews can enable designers to extract PerSFs for

further design study and to create products that resonate with customers' sustainable values.

2.1 Introduction

Designing sustainable products that are successful in the market poses a continued challenge for designers. Despite 66% of global consumers saying they are willing to pay more for sustainable products [7], it is difficult to advertise and sell to this desire as sustainable features are often hidden and unnoticed, such as energy usage or manufacturing methods [8]. Customers are also skeptical of eco-labels due to misleading marketing strategies, or “greenwashing” [9]. Designers can communicate sustainability through subtle cues in the product features. For example, a previous study by She and MacDonald demonstrated that customers think about sustainability-related decision criteria as well as prioritize hidden sustainability features when exposed to visible product features termed “sustainability triggers” [8]. These findings were based on simulated real-world decision scenarios using realistic prototypes of toasters.

The growth of online shopping introduces a new challenge for communicating sustainability. Over the past two decades, more customers are moving towards online outlets with e-commerce sales making up 9.6% of total retail sales as of 2018, up from 4.2% in 2010 [10]. Roghanizad and Neufeld show that online customers tend to rely more on intuition than rational judgement when making purchasing decisions due to higher risk of buying a product before seeing it [11]. The authors use an online bookstore shopping simulation with website, decision, and risk manipulation to investigate changes in shopping behavior. Identifying customer perceptions of

sustainable features (PerSFs) can therefore help designers increase the appeal of sustainable products for online shoppers.

Traditional approaches of understanding customer perceptions include surveys, interviews, and focus groups. These approaches use stated preference in which customers report their preference or feedback in response to a prompt given by the designer. Stated preference for sustainability is prone to Social Desirability Bias: the propensity for people to do or say the socially-acceptable thing in hypothetical situations. For example, out of 60 participants that stated they are not willing to buy non-recycled paper towels in a survey, 52 of them reported buying a towel brand with 0% recycled paper the last time they went shopping [12]. This is a large problem for sustainable product assessment. Moreover, stated preference methods are time-intensive, prone to other biases, for example, priming, and may not capture all customer needs.

An alternative source for understanding customer perceptions is through online reviews; these have become feasible for designers to tap into with advancements in natural language processing (NLP). An example of two product reviews is shown in Fig. 2.1; each review provides different PerSFs of the product. For example, features such as the environment-friendly packaging and charity donations have positive sentiment (i.e., drive customer satisfaction) while the functionality of the filter has negative sentiment (i.e., drives customer dissatisfaction). The reviews can serve as a roadmap for designers on how to communicate sustainability from a product's features while also driving customer satisfaction.

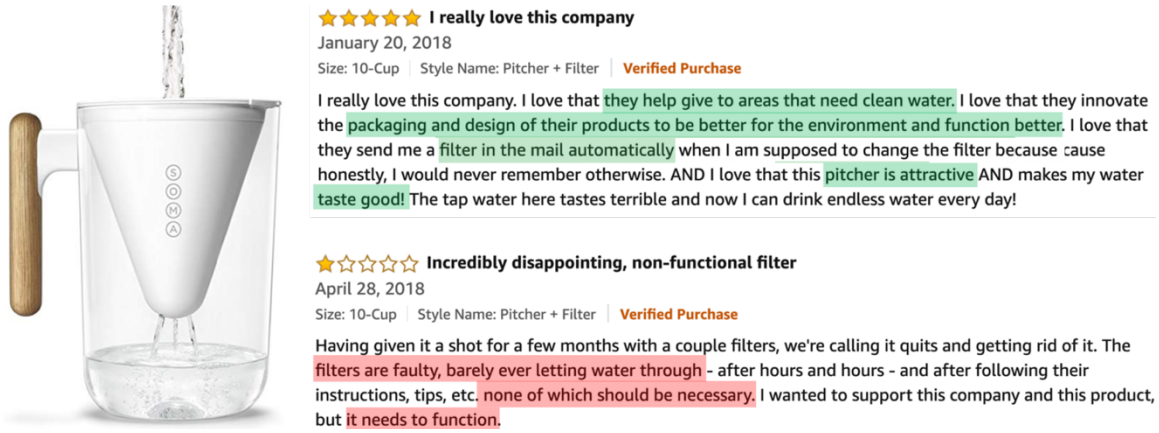

Figure 2.1: Example of a product review from Amazon. Green highlight indicates positive, red indicates negative sentiment

This study uses online reviews to identify PerSFs and to determine which of these features have positive and negative sentiment. Machine learning techniques are used to process large amounts of information. The goal is to help designers bridge the gap in perceptions and create products that satisfy both crucial sustainability design concerns and sustainability concerns as interpreted by the customer, which may in reality be superficial concerns. The rest of the paper is organized as follows: Section 2.2 presents a brief background on the use of online reviews in design, section 2.3 presents a literature review on NLP research, section 2.4 describes the method used to build a machine learning model, section 2.5 and 2.6 show the results and analysis, findings are discussed in section 2.7, and conclusions are made in section 2.8.

2.2 Background

In this section we present a background on the use of online reviews in design and the associated challenges for designers. A growing body of works is implementing techniques from NLP to address these challenges and is presented in section 2.3.

2.2.1 Online Reviews as a Resource for Designers

Online reviews are one of the largest and most accessible collections of crowdsourced customer perceptions. Ren et al. show that crowdsourcing can be used to capture perceptions of design features [13]. They recruited respondents from Amazon Mechanical Turk (MTurk) to assess perceived safety of car designs and used machine learning to capture important design features. The findings suggest that designers can use online reviews to understand perceptions that enable them to communicate cues to customers from product features.

Online reviews have been considered as both stated and revealed preferences, where revealed preferences rely on past-purchase information and not hypothetical. For example, Engström and Forsell consider online reviews as stated preference because they differentiate online reviewers from users who bought a product [14]. Netzer et al. consider online reviews as revealed preference that can be used as auxiliary input to stated preference data [15]. Online reviews have traits of both preferences as customers are not responding to a prompt but are still open to reframe their actions and choices in a more positive light (for a full discussion on the pitfalls of online reviews please refer to section 2.2.2).

Overall, it is likely that customers' assessments of sustainable features are more genuine than those in surveys and other traditional stated-preference approaches. For example, customer perceptions extracted from online reviews compare favorably with using elicitation-based methods such as surveys. Decker and Trusov demonstrate this using reviews for mobile phones [16]. Online reviews are also a source of product

innovation for designers. Qiao et al. examined frequency of App updates in the Google Play Store relative to the types of reviews written by users. They found that mildly negative and long and easy to read reviews increase the likelihood of an App update [17]. Reviews therefore provide more than just a word-of-mouth effect and provide valuable information for designers.

2.2.2 Challenges of Online Product Reviews for Designers

The availability of customer perceptions in online reviews presents both an opportunity and a challenge. While it offers a wealth of information for designers, it is difficult to synthesize useful information from it. Online reviews are unstructured, mostly written in free form, and the large quantities make them challenging to be processed by humans. The context that the reviews are written in is also unknown to the designer which can be problematic. For example, customers may have received a product for free in return for a review. It is also not possible to know if all customers paid the same price due to the fluctuating prices on websites such as Amazon, limiting the value of comments that mention words such as "affordable". In response to this challenge, industry experts have developed tools that measure the authenticity of reviews based on author history and other factors (refer to section 2.4.1 for more information).

Furthermore, customers perceive helpfulness of reviews differently from product designers. Liu et al. study the correlation between the customer helpfulness vote count of reviews from Amazon with review annotations on helpfulness to a designer [18]. The authors find a weak correlation between the two with a 35.3% mean

average error (MAE) and 29.5% root mean square error (RMSE). This suggests that there is a gap in perceptions for helpfulness of a review between customers and designers. The paper finds that longer reviews that discuss many product features are most helpful to a designer.

2.3 Developments in NLP Research

Research related to online reviews dates back to the 2000s in marketing research. Later works focused on extracting customer preferences from reviews using NLP techniques. These preferences might be explicit, where their meaning is not open to interpretation, or they may be implicit, where we would need to read between the lines to interpret them. The terms text mining, opinion mining, and sentiment analysis are often used interchangeably to refer to a group of NLP techniques. This section reviews NLP research within the field of design.

2.3.1 Extracting Explicit Customer Perceptions from Online Reviews

This section focuses on works that extract explicit customer perceptions from reviews. Rai was one of the first to identify customer preferences from online reviews with the goal of aiding designers [19]. He extracted key product features from reviews for a camcorder from epinions.com using a term-document matrix (TDM) and part-of-speech (POS) tagger. Stop-words were removed from the reviews and words were stemmed. A weighted metric considered the rate of occurrences of product features in the reviews to measure the importance of a feature. When compared to information from the website, importance levels were accurate up to the sixth ranked attribute.

Stone and Choi used Twitter as a source of customer preferences [20]. The authors used a 3-class Support Vector Machine (SVM) model for sentiment classification on 7000 Twitter messages related to smartphones, and a preference model to compare results of the SVM model with data from BestBuy (where product features are already categorized into pros and cons). Tweets were featurized using a bag-of-words model. Note that "featurizing" in this case refers to an NLP process for identifying measurable properties in text and is not related to features of a product. The results confirmed that customers share their opinions of products through Twitter and that designers can use this source to potentially inform design decisions.

Singh and Tucker used sentiment analysis to determine "must have" and "deal breaker" features for products [21]. "Must have" features are those that are popular while "deal breaker" features are those that are unpopular. Tweets related to the iPhone 5 were collected to test the method. Among the "must have" features were "light weight" and "WiFi" while the "deal breakers" included "battery", "screen", "speaker" among others. By identifying these features, designers can determine what to focus on in the next iteration of a product.

Singh and Tucker follow up on this work by investigating different machine learning models to classify reviews based on the content of the review using precision, recall, and F-scores to evaluate the model [22]. The authors manually annotated reviews to one of the following categories: function, form, behavior, service, and other content. Latent Dirichlet Allocation (LDA) was used for topic modeling to provide a benchmark for the annotators and to ensure that reviews annotated in "other" don't belong in the

other categories. LDA is a topic modeling approach which is commonly used for identifying topics in large amounts of text . The results showed that most one-star reviews were related to service, and that a product's star rating had the highest Pearson correlation with reviews related to form. By classifying reviews based on content, designers can identify which aspect of the product (function, form, behavior) needs improvement. Moreover, if a review is related to service, then it is more of a concern for the seller than the designer.

Tuarob and Tucker use social media networks to identify lead users [28]. Lead users are a group of product users that face needs ahead of the general market or population and can be a source of product innovation for designers. The authors compare product features that are discussed in social media networks with features from product specifications to identify which features do not currently exist in the market. The proposed method was tested using an iPhone case study and found the following top five latent features: waterproof, solar panel, hybrid, toothpick, and iHome. Using this method, designers can more efficiently identify lead users to help innovate new products.

2.3.2 Extracting Implicit Customer Perceptions from Online Reviews

Implicit perceptions include phrases such as, "I have to squint to read this on the screen," where explicitly this might be "the screen is too small". Tuarob and Tucker implemented a co-word network in the context of product design to capture implicit data in reviews [29]. To develop the co-occurrence network the authors first extracted explicit product features using a POS tagger. Sentiment extraction was performed using

SentiStrength [30]. A co-word network was then generated where the nodes are ranked to translate the implicit message into an explicit form. The authors used Twitter data, comprising of 390,000 Twitter messages about 27 smartphone products, to test the method. With this method designers can capture more of the available perceptions in online reviews.

Wang et al. proposed a Kansei text mining approach to capture customers' affective preferences in products from reviews [31]. Kansei engineering is a product development process that quantifies relationships between affective responses and design features [32]. Wang et al. first collected generic Kansei words using WordNet to expand on words from literature and then extracted product features using a POS tagger to identify common nouns and noun phrases. They filtered sentences from reviews so that only identified product features and Kansei words were included. The sentences were summarized based on word frequency to determine customer affective preferences. The authors used product reviews from Amazon to test this method.

The literature has yet to explore methods to identify complex topics in reviews similar to sustainability. Moreover, limited work exists on determining PerSFs from online reviews. This research aims to model PerSFs using machine learning techniques to determine which of these features are associated with positive and negative sentiment.

2.4 Method

The method described in this study combines research from identifying customer perceptions, rating design ideas, and natural language processing (Fig. 2.2). Methods for

identifying customer perceptions originate in marketing and behavioral science research and involve investigating human behavior in different purchasing contexts. Rating design ideas is a method that is commonly used in design research where concepts are evaluated through surveys or interviews either by “expert designers” or “novices”. Additional research insights on rating ideas were pulled from the field of information retrieval, specifically the idea of statements having a positive or negative emotion. Finally, we borrow algorithms from natural language processing, within the larger field of machine learning/cognitive science to codify written responses. There are many studies that use natural language processing to measure customer sentiment in online product reviews. This paper innovates on this research space by creating a new rating method to evaluate customer perceptions in product reviews with the goal of aiding designers to create more successful products. In contrast with methods proposed in literature, our approach enables designers to identify multifaceted and complex insights from online reviews beyond sentiment of product features. To the best of our knowledge this is the first rating method introduced in the design research space that combines approaches from identifying customer perceptions and natural language processing. We use sustainability as a case study to demonstrate its value. Specifically, we look towards evaluating customer perceptions of sustainability from reviews and using natural language processing to extract sustainable product insights for designers at scale.

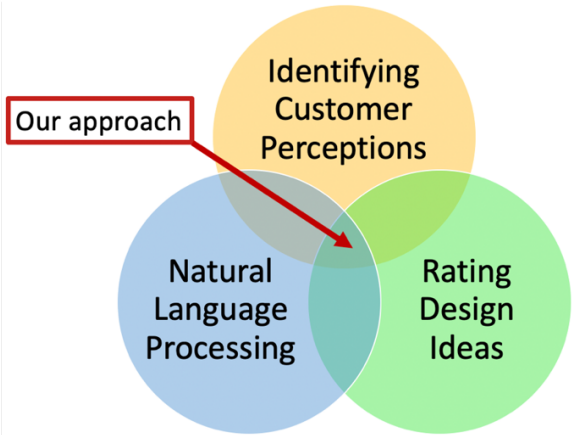

Figure 2.2: High-level overview of method topics

In this study we categorized sustainable product features into three aspects: social, environmental, and economic. The research proposition of this work is that product reviews related to these sustainability aspects contain semantic and syntactic characteristics that can be modeled. Sections 2.4.1 and 2.4.2 cover the method associated with the green and yellow regions of Fig. 2.2 while sections 2.4.3 and 2.4.4 explain the blue region of Fig. 2.2. A simplified chronological representation of the steps we took is shown in Fig. 2.3.

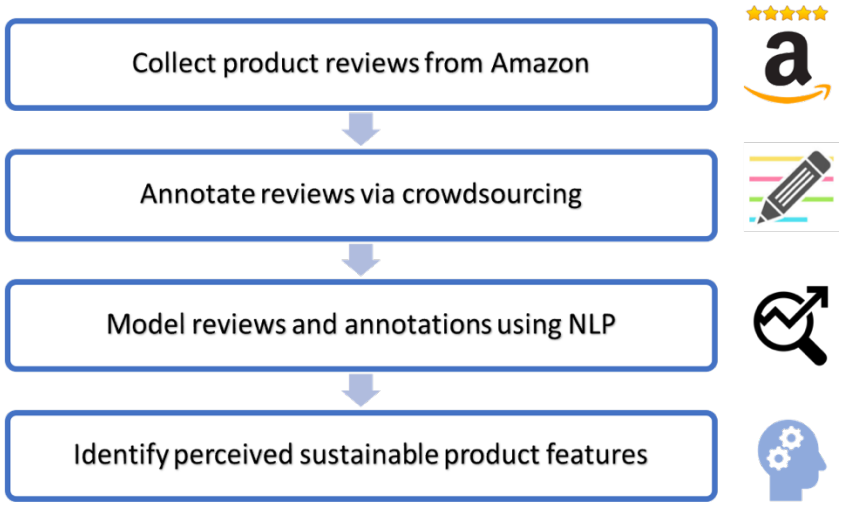

Figure 2.3: Chronological method flow

We used supervised learning techniques based on logistic classification to model the reviews. Each of the steps in Fig. 2.3 are explained below.

2.4.1 Collect Product Reviews from Amazon

We scraped a total of 1474 product reviews from Amazon for four French Press coffee makers. The intention was to select products that are ubiquitous and likely to have reviews that contain PerSFs. We used an online data analytics tool (fakespot.com) to estimate authenticity of reviews for a product and selected only products having an estimated 80% authentic reviews or higher. Very few products were rated as having 90% or more authentic reviews. The tool analyzes reviewer history patterns such as writing style, date correlation, frequency, and other factors to estimate authenticity. While up to 20% of the scraped reviews may have been fake, the number that contain sustainability aspects will be small due to fake reviews containing generic content. Therefore, any fake reviews are likely to be weeded out during the annotation process (see section 2.4.2). If any fake reviews are annotated, they are likely to be small in numbers and have a negligible effect on the models. We selected products that had similar features and were around the same price point as each other.

2.4.2 Annotate Reviews via Crowdsourcing

We recruited respondents from MTurk to annotate the collected product reviews via a Qualtrics survey, we refer to these respondents as “annotators” in this study (see section 2.4.2.3 for more information on annotators). The survey included training sessions, short quizzes, annotating reviews, checkpoints, and demographics

questions at the end. The annotations generated from the survey are used as data input to a machine learning model that identifies PerSFs from reviews (see section 2.4.3.2).

2.4.2.1 Survey Design

The survey consists of three versions to be customized for each sustainability aspect (social, environmental, economic). We distributed a total of 900 annotators evenly across each version, see Fig.2.4.

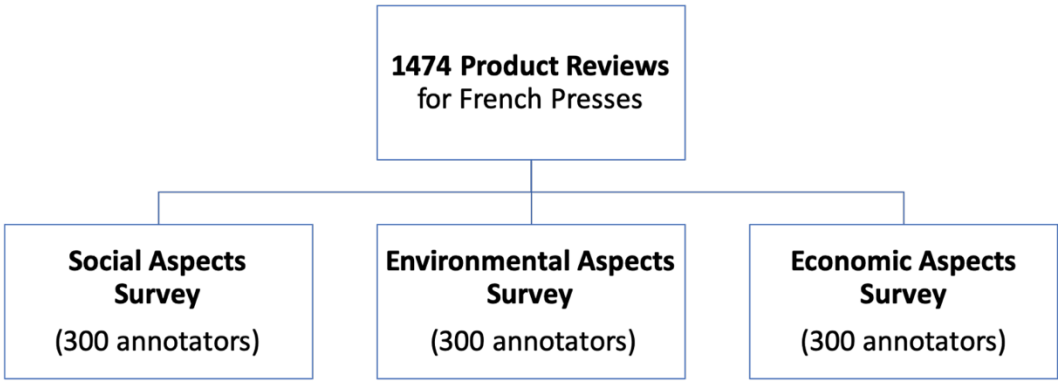

Figure 2.4: Three survey versions

In each version, annotators focus on one sustainability aspect to simplify the task as much as possible. We chose this approach after a pilot study showed that combining all three aspects in one survey confused the annotators. Each version has a customized training and testing portion. In the training portion, annotators are shown topics to look for in reviews (see Table 2.1) along with examples of annotated reviews[1]. In the testing portion, annotators choose phrases that are relevant to a sustainability aspect from

[1] http://erinmacd.stanford.edu/?attachment_id=334

example reviews. Annotators must pass this test to proceed and are given three attempts. Between the three versions, examples and test questions provided are based on similar topics to reduce potential biases.

Table 2.1: Topics to look for in reviews for each sustainability aspect

Social Aspects	**Environmental Aspects**	**Economic Aspects**
Health and safety	Material use	Product price
Family and culture	Energy and water consumption	Cost saving
Education	Product durability	Marketing
Community support	Air and water emissions	Profit and business growth
Human rights	Waste and recycling	Job creation

After passing the test, annotators are presented with 15 reviews and are asked to complete the steps shown in in Fig. 2.5.

Figure 2.5: General annotation process

Reviews are pulled from a server using weighted random sampling (see section 2.4.2.2) and displayed in the Qualtrics question. For each review, the associated product type and rating are shown. Annotators then use their best judgment to highlight phrases they perceive are “Relevant” to a sustainability aspect. Up to five relevant phrases can

be highlighted per review. Figure 2.6 shows a highlighting example for an environmental aspect.

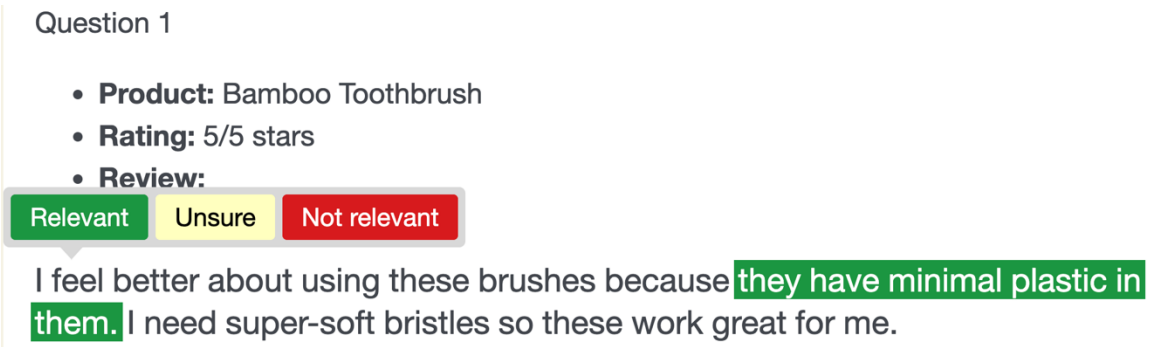

Figure 2.6: Example of highlighting a phrase

After highlighting a relevant phrase, annotators are asked to type in a product feature that is mentioned in the phrase and rate the positive and negative emotional strengths associated with the phrase (see Fig. 2.7).

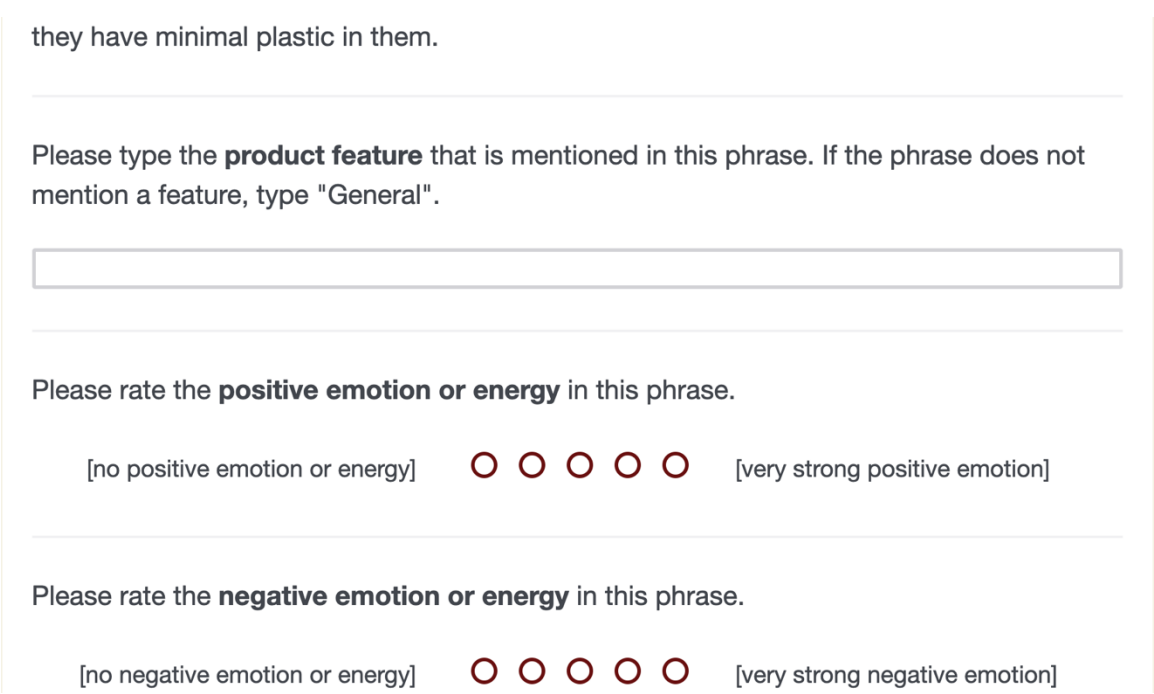

Figure 2.7: Example of questions about a highlighted phrase

If a phrase did not mention a specific product feature, annotators are asked to type “general”. The emotional strengths are rated on a 5-point Likert scale. We ran two pilot studies in which we presented annotators with reviews and asked them to evaluate the positive/negative emotions of phrases. We used two Likert questions as shown in Fig. 2.7, one for positive emotion or energy and one for negative emotion or energy. In the first pilot study we provided definitions of the terms “positive”, “negative”, and “emotion or energy” while in the second pilot study we did not. 16 annotators (eight per study) participated in total. We found that not providing definitions of the terms was less confusing (based on verbal feedback from participants) to the annotators and provided more usable responses (based on the number of similar ratings between participants, which doubled in the second study vs. the first). The overall emotional strength in a review phrase is then calculated as shown in Eq. 2.1 [27].

$$Emotional\ strength = Positive\ strength - Negative\ strength \tag{2.1}$$

If a review does not contain any relevant phrases, annotators are asked to highlight the entire review and label it as “Not relevant”. Annotators also have the option to select “Unsure” if they wish to opt out (Fig. 2.6). If either of these options are selected, annotators skip the questions in Fig. 2.7 and are presented with the next review. Note that only phrases highlighted as “Relevant” are used in the machine learning model. These questions required custom features in Qualtrics which we created using JavaScript. We decided to add the highlighting feature to gain more granular

annotations. An initial study showed that having a single annotation for a full review resulted in generic outputs from the machine learning model.

Despite the annotator training sessions in the surveys, the subjective nature of sustainability means it is unlikely to have consistent behavior among all annotators. We mitigate this by having three annotators for each review, therefore increasing the probability of an annotator catching a relevant phrase that was missed by another annotator. Moreover, if multiple annotators are highlighting the same phrase, then we can assume more confidence in the accuracy of the annotation.

2.4.2.2 Server Implementation

To control which reviews are annotated by whom, we hosted reviews on a server that Qualtrics requests reviews from via a JavaScript-built custom feature. The server uses a weighted random sampling method to select a review that it sends back to Qualtrics. The sampling method considers how many times a review has been previously selected and prioritizes reviews that have fewer annotations. Eq. 2.2 provides a mathematical representation of this:

$$S(r) = \left(1 - \frac{counter(r)}{3}\right) * random() \tag{2.2}$$

where r represents a review, counter(r) is the number of times a review has been selected, random() generates a random number between 0 and 1, and S(r) is the probability that a review is selected. If a review has not been selected before, it has a uniform probability of being selected, otherwise it is less likely to be selected until all other reviews have been selected the same number of times.

2.4.2.3 Annotators

A total of 900 annotators participated in the study (300 annotators per version of the survey) and each annotator spent an average time of 20 minutes to complete the survey for a compensation of $4. We used online instead of in-person annotators to efficiently annotate many reviews within reasonable time constraints. Moreover, we recruited respondents from MTurk instead of expert judges so that the demographics of the annotators match the demographics of online users in terms of age and education levels [33]. This is important such that the PerSFs identified by annotators can match as close as possible to those of the online reviewers.

To increase data reliability, we limited annotators to respondents in the United States that were on a desktop/laptop and had a minimum 97% approval rate. High approval rates are correlated most strongly with data quality [33]. Respondents based in the US also provide the highest response quality on average [34]. Moreover, after pilot testing, we found that the survey formatting on mobile devices was cumbersome and affected response quality, so we placed a laptop restriction. The surveys were launched on weekday mornings Pacific Standard Time to align with better responses from respondents during regular working hours [34]. The surveys were launched using Human Intelligence Tasks (HITs) on the MTurk platform.

We approved 871 responses out of the 900 total annotators. We used two criteria to approve responses: 1) time to completing the survey (t) is within 1 standard deviation (σ) of mean completion time (μ) or longer (i.e., $t \geqslant \mu - \sigma$) and 2) passing a checkpoint question. For participants that did not meet the first criteria, we approved

their response contingent on them answering the checkpoint question correctly. By relying on the checkpoint question as a final decider, we limit the chances of unfairly rejecting responses. For example, certain annotators may have received shorter reviews on average resulting in a shorter completion time.

2.4.3 Model Reviews and Annotations using NLP

We used logistic classification to analyze the acquired data and identify PerSFs. The model predicts if a given phrase has a positive or negative sentiment using (1) phrases that are highlighted as relevant (i.e., contain sustainability aspects) and (2) the typed-in product features by annotators. We first featurize the annotations and then build a logistic classifier model. Note that the term “featurize” here refers to an NLP process and is not related to product features. The steps involved are outlined below.

2.4.3.1 Featurize Annotations

We featurized the annotated review phrases, called "annotations" and associated words to identify measurable properties that can be stored in a matrix for input to a classifier model. The following data was featurized: the highlighted phrases, the typed-in product features, and the emotional strength scores.

We featurized the highlighted phrases using a standard bag-of-words (BOW) model as well as bigrams and trigrams [35]. Note that only phrases that were highlighted as “Relevant” (i.e., contained sustainable aspects) are used in the model. Text that was highlighted as “Not Relevant” or “unsure” was not used. In a BOW model, the rows consist of all the phrases while the columns consist of the vocabulary for the entire collection of phrases. The matrix then tabulates the number of times a certain

word occurs in each phrase. Table 2.2 shows an example. Bigrams and trigrams are modeled similarly except that we count the occurrences of two and three consecutive words respectively instead of the occurrences of individual words.

Table 2.2: Simple BOW model example

	Bamboo	**Handle**	**Stainless**	**Steel**
Bamboo handle	1	1	0	0
Stainless steel handle	0	1	1	1

The product features typed in by the annotators were featurized using LDA to identify a set of overarching product features. In this case, the topics are the product features, and the documents are the compiled texts typed in by the annotators. The number of topics is pre-defined and tuned for optimal results. The LDA model is presented mathematically in Eq. 2.3 [36]:

$$P(t_i|d) = \sum_{j=1}^{|Z|} P(t_i|z_i = j) * P(z_i = j|d) \tag{2.3}$$

where t_i represents a term from the total terms T, d represents a document from a collection of documents D, z_i is a topic to be identified, $|Z|$ is the total number of topics which is predefined, $P(t_i|z_i=j)$ is the probability of finding term t_i in topic j, and $P(z_i=j|d)$ is the probability of finding a term from topic j in document d. The LDA model is used to maximize the probability $P(z|d)$ which is the probability of a topic given the document. We hot-encoded the identified product features so that they are machine readable. For

example, if we identified “lid”, “handle”, and “glass” using LDA, we would input them to model as [1,0,0], [0,1,0], and [0,0,1] respectively for each phrase.

While the highlighted phrases and the typed in product features are inputs to the model, the emotional strength scores are outputs to the model. We used a two-class model which means that the output must be binary. In this case, the binary options are positive sentiment and negative sentiment. We initially ran a multi-class model but due to having less labeled data per class, the explanation power was too limited to draw conclusions. We therefore proceeded with a two-class model. A two-class model also allowed us to interpret the generated parameters and identify positive and negative PerSFs (see section 2.4.4). Implementing a multi-class model would have reduced the model performance without a clear benefit in terms of understanding what PerSFs drive customer satisfaction or dissatisfaction. We treated emotional strength scores above 0 as positive sentiment and scores at 0 or below as negative sentiment.

2.4.3.2 Build a Logistic Classifier

We implemented logistic classification in this study to predict if a phrase with sustainable aspects had positive or negative sentiment. We built three separate models to account for each sustainability aspect (social, environmental, and economic). The logistic function produces an S-shaped curve bounded between 0 and 1 such that the output is always meaningful for our purpose; negative sentiment has a value of 0 while positive sentiment has a value of 1. This model has proven to be a simple yet highly effective model in natural language understanding. The model for logistic classification is shown in Eq. 2.4:

$$p(Y = 1|X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}} \quad (2.4)$$

where X is a matrix with rows consisting of the phrases and columns consisting of the following information for each phrase: (1) BOW model, bigrams, trigrams and (2) product feature from LDA.

The term p(Y=1|X) is the probability that a given phrase belongs to class Y = 1 (i.e. that the phrase has positive sentiment) [37]. The βs are fitting parameters that are optimized using a maximum likelihood function shown in Eq. 2.5:

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} p(x_i)^{y_i} \left(1 - p(x_i)\right)^{1-y_i} \quad (2.5)$$

where $p(x_i)$ is the probability that review x_i belongs to class y_i. The intuition behind the maximum likelihood function is that betas are selected such that plugging them into Eq. 2.4 yields a number close to 1 for reviews that have positive sentiment and a number close to 0 for reviews that have negative sentiment.

We implemented logistic classification in Python using the Scikit package. The matrix generated from featurizing annotations consisted of several thousand columns that the logistic model used as information to predict customer sentiment. To avoid overfitting, the model uses penalty terms to shrink fitting parameters based on Ridge regularization. We used hyperparameter optimization with five-fold cross validation to optimize penalty terms.

2.4.4 Identify Features Perceived as Sustainable by Customers

After building and evaluating the logistic classification model, we examined beta parameters and p-values to identify the variables that have the largest influence on the model. The two-class model in this case lends itself for interpretability. For example, a positive parameter would indicate that a variable has a positive emotional score while a negative parameter would indicate that a variable has a negative emotional score. This interpretation would have been less clear with a multi-class model. Similarly, variables with a p-value of 0.05 or less were identified as statistically significant for having a relationship with the dependent variable (sentiment). As described in section 2.4.3.1, the explanation power from a multi-class model was too limited to draw conclusions due to the data structure.

P-values were measured using the Chi-squared test to measure dependence between variables. Note that we did not apply Bonferroni corrections as we used Ridge regularization with penalty parameters to address the high-dimensionality issue in the models. Through these indicators we can determine which PerSFs have positive or negative sentiment.

2.5 Pre-processing and Model Evaluation

Before featurizing the annotations, we first pre-processed the text data collected. This includes the phrases highlighted as relevant and the product features typed in by the annotators. Pre-processing text is done to minimize the amount of noise in the data by removing information that is unlikely to add value. The following pre-processing steps were taken: lowercasing, removing punctuation, removing stop-words

(words including “to”, “from”, “but”, “as”, etc.), and stemming (breaking down words to their root version).

We split 70% of the featurized annotations into a training set and the remaining 30% into a test set. The training data is used to train the model while the testing data is used to evaluate the predictive abilities of the model. By having two sets of data, we reduce the chances of overfitting as the model is evaluated on new data. We used five-fold cross validation on the training set. To measure how effective the model is, we used three metrics commonly used in NLP: precision, recall (also known as specificity), and F1 score. These are shown in Eqs. 2.6, 2.7, and 2.8, respectively.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{2.6}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{2.7}$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{2.8}$$

Precision and recall provide different perspectives about how well the model performs while F1 is a harmonic average of the two. Precision indicates how many of the predictions made by the model were correct while recall indicates how well the model was able to predict available information. For example, if there are 5 reviews with positive sentiment and the model predicts that only 2 of them are positive, it would have a 100% precision score while the recall would only be 40%.

The precision, recall, and F1 scores are shown in Tables 2.3-2.5 for social, environmental, and economic aspects, respectively. These scores evaluate how well the

model predicts positive and negative sentiment in phrases that contain sustainability aspects.

Table 2.3: Precision, recall, and F1 scores for social aspects

	Precision	Recall	F1
Positive Sentiment	0.85	0.87	0.86
Negative Sentiment	0.70	0.66	0.68

Table 2.4: Precision, recall, and F1 scores for environmental aspects

	Precision	Recall	F1
Positive Sentiment	0.83	0.86	0.85
Negative Sentiment	0.72	0.66	0.69

Table 2.5: Precision, recall, and F1 scores for economic aspects

	Precision	Recall	F1
Positive Sentiment	0.85	0.95	0.90
Negative Sentiment	0.72	0.42	0.53

The F1 scores for predicting positive sentiment are consistently high (between 0.85 to 0.90) while they are lower for predicting negative sentiment (between 0.53-0.69). This is likely because there were more annotated phrases related to sustainability that have positive sentiment compared to negative across the three sustainability aspects. Moreover, the results indicate that there may be false positives and false negatives, pointing to the need to validate the features extracted (refer to Chapter 4). Nonetheless, the scores suggest that we can have confidence in the value derived from the model and that designers can extract meaningful PerSFs from them, thus supporting our research proposition. The following section presents the PerSFs extracted in this study.

2.6 Analysis and Results

This section is split into two parts: in the first we analyze the annotation patterns in the survey, and in the second we report the outputs from the logistic classification models.

2.6.1 Analysis of Annotations

A total of 5189 phrases were highlighted as relevant to a sustainability aspect. Out of these phrases, 707 of them were highlighted by multiple annotators with an average difference in the positive ratings of 1.06 and in the negative ratings of 1.12 (evaluated on 5-point Likert scales) with standard deviations of 1.18 and 1.22 respectively across all three surveys. This suggests that the annotators had a consistent understanding of the questions on positive and negative energy.

Figure 2.8 shows the distribution of the number of relevant reviews annotated by annotators for each survey version. All three versions follow a similar skewed normal trend, averaging at about 6 relevant reviews per annotator followed by a spike at 15 reviews. The distributions are skewed towards 0 because overall there are less reviews that are relevant to sustainability aspects than reviews that are not relevant. The spike at 15 relevant reviews suggests that a subset of annotators was annotating more than needed, because this indicates that 15-20 annotators marked each review they saw as relevant, which is unlikely to be the case.

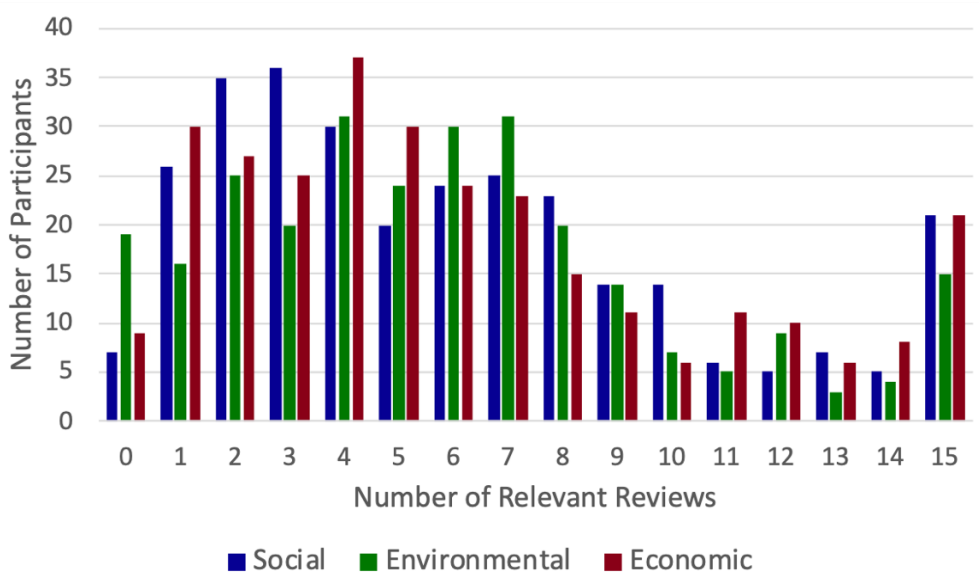

Figure 2.8: Number of relevant reviews per annotator

Figures 2.9-2.11 show distributions of the number of relevant phrases highlighted by annotators for social, environmental, and economic aspects respectively. These show more granular information than looking at the reviews overall. Most annotators highlighted between 0 and 20 relevant phrases with a handful of outliers in each survey. We manually checked the outliers and found that these annotators were still following guidelines for what is relevant to sustainability but chose to highlight shorter phrases with more frequency. The distributions in Figs. 2.8-2.11 do not follow a perfect normal curve which suggests that there is variability in the behavior of the annotators, as expected. This confirms the need for having multiple annotators per review to identify relevant aspects of sustainability in reviews.

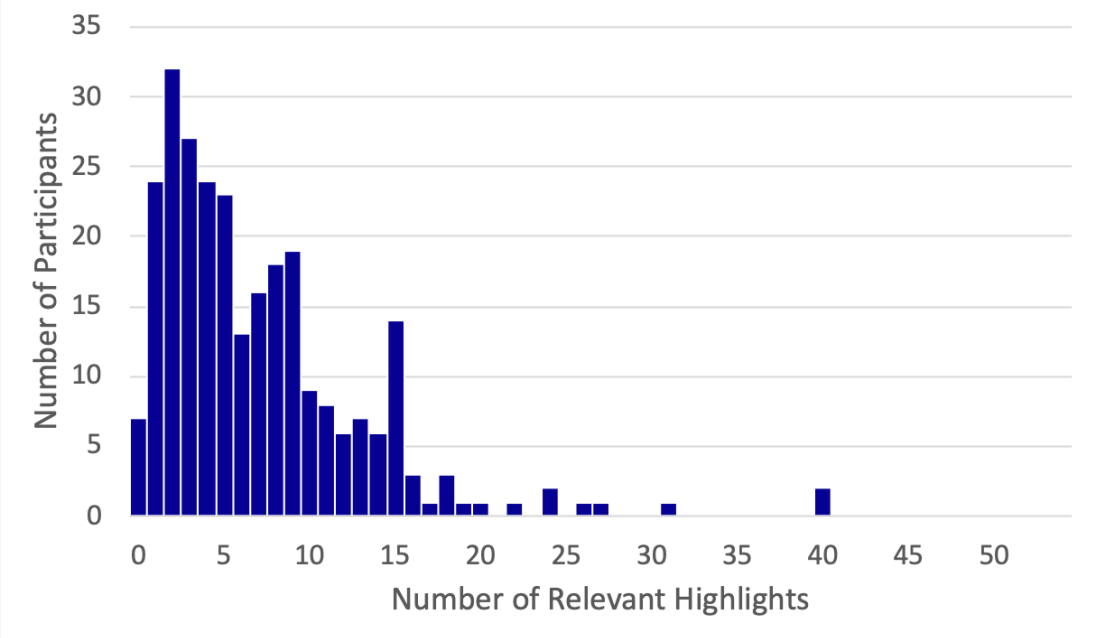

Figure 2.9: Number of highlights per annotator for social aspects

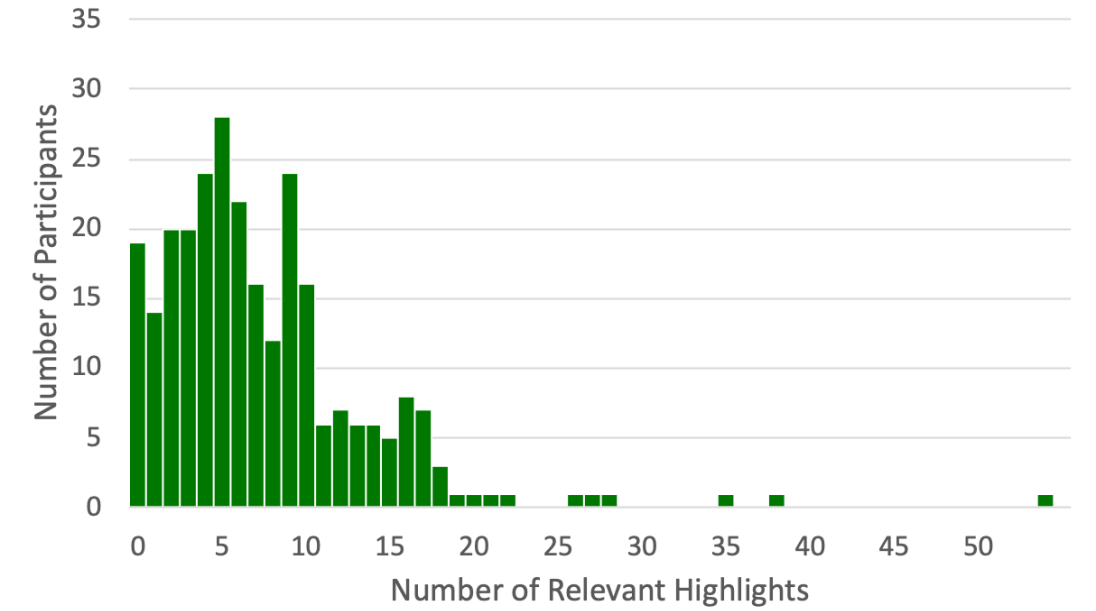

Figure 2.10: Number of highlights per annotator for environmental aspects

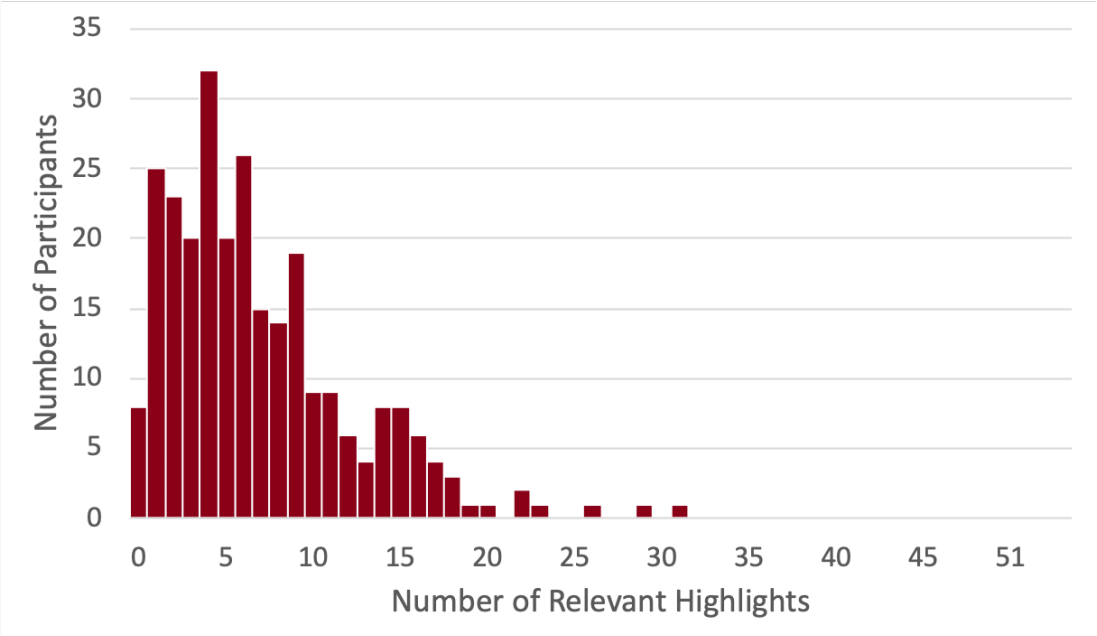

Figure 2.11: Number of highlights per annotator for economic aspects

2.6.2 Analysis of Classification Models

This section presents the product features obtained using topic modeling followed by the results from the logistic classification models.

2.6.2.1 Topic Modeling Output

Table 2.6 shows the extracted product features using the topic modeling approach outlined in section 2.4.3.1. The features are in order of highest occurrence in the annotated phrases. Note that we manually categorized the product features shown in Table 2.6 based on the cluster of words generated from the LDA model. For example, the cluster of words generated for topic 10 in the Economic model included “great”, “so good”, “love it”. We categorized these as “liking the product”.

Table 2.6: Product features generated from topic modeling

	Social	Environmental	Economic
1	General	General	General
2	French Press	French Press	Brand and marketing
3	Health and safety	Product durability	Cost saving
4	Liking the product	Plastic use	Durability
5	Glass carafe	Energy and water consumption	Quality
6	Easy use	Material use	Product design
7	Family and culture	Glass	Price
8	Coffee	Quality	Carafe
9	Plunger	Water waste	Glass
10	Filter	Metal	Liking the product
11	Size	Filter	Purchasing
12	Handle	Lid	-
13	Screen	Plunger	-
14	Lid	Size	-
15	Metal	-	-

The product features generated from the LDA model include a combination of general concepts presented from the training (such as “health and safety”, see Table 2.1) and specific product features generated by the annotators (such as “glass carafe”). Product features for social aspects revolve around safety, convenience, and generally liking the product. For environmental aspects the product features revolve around durability, material use, and energy and water consumption. Features for economic aspects revolve around price, quality, durability, and advertising. From Table 2.6 we can see that features tend to become more product-specific further down the list for social and environmental aspects. For the economic aspects, most of the product features are not product-specific. The product features from the LDA model provide an initial indication for a designer on where they should focus their efforts for a given sustainability aspect.

2.6.2.2 Logistic Classification Output

The largest and smallest logistic classification parameters from each of the sustainability aspect models are shown in Figs. 2.12-2.14. The larger (positive) parameters correspond to features that the model predicts have positive sentiment while the smaller (negative) parameters correspond to features that the model predicts have negative sentiment. Note that the features displayed in the figures have been stemmed as part of pre-processing the highlighted phrases such as “bought thi” in Fig. 2.12, which may originally have been “bought this” or “bought these” (see section 2.5 on stemming). Moreover, note that synonyms are present in the results (for example, “great valu” and “worth money” in Fig. 2.14). These synonyms may have been reduced by implementing vector representation of words to determine word similarities, however we avoided this to retain interpretability of the outputs of the model (i.e., to keep the outputs of the model as words instead of vectors).

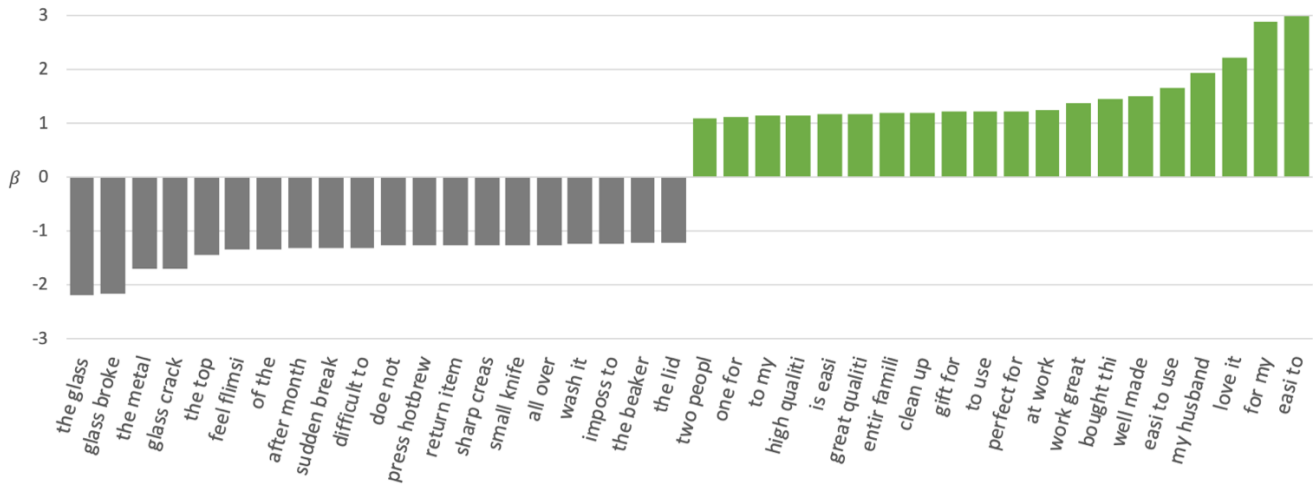

Figure 2.12: Top 20 most positive (green) and negative (grey) logistic classification parameters for social aspects

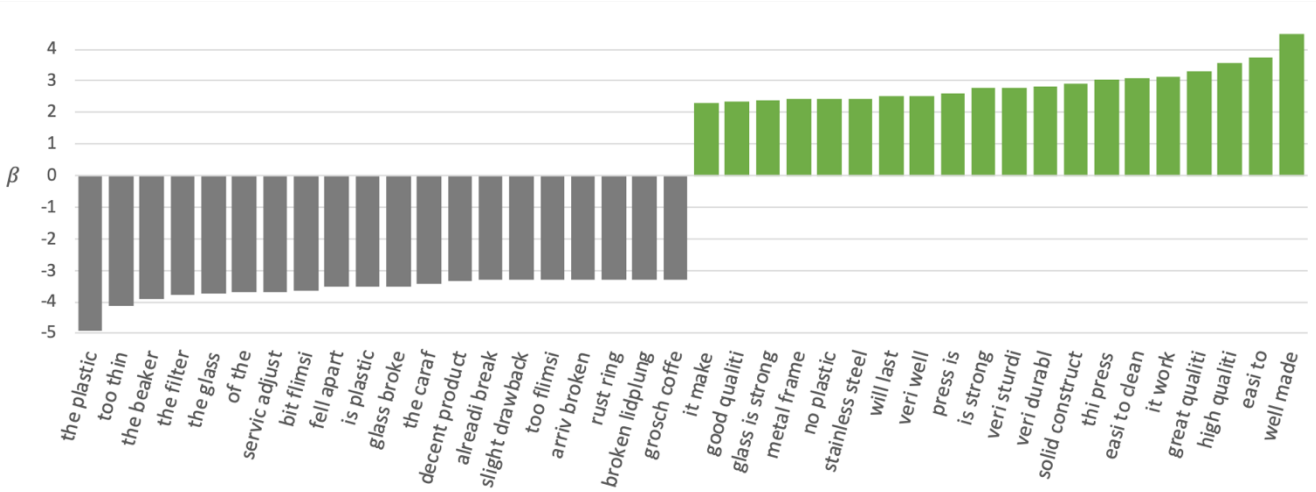

Figure 2.13: Top 20 most positive (green) and negative (grey) logistic classification parameters for environmental aspects

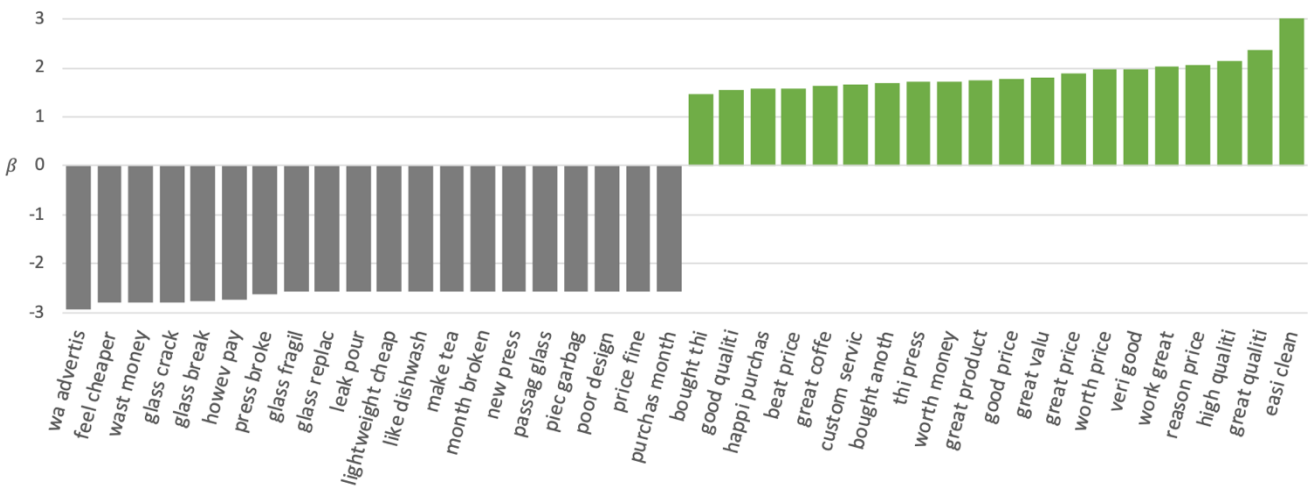

Figure 2.14: Top 20 most positive (green) and negative (grey) logistic classification parameters for economic aspects

Table 2.7 shows the features that are statistically significant at p=0.05 to customer sentiment for each sustainability aspect. For the most part these words can also be found from the parameters in Figs. 2.12-2.14, or are otherwise related, therefore indicating reliability in the results. For example, “after month” in the environmental column is related to the durability of the product over time. It is interesting to note that environmental aspects had the greatest number of significant

words, suggesting that customers have more consistent perceptions of product features related to environmental aspects than social aspects.

Table 2.7: Statistically significant words

Social	**Environmental**	**Economic**
Easy to	Easy to	Was advertised
Easy to clean	Well made	Feel cheaper
Glass broke	Easy to clean	Waste money
To clean	The glass	Glass crack
Glass crack	After month	Glass break
After month	Glass broke	Press broke
For my	To clean	-
The glass	Month of	-
Easy to use	Too thin	-
-	The plunger	-
-	High quality	-
-	Flimsy	-
-	Carafe	-
-	Plastic	-
-	Lid	-

2.7 DISCUSSION AND LIMITATIONS

The words, or PerSFs, identified by this study point to useful directions in sustainable design. To reiterate, it is important to design not only for "real" sustainability, but also to include features that customers perceive as sustainable. Whether actually beneficial for the planet or not, these perceived beneficial features create cognitive alignment and trust for customers when they evaluate sustainable products for purchase [38]. The PerSFs serve as useful inputs for product experiments with customers to create sustainable products with mass-market appeal. Here, we will review the PerSFs identified and point to some associated design directions.

It is important to note that several crucial sustainability concerns for environmental aspects were identified by the LDA model, which means that they were

mentioned in reviews, but they were not identified as critical to positive and negative sentiment. For example, energy and water consumption or recycling did not have a significant effect on the environmental aspects model in Fig. 2.13. To investigate this further, we performed a life cycle analysis (LCA) using Sustainable Minds[2] on a standard French Press and found that the biggest environmental impacts in terms of carbon footprint are associated with: (1) transportation of the product from the manufacturing site to the customer and (2) energy and water consumption while the product is being used (Fig. 2.15). The manufacturing of the French Press turns out to have a relatively low impact on the environment over an estimated 5-year lifespan of the product.

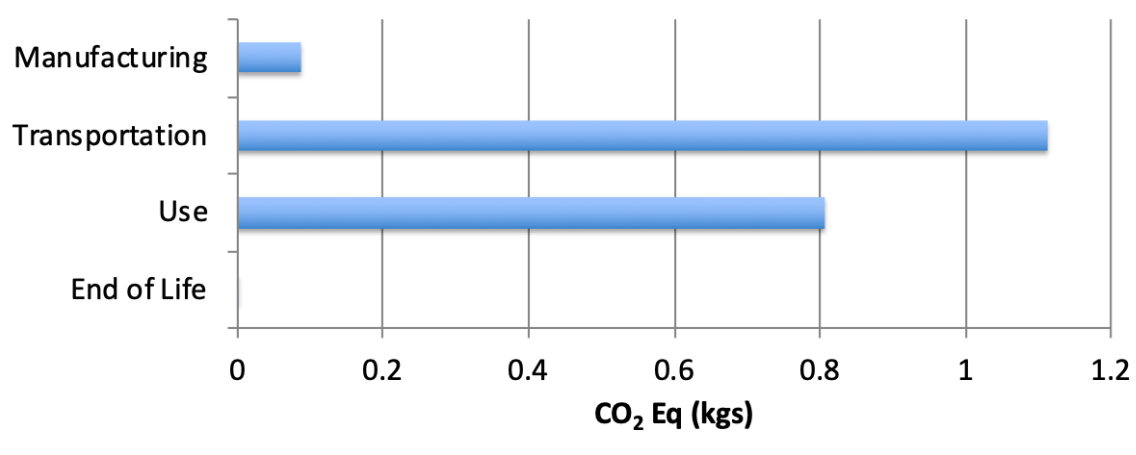

Figure 2.15: Life Cycle Analysis of French Press

A deeper look into the carbon footprint of materials in the French Press shows that choosing plastic at times can have a lower impact on the environment than stainless steel. Table 2.8 shows the carbon footprint for materials of two French Presses; the first is the original design from Fig. 2.15, the second replaces plastic parts with

[2] http://www.sustainableminds.com/

stainless steel. We can see that the design with more stainless steel and less plastic has a larger carbon footprint. This is contrary to the PerSFs identified for environmental aspects and supports existing literature that customer perceptions of pro-environmental designs can differ from actual pro-environmental designs [3,38]. This also demonstrates the gap in perceptions between designers and customers and the need for meeting both real sustainability concerns and concerns as interpreted by the customer.

Table 2.8: CO2 eq. emissions by material of product part

Original		**Modified**	
Material	CO_2 eq. kg/ function unit	**Material**	CO_2 eq. kg/ function unit
Glass, flat, uncoated	0.0943	Glass, flat, uncoated	0.0943
Stainless steel, austenitic	0.0263	Stainless steel, austenitic	0.0414
Polypropylene, PP	0.0149	Stainless steel, austenitic	0.0263
Stainless steel, austenitic	0.00993	Stainless steel, austenitic	0.0129
Stainless steel, austenitic	0.00993	Stainless steel, austenitic	0.00993
Stainless steel, austenitic	0.00993	Stainless steel, austenitic	0.00993
Polypropylene, PP	0.00465	Stainless steel, austenitic	0.00993
Total	**0.170**		**0.205**

Turning to the PerSFs that were found to have significant effect, we will now offer some recommendations for designers. For social aspects, in Fig. 2.12, the extracted PerSFs that are positive tend to relate to people, such as “for my”, “perfect for”, “entire family”. Other positive PerSFs include, quality, ease of use, and something that can be brought to work. These features relate more to the general experience of the product rather than a tangible feature. When looking at negative PerSFs for social sustainability, however, the features become more tangible such as the “glass crack”, “metal”, and. “sharp crease”. These features are potentially unsafe to the user. We also see features such as “beaker” and “lid” which can be tied to “glass crack” or “sharp crease”. Other negative PerSFs include difficulty of use such as “small knife” or “impossible to”.

For environmental aspects in Fig. 2.13, the extracted PerSFs are tangible features for both positive and negative parameters. Some of the features with positive sentiment include “glass is strong”, “no plastic”, “stainless steel”, as well as more general features such as “sturdy” or “high quality”. Looking at the features with negative sentiment, most of them are about the product breaking, which relates to durability. These include the carafe, filter, and glass breaking. The use of plastic also has negative sentiment. In some products, avoiding plastic in the external parts of the product may help it resonate with customers as sustainable.

For economic aspects in Fig. 2.14, the extracted PerSFs that have positive sentiment include that the product works overall and that it is worth the money. The features with negative sentiment include advertisements, feeling cheap, breaking, or if the product is not worth the money. These findings show that the number of tangible features for economic aspects is limited.

The results show potential in enabling designers to extract PerSFs from online reviews. For the case of French Presses, we recommend that designers communicate social aspects of sustainability by focusing on intangible features, such as making the product gift-friendly. Moreover, designers should ensure that the tangible features are perceived as safe for the user. For environmental aspects, designers can communicate this aspect by avoiding the use of plastic and instead using "reliable" materials such as metal. Designers can perform further semantic testing to identify metals and finishes that read as "reliable." Glass can also be perceived as positive if it does not impair durability of the product. For economic aspects, PerSFs revolve around how well the

product works in general and if it is a good price, but we could not identify tangible product features. Therefore, from a designer's perspective, the economic aspect of sustainability serves mainly as a price constraint for meeting the perceptions of social and environmental sustainability of a product. Using these insights, designers can communicate different aspects of sustainability to customers through the design of product features.

There are a few limitations in the study. The PerSFs extracted in this study were generated from reviews of French Presses and may not apply to other products. Testing the method on different products could help identify patterns in PerSFs between different products (refer to Chapter 5 for a generalizability study). Moreover, there are several words that overlap between sustainable aspects. For example, the glass breaking was common to all three aspects because it is interpreted as unsafe for social aspects, waste of material for environmental aspects, and low value for money for economic aspects. Therefore, it is important to keep in mind the context that the phrases were highlighted in. Moreover, using annotators to interpret the reviews instead of directly asking the authors of the reviews adds uncertainty. Finally, the lower scores for negative sentiment in Tables 2.3-2.5 suggest that there is noise in the features associated with negative sentiment, which could explain why terms such as "dishwasher" and "make tea" appear as negative features for economic aspects (Fig. 2.14). Annotating reviews that have a more balanced distribution between positive and negative sentiment could help address this. Moreover, we could achieve more

consistent annotation patterns in Figs. 2.8-2.11 by simplifying questions in the survey and emphasizing highlighting instructions.

2.8 CONCLUSION

This study shows that customer perceptions of sustainable features (PerSFs) can be extracted using annotations of online reviews and machine learning for the three pillars of sustainability: environmental, social, and economic aspects. We used reviews of French Presses to demonstrate the proposed method. Reviews were annotated by MTurk respondents using a Qualtrics survey and logistic classification was used to model the annotations. In terms of social aspects, positive PerSFs for a French Press include intangible features, for example, giving the product as a gift to a relative, while negative PerSFs include tangible features that could be unsafe to a user, for example, “glass cracking”. For environmental aspects, customers associate “stainless steel” and “strong glass” in French Presses with positive PerSFs and the use of plastic or product breaking with negative PerSFs. For economic aspects, customers relate product quality and value for money as relevant features. Importantly, features typically associated with "real" environmental benefit, such as energy use and water use, were identified, but under-represented as compared to "perceived" features that are not necessarily beneficial to the environment.

The logistic classification models performed well for predicting positive sentiment in phrases containing sustainable aspects, while there is room for improvement for predicting negative sentiment. Annotating reviews that have a more balanced distribution of positive and negative reviews would help address this.

Moreover, noise in the annotations can be reduced by simplifying some of the questions in the survey. For example, a single 5-point Likert scale would have been sufficient to measure the positive and negative sentiment in reviews. Emphasizing highlighting instructions could also have helped outlier behaviors shown in Fig. 2.8.

Moving forward, we want to investigate how the identified PerSFs can feed into design methods that validate the machine learning results and be used by designers in their products to communicate sustainability to customers. We also want to test if the identified PerSFs can affect customer purchasing behavior and increase demand for sustainable products.

3. Chapter 3

INVESTIGATING INTER-RATER RELIABILITY OF QUALITATIVE TEXT ANNOTATIONS IN MACHINE LEARNING DATASETS

Abstract

An important step when designers use machine learning models is annotating user generated content. In this study we investigate inter-rater reliability measures of qualitative annotations for supervised learning. We work with previously annotated product reviews from Amazon where phrases related to sustainability are highlighted. We measure inter-rater reliability of the annotations using four variations of Krippendorff's U-alpha. Based on the results we propose suggestions to designers on measuring reliability of qualitative annotations for machine learning datasets.

3.1 Introduction

The rapid growth in online user generated content and advancements in machine learning algorithms have enabled new approaches for designers to learn about customer needs. Designers traditionally conduct interviews, surveys, focus groups, or simply observe customers in a target context to better understand their needs [39]. Designers are also now able to identify important customer insights from sources such as product reviews or tweets using machine learning models and natural language processing techniques. These approaches are potentially faster, more cost-effective, and address some biases compared to traditional approaches, for example, surveys or interviews, but also introduce new challenges [28].

In supervised learning designers provide samples of input and output data to build a model. For example, Stone and Choi annotate tweets about phone products based on positive, negative, and neutral emotions in the tweets [40]. In this example the inputs are the tweets and the outputs are the annotations. A common challenge with this type of dataset is measuring the reliability of the annotations since the quality of the model depends on it. In machine learning research, a common way to evaluate the dataset is by looking at the evaluation metrics of the model such as precision, recall, F1 [35].

In traditional design research, designers request ratings from "expert" judges or "novices" (see for example, [41,42]). A common approach for assessing reliability of design ratings is inter-rater reliability (IRR) which provides an internal validity check. With IRR the responses from different annotators are compared using statistical analyses [43]. For example, Toh et al. use IRR to measure the agreement between two annotators rating a set of electric toothbrush design concepts [44]. By achieving high IRR measures, the authors can have confidence in their research approach and results collected. For an overview of IRR, please refer to section 3.2.

In this paper we explore IRR as a measure of reliability of qualitative annotations in machine learning datasets. The goal is to determine best design practice metrics to assess reliability of annotator data. We work with previously collected annotations of product reviews from Amazon where phrases relevant to sustainability are identified and highlighted [45]. This is a highly qualitative annotation task because sustainability is a complex and often subjective concept. We use IRR to measure the degree of

agreement among annotators and discuss the results given our context. The rest of the paper is organized as follows: in section 3.2 we provide an overview on IRR, in section 3.3 we describe our research approach, in section 3.4 we present the results, we discuss the results in section 3.5, and we conclude the paper in section 3.6.

3.2 Overview of Inter-Rater Reliability Measures

Inter-rater reliability (IRR) is a statistical measure of the degree of similarity between the results of different raters' (in this case, annotators), judging tasks. These tasks may involve sorting, judging on a scale, and parsing phrases. Based on these different tasks, raters may also be known as "coders", "judges", "observers", or "annotators". The IRR scale ranges from below 0 (denoting no agreement) up to 1 (denoting perfect agreement). The idea behind IRR is that the more agreement there is between the raters, the higher the confidence we can have in what the raters provide. Several measures exist for IRR, the simplest being a joint probability agreement which measures the percentage of observed agreement. Most IRR measures correct for expected agreement by chance and are considered to be a more robust estimate of the agreement, otherwise the agreement measure is overestimated [46]. We discuss some of the commonly used IRR measures below.

3.2.1 Cohen’s Kappa

Cohen's kappa measures the IRR between two raters for categorical items [47]. Recent examples of research using Cohen's kappa include studying reliability of coders assigning categories to audio files [48], or categorizing photos based on visitor behavior in public

parks [49]. Cohen's kappa is a function of p_o, the relative observed agreement, and p_e, the expected hypothetical agreement by chance, as shown below (Eq. 3.1).

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{3.1}$$

The observed agreement is calculated using Eq. 3.2,

$$p_0 = \frac{count\ of\ agreements}{count\ of\ agreements + count\ of\ disagreements} \tag{3.2}$$

and the expected agreement by chance is calculated using Eq. 3.3,

$$p_e = \frac{1}{N^2} \sum_k n_{k_1} n_{k_2} \tag{3.3}$$

where k is the number of categories, N is the number of items, n_{k_1} is the number of times rater 1 selected category k, and n_{k_2} is the number of times rater 2 selected category k. The advantage of Cohen's kappa is that it corrects for the expected agreement by chance and is therefore a robust estimate, but the disadvantage is that it is limited to only two raters for categorical items.

3.2.2 Fleiss' Kappa

Fleiss' kappa extends Cohen's kappa to work with any fixed number of raters for categorical items [50]. Recent examples of research using Fleiss' kappa include studying how well participants can read facial expressions [51], and evaluating psychometric perceptions of satisfaction questionnaires for patients and family members [52]. Fleiss' kappa takes the same form as Eq. 3.1, however, the observed and expected agreements are calculated differently as shown in Eqs. 3.4 and 3.5, respectively.

$$p_0 = \frac{1}{Nn(n-1)} \left(\sum_{i=1}^{N} \sum_{j=1}^{k} n_{ij}^2 - Nn \right) \tag{3.4}$$

where N is the number of raters, n is number of items, k is the number of categories, i is the index for each rater, and j is the index for each category.

$$p_e = \sum_{j=1}^{k} \left(\frac{1}{Nn} \sum_{i=1}^{N} n_{ij} \right)^2 \quad (3.5)$$

The advantage of Fleiss' kappa is that it is not limited to only two raters, but the disadvantage is that it can only be used to measure reliability of categorical items.

3.2.3 Krippendorff's U-alpha

Krippendorff's U-alpha measures the IRR for any number of raters and different types of data including both nominal and ordinal [53]. It is commonly used for measuring reliability of qualitative text analysis data such as highlighted text. Recent examples of works that have used Krippendorff's U-alpha include identifying arguments in portions of text [54], and identifying policy issues in news articles [55].

For a given text of length L, Krippendorff's U-alpha quantifies highlighted text by measuring where a highlight starts, *b*, and how long the highlight is, *l*, for each category (see Fig. 3.1).

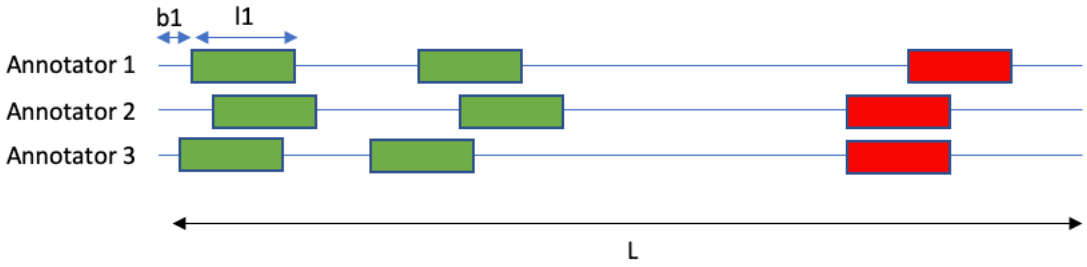

Figure 3.1: Quantifying text annotations for Krippendorff's U-alpha

In Fig. 3.1 there are three annotators for a text of length L and two categories (red and green). The highlights are quantified in terms of *b* and *l* and the differences

between annotators is measured to calculate agreement For a given category c, Krippendorff's U-alpha is calculated using the observed disagreement, D_{oc}, and expected disagreement, D_{ec}, as shown in Eq. 3.6.

$$\alpha_c = 1 - \frac{D_{oc}}{D_{ec}} \tag{3.6}$$

For a given category c, D_{oc} is calculated as shown in Eq. 3.7,

$$D_{oc} = \frac{\sum_{i=1}^{m} \sum_g \sum_{j=1|j\neq i}^{m} \sum_h \delta_{cigjh}^2}{m(m-1)L^2} \tag{3.7}$$

δ_{cigjh}^2 is the squared difference between annotation g and annotation h corresponding to any two observers i and j, respectively, m is the number of raters, and L is the length of the given data. Length L and difference δ_{cigjh} is typically measured using letter counts as a unit of length. The difference δ_{cigjh} is calculated in Eq. 3.8 as follows:

$$\delta_{cigjh} =$$

$$\begin{Bmatrix} \left(b_{cig} - b_{cjh}\right)^2 + \left(b_{cig} + l_{cig} - b_{cjh} - l_{cjh}\right)^2 \; iff \; v_{cig} = v_{cjh} = 1 \; and - l_{cig} < b_{cig} - b_{cjh} < l_{cjh} \\ l_{cig}^2 \; iff \; v_{cig} = 1, v_{cjh} = 0 \; and \; l_{cjh} - l_{cig} \geq b_{cig} - b_{cjh} \geq 0 \\ l_{cjh}^2 \; iff \; v_{cig} = 0, v_{cjh} = 1 \; and \; l_{cjh} - l_{cig} \leq b_{cig} - b_{cjh} \leq 0 \\ 0 \; otherwise \end{Bmatrix} \tag{3.8}$$

where b denotes the beginning of a highlight for a given rater, l is the length of a given highlight, and v is binary denoting if a section is a highlight $(v = 1)$ or not a highlight $(v = 0)$. For text data, b and l are defined in terms of letter counts. The expected disagreement is then defined as shown in Eq. 3.9,

$$D_{ec} = \frac{\frac{2}{L}\sum_{i=1}^{m} \sum_g v_{cig} \left[\frac{N_c - 1}{3}\left(2l_{cig}^3 - 3l_{cig}^2 + l_{cig}\right) + l_{cig}^2 \sum_{j=1}^{m} \sum_h (1 - v_{cjh})(l_{cjh} - l_{cig} + 1) \; iff \; l_{cjh} \geq l_{cig}\right]}{mL(mL-1) - \sum_{i=1}^{m} \sum_g v_{cig} l_{cig}(l_{cig} - 1)} \tag{3.9}$$

where L is the total length of the text (letter counts). To calculate Krippendorff's U-alpha for multiple categories, the observed and expected disagreements for each category are summed as shown in Eq. 3.10.

$$\alpha = 1 - \frac{\sum_c D_{oc}}{\sum_c D_{ec}} \tag{3.10}$$

Based on the above explanations of different IRR metrics, we choose to focus on Krippendorff's U-alpha because it is the most generalizable approach for different types of data and enables us to calculate reliability of qualitative highlighted text. Krippendorff's U-alpha was created to measure agreement between raters for qualitative text, but it is unclear if a high degree of agreement is desirable in the context of machine learning datasets. In this paper we investigate the implications of Krippendorff's U-alpha measure when annotating text for machine learning datasets.

3.3 Research Approach

In this study we use annotations of product reviews of French Press coffee makers collected in [45]. Annotators were recruited from Amazon Mechanical Turk and participated in a Qualtrics survey where they were briefly trained on either social, environmental, and economic sustainability. There were three versions of the survey to account for each sustainability aspect. After completing the training, annotators were asked to highlight phrases related to a sustainability aspect in product reviews and to rate the positive and negative emotions in the phrases they highlighted. The authors adhered to common practice for high quality responses from Amazon Mechanical Turk as outlined by Paolacci and Chandler [33] and Goodman and Paolacci [34]. Bonus compensation was also offered for annotators to incentivize high quality work. Some of

the reviews were annotated by two or more participants: social (449 reviews), environmental (404 reviews), and economic (436 reviews), for a total of 1289 reviews that we use in this IRR study. Note that for these 1289 reviews, there are two to three annotations per review.

We calculate IRR measures on the annotated reviews using Krippendorff's U-alpha as defined in Eqs. 3.7 – 3.10. The annotations are split into two categories, positive emotion and negative emotion. As a baseline we use letter counts in Eq. 3.8 to measure differences between annotations and calculate Krippendorff's U-alpha. In addition to the baseline, we implement some variations so that we may tune how sensitive the IRR measure is to differences between annotators. The baseline measure and variations implemented are explained below.

Baseline (Letter counts): We calculate Krippendorff's U-alpha using letter counts as unit of difference measure between annotations (the smallest unit of length). The difference between annotator highlights is counted by letters.

Letter counts with natural language processing (NLP): Similar to the baseline, we use letter counts to measure length and differences between annotations, however we first pre-process the reviews with natural language processing. This includes lowercasing, removing white spaces, numbers, punctuation, and stop words, lemmatizing, and stemming the words in the reviews. The intuition of using NLP is that it can remove potential noise in the annotations.

Word counts: We calculate Krippendorff's U-alpha using word count to measure length and differences. Although Krippendorff's alpha adjusts based on length of the

overall text, the intuition of using word counts is that it may make the overall calculation less sensitive to distances between annotator highlights.

Word counts with NLP: We calculate Krippendorff's U-alpha using word count to measure length and differences, and pre-process the reviews with natural language processing. We implement the same NLP steps as in "Letter counts with NLP" but using word counts instead. We intuit that NLP may have a bigger impact when looking at word counts and better allow us to tune the outputs as needed.

For this study we calculate Krippendorff's U-alpha for three sets of 400 to 450 reviews (a set for each sustainability aspect) using the above four measures. We developed a Python code to calculate Krippendorff's U-alpha measures on the annotations, available on GitHub[3]. We compare the different IRR variations to determine if we can tune the output to provide insight on the reliability of the annotator data. Based on the results, we then discuss these measures in the context of qualitative annotations for machine learning datasets.

3.4 Results

The mean IRR scores from 400 to 450 reviews for each sustainability aspect are shown in Fig. 3.2 below. Along the horizontal axis we have three sets of horizontal bars, one for each sustainability aspect. Within each set are the results of the different variations of Krippendorff's U-alpha variations described in section 3.3.

[3] https://github.com/ndehaibi/krippendorff-alpha-irr

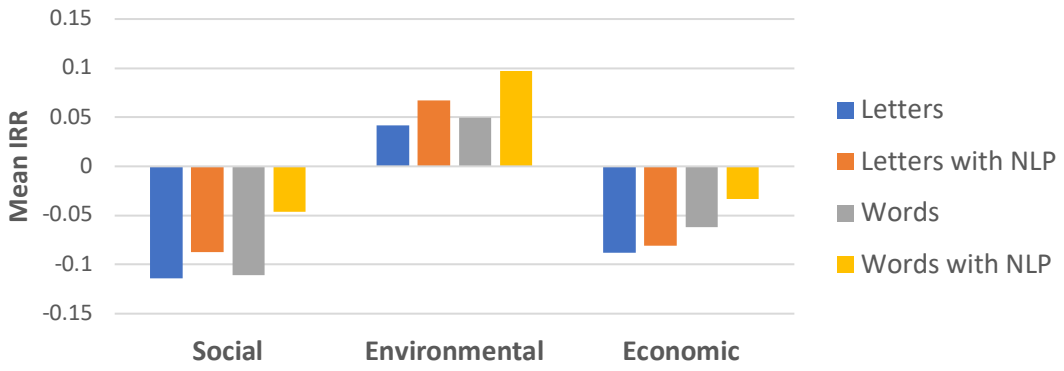

Figure 3.2: Mean IRR scores for each sustainability aspect

In Figure 3.2 we see the IRR scores for environmental sustainability in the middle set were highest on average, ranging from 0.042 to 0.097. The IRR scores for economic sustainability in the right set were second highest on average, ranging from -0.088 to -0.033. The IRR scores for social sustainability in the left set were the lowest on average, ranging from -0.114 to -0.046. We also see that pre-processing the reviews with natural language processing and looking at word counts resulted in the highest IRR scores on average. Pre-processing reviews with NLP and looking at letter counts also increased scores, but not as much.

The difference in Krippendorff's U-alpha between word counts compared to letter counts is negligible in the absence of NLP. For example, the mean IRR scores for environmental sustainability are 0.042 and 0.0498 for letter counts and words counts respectively. The negligible difference without NLP was expected because despite having smaller distances with word counts, the overall difference gets normalized by a smaller review length compared to when looking at letter counts.

We also see in Figure 3.2 that on average the IRR scores revolve around 0; environmental sustainability was slightly above 0 on average while social and economic

sustainability were slightly below 0. In sections 3.4.1 to 3.4.3 we present the distributions of the IRR scores for each sustainability aspect, and in section 3.5 we offer insights about these results.

3.4.1 Social Sustainability

The mean IRR scores and standard deviations for social sustainability are presented in Table 3.1.

Table 3.1: Mean IRR scores and standard deviations for social sustainability

	Review Count	IRR Mean	IRR Standard Deviation
Letters	449	-0.114	0.848
Letters with NLP	441	-0.087	0.860
Words	449	-0.111	0.853
Words with NLP	441	-0.046	0.813

Histograms of social sustainability IRR scores for each Krippendorff's U-alpha measure are shown in Figure 3.3.

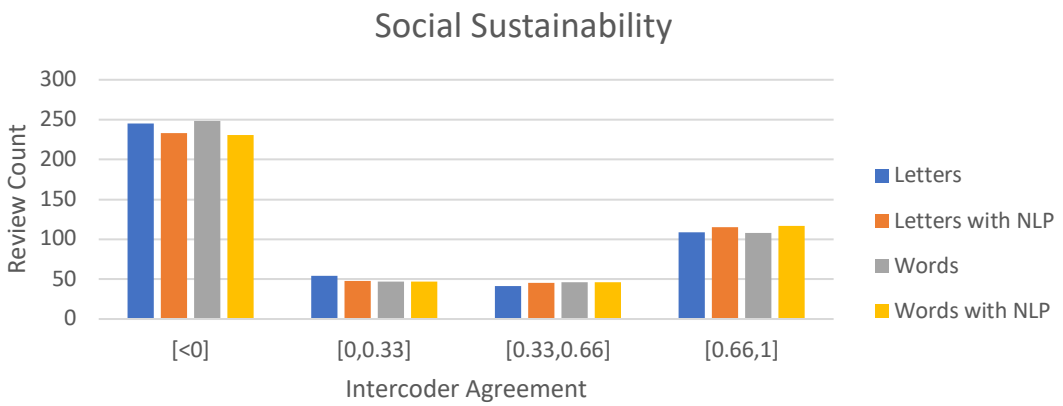

Figure 3.3: IRR for social sustainability

From Table 3.1, we can see that the review counts change from 449 to 441 when they are pre-processed with natural language processing. This is because some reviews may have annotations that contain only numbers or stop words; pre-processing in these

cases would remove the annotation entirely. Table 3.1 also shows that the standard deviations are large relative to the mean. This is demonstrated by the distributions in the histograms shown in Figure 3.3 of IRR scores for each Krippendorff's U-alpha measure. Despite a mean score of around 0, the IRR scores for social sustainability range from about -3 to 1.

The histograms follow a similar distribution for all the Krippendorff's U-alpha measures; there is a spread from scores below 0 to 1 with the highest count of reviews being closer to 1. We see slight improvements in scores with measures that include NLP.

3.4.2 Environmental Sustainability

The mean IRR scores and standard deviations for environmental sustainability are presented in Table 3.2.

Table 3.2: Mean IRR score and standard deviations for environmental sustainability

	Review Count	IRR Mean	IRR Standard Deviation
Letters	404	0.042	0.773
Letters with NLP	399	0.067	0.814
Words	404	0.0498	0.757
Words with NLP	399	0.097	0.779

Histograms of IRR scores for each Krippendorff's U-alpha measure are shown in Figure 3.4. While the scores are higher here than social sustainability, the distributions are very similar.

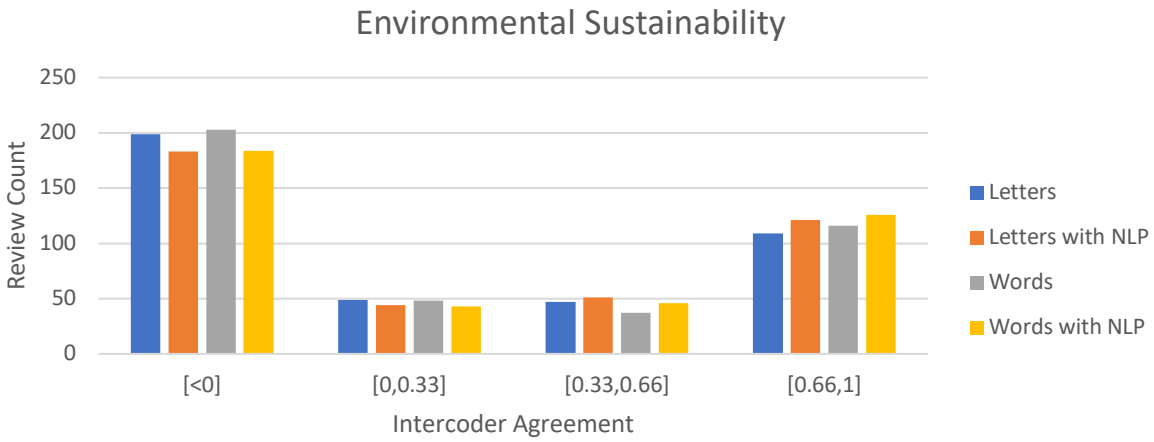

Figure 3.4: IRR for environmental sustainability

3.4.3 Economic Sustainability

The mean IRR scores and standard deviations for environmental sustainability are presented in Table 3.3.

Table 3.3: Mean IRR score and standard deviations for economic sustainability

	Review Count	IRR Mean	IRR Standard Deviation
Letters	436	-0.088	0.905
Letters with NLP	433	-0.081	0.920
Words	436	-0.062	0.882
Words with NLP	432	-0.033	0.865

Histograms of IRR scores for each Krippendorff's U-alpha measure are shown in Figure 3.5. Again, we see a very similar distribution compared to the other two sustainability aspects.

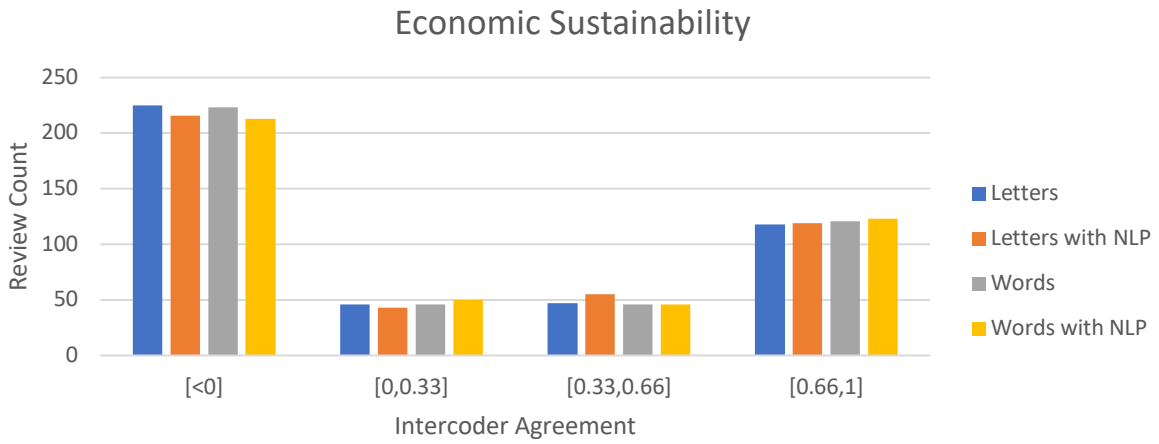

Figure 3.5: IRR for economic sustainability

3.5 Discussion

Considering these results, we present a discussion on how designers can use IRR scores in the context of assessing reliability of qualitative text annotations for machine learning datasets. The Krippendorff's U-alpha variations we implemented did not have a large effect on the IRR scores and so we examined the annotations that had the lowest IRR scores to better understand the results. Below is an example of one of these annotations that received an IRR score of -3:

Review: Did not last a month of light use (every other day or so). The plastic nub that holds the strainer in place broke and now it's useless.

Annotator 1 highlight: The plastic nub that holds the strainer in place broke and now it's useless.

Annotator 2 highlight: Did not last a month of light use (every other day or so).

In this example, the first annotator highlighted the second half of the review while the second annotator highlighted the first half of the review. From the lens of Krippendorff's U-alpha, these annotations have no overlap and span different halves of

the overall review. This suggests there are systematic differences between the annotators. Looking at this pair of annotations however we see that, semantically, both annotations revolve around durability of the product. The first sentence is a general statement about the durability, while the second provides more detail to explain the first statement. Therefore, there is some redundancy in the review and the annotations might be more closely related than IRR suggests. For this reason, we suggest that having a low Krippendorff's U-alpha score in the context of qualitative annotations for machine learning may not necessarily reflect a low agreement between the annotators.

We also propose that having a high agreement score may not be desirable or effective when building an annotated dataset for machine learning. Referring to the annotation example above, we see that one annotator highlighted the first half of the review which was a general comment, while the other annotator highlighted the other half which was more specific. If both annotators had highlighted the general phrase, we would not have gained as much useful information despite having a higher agreement score. Therefore, in the context of machine learning, we suggest that the annotation task becomes more effective as a hunting exercise where we collect as much relevant information as we can. This is particularly the case with NLP tools, for example, term-frequency inverse-document-frequency (TF IDF), that can reduce the importance of redundant terms and increase the importance of unique and specific terms in models. TF IDF is the product of term frequencies and inverse document frequencies (Eq. 3.11) (Jurafsky and Martin, 2017).

$$w_{ij} = tf_{ij} * \log\left(\frac{N}{df_i}\right) \tag{3.11}$$

Equation 3.11 shows the TF IDF weight w_{ij} for word i in document j where N is the total number of documents and df_i is the number of documents where the word i occurs. The TF IDF transformation gives a higher weight to words that occur only in a few documents. Therefore, having a high agreement score becomes less important in this context when machine learning can mitigate annotations with less useful information while also benefiting from a larger dataset.

Particularly with qualitative topics such as sustainability, it is expected that people will have different perspectives even if annotation training is provided. While a high IRR score may not be desirable in this context, we suggest that IRR can still provide useful information for designers. Looking at Figure 3.2, we see that on average environmental sustainability has a higher IRR than the other two sustainability aspects. This could suggest that annotators have a slightly more united perspective on what environmental sustainability is compared to social and economic sustainability. Environmental sustainability is generally the more prevalent aspect of sustainability and users may be more familiar with their perception of it, therefore reducing redundancies in reviews. This can inform designers on how they chose to design products involving environmental aspects. Moreover, looking at the histogram distributions in Figures 3.3 to 5, we see that there are four buckets of IRR scores ranging from below 0 to 1. These distributions could be useful to designers as they cluster annotations with high agreement and lower agreement, therefore helping designers identify perceptions that are more prevalent and perceptions that are more niche (perceptions on sustainability in this case). For example, using the clusters designers could identify sustainability

perceptions that have a consensus among customers, or focus on smaller market segments by looking at disagreements in the clusters.

Coming back to measuring reliability of qualitative annotations in machine learning datasets, we suggest that external validity metrics of the model, for example, accuracy, precision, and recall, are more effective measures despite being foreign in the design research space. To calculate these metrics, we would split the data into training, validation, and test sets and train the model to make sure that it is working, and outputting results as expected [35]. Based on the metric scores of the model we would then be able to infer it the annotation dataset is reliable.

3.6 Conclusion

The goal of this study is to help designers measure reliability of qualitative annotations in the context of machine learning datasets using metrics that are common in the design research space. We investigated inter-rater reliability (IRR) as an internal validity measure by leveraging annotations of text data from a previous study where annotators highlighted social, environmental, and economic aspects of sustainability in online product reviews of French Press coffee makers. We calculated IRR scores of the annotations using four variations of Krippendorff's U-alpha: the first is the baseline where we looked at letter counts to measure differences between annotations, the second is where we looked at word counts to measure differences, and the third and fourth are the same as the first and second but with natural language processing of the reviews. The purpose of the variations was so that we may tune how sensitive the IRR measures are in different annotation scenarios.

We found that, while the variations slightly increased IRR scores from the baseline, the IRR scores on average ranged between -0.1 to 0.1. We examined annotations with the lowest scores and found that a low IRR score in the context of qualitative annotations may not necessarily reflect a low agreement between annotators due to potential redundancies in semantics. Moreover, we found many examples where, despite having low IRR scores, the annotators still captured useful information. In the case of machine learning datasets, we suggest that having a low IRR score might be preferable over high agreement between annotators to provide more unique data for a model to learn from. We discussed how this is particularly the case when tools such as TF IDF can help balance for annotations with less useful content. Based on the results we propose that IRR can still be useful for designers in this context by clustering customer perceptions based on how well users agree or disagree on them. In terms of measuring reliability of this type of dataset, we propose that using external validation metrics, for example, accuracy, precision, and recall, are a better indicator of data quality despite them being foreign in design research. The results and discussions in this study are limited to the context of highly qualitative annotation tasks that are used as machine learning datasets.

4. CHAPTER 4

VALIDATING PERCEIVED SUSTAINABLE DESIGN FEATURES USING A NOVEL COLLAGE APPROACH

Abstract

Designers are challenged to create sustainable products that resonate with customers, often focusing on engineered sustainability while neglecting perceived sustainability. We previously proposed a method for extracting perceived sustainable features from online reviews using annotations and natural language processing, testing our method with French press coffee carafes. We identified that perceived sustainability may not always align with engineered sustainability. We now investigate how designers can validate perceived features extracted from online reviews using a relatively new design method of collage placement where participants drag and drop products on a collage and select features from a drop-down menu. We created collage activities for participants to evaluate products on the three aspects of sustainability: social, environmental, and economic, and on how much they like the products. The activity used French presses as a case study where participants placed products along the two axes of the collage, sustainability and likeability, and labeled products with descriptive features. We found that participants more often selected our previously extracted features when placing products higher on the sustainability axis, validating that the perceived sustainable features resonate with users. We also measured a low correlation between the two-axes of the collage activity, indicating that perceived sustainability and likeability can be measured separately. In addition, we found that product perceptions across sustainability aspects may differ between demographics. Based on these results,

we confirm that the collage is an effective tool for validating sustainability perceptions and that features perceived as sustainable from online reviews resonate with customers when thinking of various sustainability aspects.

4.1 Introduction

With the growth of e-commerce platforms, designers are challenged to create products that resonate with customers so that they stand apart from the competition. When creating sustainable products, designers typically rely on engineered sustainability tools, for example, life cycle analyses (LCA), to guide their decisions. Perceived sustainability, however, is an often-missed factor that can help designers differentiate their products and influence purchasing decisions [56]. Perceptions of sustainability reflect what customers think is sustainable, which may not always align with engineered sustainability. We showed previously that customers can perceive sustainability differently from engineered sustainability [45]. For example, customers perceive that "natural" materials such as stainless steel and glass are what make a coffee carafe sustainable, but, it is the engineered sustainability such as the auto-off energy saving feature that has the most benefit to the environment.

This disconnect between perceived and engineered sustainability can lead to misinformed purchasing decisions, such as a customer not purchasing a sustainable product because they perceive it as not sustainable [57]. Perceived sustainability is often overlooked when making sustainable engineering decisions. Following the coffee carafe example above, a carafe that includes an auto-off feature should also include natural materials, instead of being all plastic. This creates an alignment for customers

that the product is sustainable. With this alignment, designers can potentially differentiate the product in the market and drive purchases. The benefits of sustainable features are limited if the product does not also have market success, and achieving that success relies on both perceived sustainability as well as engineered sustainability.

Designers can differentiate their products by adding features that resonate with customers compared to other options. For customers to resonate with sustainable products, they must (1) identify them as sustainable and (2) like them. The current approach for designers to differentiate sustainable products is to provide information about engineered sustainability, either through online product descriptions or eco-label packaging [58]. Engineered sustainability information, however, can lead to anxiety or confusion when a person has a limited understanding of it [38]. Eco-labels can also trigger the altruism-sacrifice heuristic, where customers expect to sacrifice performance for sustainability [59]. O'Rourke and Ringer investigated how sustainability information affected 40,000 purchase interactions on GoodGuide.com over a 12-month period [60]. The authors found that engineered sustainability information tends to be significant only for those that directly seek it, which is not enough to influence mainstream customer behavior. Therefore, only providing engineered sustainability information has a limited effect on consumers.

An alternative approach for designers to differentiate sustainable products is to design-in perceived sustainability. Although perceived sustainability features, such as using natural materials, may not contribute towards engineering sustainability goals, such as low energy use, they resonate with customers as "sustainable," and such

features have been shown to change purchase intentions [38]. She and MacDonald investigated visible product features that capture sustainability perceptions termed "sustainability triggers" [2]. The authors found that the triggers led customers to think about sustainability-related criteria as well as prioritize sustainability features in simulated decision scenarios of realistic toaster prototypes. Designers can therefore use perceptions to help lead customers to accurate information about a product.

As more customers rely on online shopping, designers have access to a growing source of customer perceptions in the form of product reviews. These perceptions can offer designers insights and help guide their decisions. A growing body of research is developing methods for designers to tap into perceptions in online reviews. For example, Joung and Kim filtered online reviews for product feature perceptions using Latent Dirichlet Allocation (LDA) to identify automated keywords and validate their method using Amazon reviews of Android smart phones [61]. They found that their approach with LDA yielded better topic coherence than previous methods. Moreover, Hou et al. captured changes in customer perceptions over time using a rule-based natural language processing (NLP) method to extract features and conjoint analysis to categorize the features, validating the approach using reviews of two generations of a Kindle [62]. Their case study demonstrated how designers can use their method to improve new and existing products. In a previous paper, we developed a method for designers to extract features that are perceived as sustainable using annotations of online reviews and natural language processing, testing the method with Amazon reviews of French press products (see Section 4.2.2) [45]. While we demonstrated that

there is a gap between engineered and perceived sustainability, we did not validate if perceived sustainable features extracted from online reviews resonate with customers as sustainable.

In this study, we propose a method to validate the effectiveness of features perceived as sustainable from online reviews to help customers resonate with sustainable products. With this method sustainable designers can confirm whether features extracted from online reviews resonate with customers as sustainable; thus, enabling designers to confidently consider perceived sustainability in their products. The method involves a novel collage approach where participants drag and drop products onto a two-by-two axis and select features from a drop-down menu to describe the products. Specifically, we asked participants to evaluate products and features based on their perceived sustainability and likeability. The collage activity has previously been used in design to gain insights into creating sustainable products that resonate with customers (see section 4.2.3 for more on collages).

The rest of the paper is organized as follows: Section 4.2 presents a background on customer perceptions in design, Section 4.3 outlines the research propositions and hypotheses, we describe our method in Section 4.4, Section 4.5 presents the results and analysis, our findings are discussed in Section 4.6, and we make conclusions in section 4.7.

4.2 Background

Humans develop perceptions using sensory input information and interpret them based on available information, their thoughts, and prior experiences [63].

Perceptions can differ from person to person and change over time [64]. MacDonald et al. showed how customers develop product perceptions on a case-by-case basis [65]. The authors investigated perceptions of paper towels using a discrete choice survey and found inconsistencies in preferences when provided with crux (complex) attributes versus sentinel (simple) attributes. The findings highlight the designer's role to communicate relevant product information to customers when making purchase decisions.

Due to the subjective nature of perceptions, they may or may not accurately represent the context. Understanding perceptions is therefore critical for designers to communicate information accurately to customers. Below we provide a review on (1) works that investigate customer perceptions in sustainable design, (2) our previous work on developing a method to extract customer perceptions from online product reviews, and (3) the use of collages in the design space to assess customer perceptions.

4.2.1 Customer Perceptions in Sustainable Design

In this section we provide a literature overview on understanding customer perceptions in design.

Borin et al. investigated the effects of positive, negative, and no environmental information on consumer perceptions [66]. The authors recruited 329 participants and evaluated products in different categories including apples, bath soap, MP3 headphones, and printed paper. Within each product category there were five environmental messages ranging from very positive to very negative. The authors found that the positive environmental information did not change customer perceptions or

purchase intent compared to having no environmental message, however participants viewed products with positive environmental information better than those with negative environmental information. Therefore, highlighting negative features that a sustainable product avoids may be more effective than highlighting its positive features.

Maccioni et al. investigated the difference between conscious and unconscious perceptions of sustainable products [67]. They recruited 43 participants to evaluate 20 baseline products and 20 sustainable products in the same categories. The authors measured conscious perceptions with self-assessments and unconscious perceptions with biometric measurements. They found participants did not experience any emotional reactions to eco-design efforts because they could not identify sustainability. The authors also found that baseline products were perceived as more functional and reliable, while only participants showing high interest in sustainability perceived sustainable products as more innovative.

Steenis et al. investigated the role of product packaging on perceived sustainability [68]. The authors recruited 249 participants and tested their perceptions of soup products with different packaging material and graphics. They found that packaging is not a strong contributor to sustainability perceptions of a product, but that it is a strong contributor to perceptions of product quality and taste. The authors recommend that sustainable packaging can be effective if it enhances perceptions of quality and taste.

Catlin et al. explored the differences between user perceptions of social and environmental sustainability [69]. They recruited 422 participants and asked them to

pick between two chocolate bars, one was promoted as socially sustainable while the other was promoted as environmentally sustainable. The participants were then asked to explain their choices. The authors found that participants perceive social sustainability more with affective, short-term, and local considerations while they perceived environmental sustainability with analytical, long-term, and global considerations. Since consumers tend to focus on short-term needs over long-term, the authors suggest that social sustainability is more likely to resonate with consumers over environmental sustainability.

4.2.2 Extracting Feature Perceptions from Online Reviews

In this section we provide a background on identifying perceptions of product features from online reviews. Rai was one of the first in the design space to develop automated methods for extracting value from online reviews [70]. The author used a part-of-speech tagger with a term-document matrix to identify salient features and validated the method using reviews of a camcorder. Several works since then have developed methods using machine learning models to derive design value from reviews, but a gap remained in identifying the differences between perceived and engineered features.

Motivated by this gap, we previously proposed a four-step method to help designers extract product features perceived as sustainable from online reviews (Fig. 4.1) [45]. We tested the method using online reviews of French presses and demonstrated that the method enables designers to uncover multifaced insights beyond sentiment of product features. To the best of our knowledge, this was the first design

method that integrated research areas from rating design ideas, identifying customer perceptions, and natural language processing. We provide details on our previously proposed method here as we build off it in this paper.

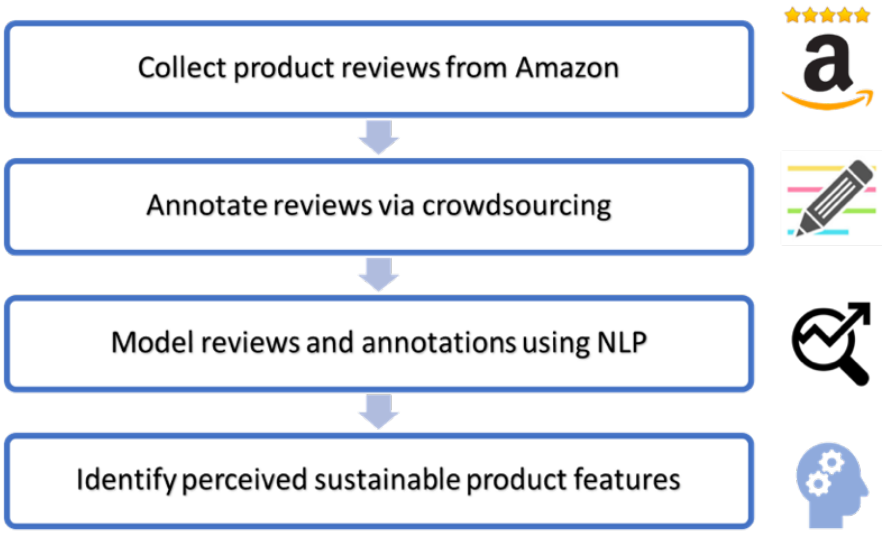

Figure 4.1: Extracting customer perceptions method flow

The method involved crowdsourcing annotations of online reviews to build a natural language processing algorithm and then extracting features perceived as sustainable from the parameters of the model. To test this method, we recruited 900 Amazon Mechanical Turk (MTurk) respondents to annotate 1474 reviews of French presses by highlighting phrases based on what they perceive is sustainable and indicating the sentiment in the phrases. We collected annotations for each of the three sustainability aspects: social, environmental, and economic. Using a machine learning NLP algorithm, we identified the most salient features perceived as sustainable from the highlighted phrases that drove positive and negative sentiment for each aspect. See Table 4.1 for most salient positive features and Table 4.2 for most salient negative features extracted from this approach. Note that these features represent how users

perceived sustainability based on the annotations and may not actually contribute to engineered sustainability. Our method captures features in the form of text directly from reviews, enabling designers to identify what influences sustainability perceptions.

Table 4.1: Positive features of French presses perceived as sustainable

Social Aspects	**Environmental Aspects**	**Economic Aspects**
Easy to use	Well made	Easy to clean
Love it	Easy to use	Great quality
Nice gift	Strong glass	Want more than one
Good for my family	Easy to clean	Reasonable price
Perfect for two	Solid design	Works great
Use with my spouse	Will last	Worth the price
Take to work	Stainless steel	Good customer service
Easy to clean	No plastic	Great value
High quality	Metal frame	Best price
Works great	Sturdy	Hard to beat

Table 4.2: Negative features of French presses perceived as sustainable

Social Aspects	**Environmental Aspects**	**Economic Aspects**
Difficult to use	Too much plastic	Advertised falsely
Looks flimsy	Glass is too thin	Looks cheap
Difficult to wash	Falls apart easily	Waste of money
Glass breaks easily	Glass breaks easily	Glass shatters easily
Sharp corners	Difficult to take apart	Poor design
Metal rusts	Too fragile	Poor customer service
Falls over easily	Plunger leaks	Don’t like this brand
Handle hurts	Handle is plastic	Won’t buy this
Fragile glass	Does not last	Too expensive
Too small	Rusts easily	Not worth the money

Positive features perceived as socially sustainable were mainly intangible, such as “nice gift” or “good for family”, while negative features were mostly tangible relating to safety issues, such as “glass breaks easily” or “sharp corners”. Positive features perceived as environmentally sustainable were more tangible, such as “stainless steel” or “no plastic” while negative features related to the durability of the product, such as

“glass breaks easily” or “too fragile”. Positive features perceived as economically sustainable related to the product being of good value while negative features related to the product not being worth the price. It is important to consider context when assessing perceived features. For example, “easy to use” for social aspects may relate to safety while for environmental aspects it may relate to reliability.

Tables 4.1 and 4.2 show that material is a salient perceived environmental concern for French press carafes while energy and water consumption features are not. In reality, energy and water consumption have a much higher environmental impact for French press carafes [45]. This demonstrates the gap between perceived and engineered sustainability and highlights the importance for designers to consider both when creating sustainable products.

The literature has yet to explore how extracted features from online reviews that are perceived as sustainable resonate with customers when thinking of various sustainability concerns. We aim to fill the research gaps by proposing a method to validate if extracted perceived features in Tables 4.1 and 4.2 will resonate with customers. The goal is to provide designers a method for validating perceived features to communicate sustainability more accurately to customers and better differentiate sustainable products in the market.

4.2.3 Evaluating Products using a Collage

A collage in design research is a set of two axes that range on specific criteria. For example, one axis might range from relaxing to not relaxing while the other axis ranges from like to dislike. Participants then place items on the collage to evaluate

based on the criteria. Using the responses designers can identify customer perceptions for the selected criteria. In this section we cover two applications in design research for using the collage approach to identify customer perceptions.

Guyton was one of the first to use a collage as a systematic method for designers to gain insights into creating sustainable products that resonate with customers [71]. Rather than evaluating products on a single axis, the collage consisted of two axes ranging from “unsustainable to “sustainable” and “dislike” to “like”. To validate the method, participants placed images of products on the collage and selected words from a vocabulary list to describe the products. Several products were tested including spatulas, mugs, and boots. Based on the product placements and vocabulary used, Guyton demonstrated the collage activity as an effective method for capturing sustainability perceptions of products.

Liao et al. build on this approach by using the collage as an evaluation tool for participants to express their product preferences [72]. The authors explored how users perceive products based on a product’s form and visible characteristics using a two-axis interactive collage tool to evaluate eight wearable products. The authors evaluated comfort, delight, and usefulness on three separate collages, respectively. They generated a list of emotional descriptive words to identify perceptions. The authors developed the collage activity into a webapp where participants drag and drop product images and select descriptive words from a dropdown list (Fig. 4.2). Based on responses from 400 participants, the tool revealed relationships between product characteristics and user perceptions. For example, wearables that resembled clothes were perceived

as more delightful and comfortable. Moreover, the authors found a high correlation between likeability and the other axes of the collage: comfort, user delight, and usefulness.

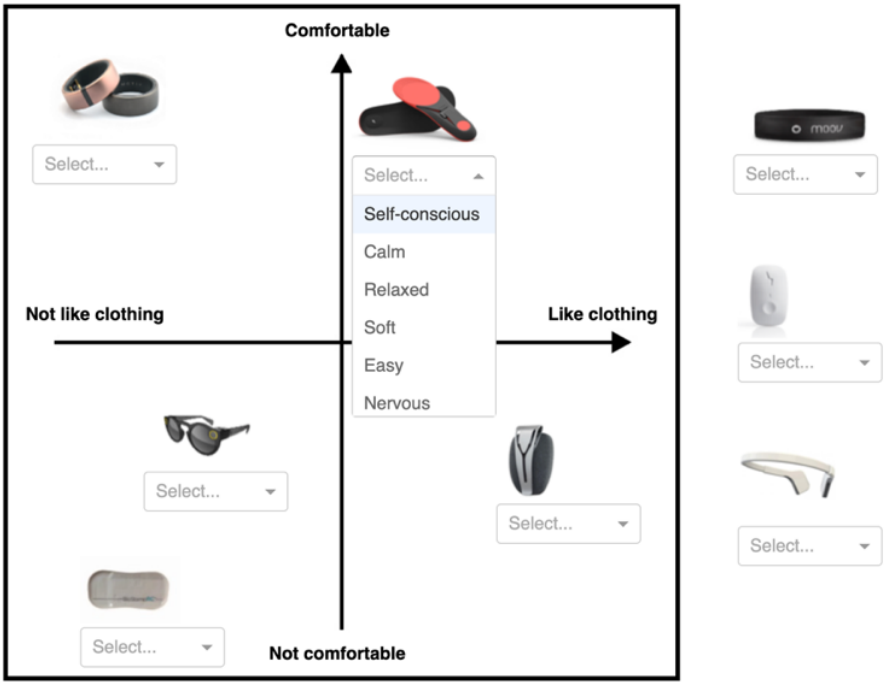

Figure 4.2: Example of a collage tool from Liao et al. [72]

Based on these previous studies, we used the collage in this study as an engaging way for participants to evaluate product sustainability. The collage tool has proven itself as an intuitive way to get input from users and to measure the relationship between products and user emotions. By allowing participants to actively choose one or more features without drawing attention to them, we can determine if participants resonate with those features. Moreover, evaluating the placement of products on two axes enables us to separately study how participants like a product and their perceived sustainability of the product.

4.3 Research Propositions and Hypotheses

This work proposes a method to validate whether perceived sustainable features extracted from online reviews resonate with customers using a collage tool. The two axes used in the collages are likeability and sustainability. While perceived sustainability features may not contribute to engineered sustainability, they can register as sustainable to customers and help them create cognitive alignment with sustainable products. The selection and position of the features on the collage are recorded and tested to validate that perceived sustainable features are evaluated as sustainable. We also validate whether perceived sustainability is associated with likeability by recording and testing the position of the products on the collage. The following propositions and hypotheses are tested.

Proposition 1: Designing-in perceptions can help customers create an alignment between perceived sustainability and sustainable products. Based on this, we propose that customers will resonate with perceived sustainable features as being sustainable.

Hypothesis 1: Participants evaluating product sustainability on a collage will select features perceived as sustainable for products that they place higher on the "sustainability" axis of the collage.

Proposition 2: Customers tend to like products that create cognitive alignment for them, and perceptions can help them achieve that. We therefore propose that perceptions of product sustainability contribute to how much customers like a sustainable product. This is motivated by prior research that 73% of millennials are willing to pay more for sustainable products [7].

Hypothesis 2: A statistically significant relationship exists between the placement of a product on the "sustainability" axis of the collage, and the "like" axis of the collage.

4.4 Method

To test the hypotheses, we designed an activity for 1200 respondents from Amazon Mechanical Turk (MTurk) to evaluate French press products based on sustainability criteria. MTurk is a crowdsourcing platform to recruit workers for completing tasks. We refer to the respondents as participants in this paper (see Section 4.4.4 for more information on participants). The activity consisted of three parts: (1) a pre-survey to learn about the sustainability criteria and get familiarized with the products, (2) a collage tool that was adapted from previous work by Guyton [71] and Liao et al. [72], and (3) a post-survey where participants answered follow-up questions about the products and demographics (Fig. 4.3).

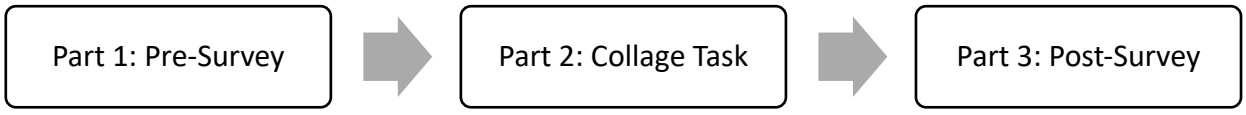

Figure 4.3: Breakdown of the three parts of the activity

We designed three versions of this activity to evaluate products on the three sustainability aspects separately: social, environmental, and economic (Fig. 4.4). We randomly assigned participants to one of the three versions. Our choice to have participants focus on one sustainability aspect was motivated by our previous study where a pilot test showed that this led to better clarity for participants and more usable responses [45]. While focusing on one aspect is not realistic for a purchasing scenario, it

provides participants clarity for evaluating products on the collage which is crucial for our study.

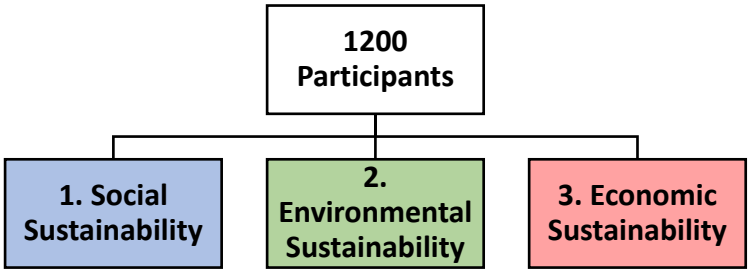

Figure 4.4: Participants distributed across three activity versions

The pre-survey, collage task, and post-survey are described in detail below.

4.4.1 Pre-Survey

In the pre-survey, participants were screened for eligibility (check section 4.4.4 for information on participants), trained and tested on sustainability criteria, and familiarized with the products and the collage activity. Sections 4.4.1.1 and 4.4.1.2 provide more information on the criteria and products, respectively.

4.4.1.1 Sustainability Criteria

Participants were trained on sustainability criteria during the pre-survey to prepare them for evaluating product sustainability in the collage task. Each version of the activity had a customized training portion based on one of the three sustainability aspects. To train participants we displayed to them sustainability evaluation criteria as defined in El Dehaibi et al. [45], shown in Fig. 4.5. We chose these definitions because they provide participants with guidelines on how to evaluate products without overriding their personal opinions. For example, while the social sustainability definition includes health and safety as a criterion, participants can determine on their own what

makes a product healthy or safe. We trained participants to focus on a specific sustainability aspect and ignore others. As an example, if participants were working on the social sustainability version of the activity, they were trained to focus on social sustainability and not to consider environmental and economic aspects. We chose this approach after a pilot study revealed that participants were still evaluating products using mixed sustainability criteria when we specified what criteria to focus on only.

Figure 4.5: Sustainability aspect definitions and training

We did not provide training on how to evaluate "likeability" of the products. While sustainability is a multi-faceted concept, likeability is more of a feeling that humans can detect. We therefore let participants decide how to evaluate likeability and what it means for them.

Following the training portion, participants completed a test that they had to pass to make sure they had understood the training. The test consisted of two multiple choice questions: the first question asked participants to select factors they will evaluate according to their sustainability aspect, and the second question asked participants to select factors they will not evaluate according to their sustainability aspect. For example, in the social sustainability version, participants could choose

"family and culture" for the first question and "energy and water consumption" for the second question.

Six French presses were used in this activity (shown in Table 4.3). We chose French press products to build off our previous work [45]. French presses are ubiquitous and likely to contain features related to sustainability. We selected the French presses in Table 4.3 for this study based on their varying aesthetic features and materials that cover the design space (e.g., stainless steel, plastic, glass, wood, etc.) as well as their varying price points, number of ratings, and reviews. Our goal was to provide participants with enough variety so they can evaluate products differently on the collage. Participants were presented with direct links to the Amazon product pages in the pre-survey so that they could get familiarized with the products. The order of the products was randomized between each participant. We asked the participants to consider the product features, price, and reviews for each product. The Amazon links opened as popups instead of new windows or tabs to reduce the number of participants we may lose from completing the activity. The Amazon popup was equivalent to opening a browser in "Incognito mode" so that participants' prior browsing history did not influence dynamic content shown on the pages, such as recommended Amazon products or reviews. This facilitated having a common baseline of live Amazon pages between participants. Other factors may still bias the dynamic content shown such as geographic location of participants, but the random choice of the MTurk participants likely limited the influence of these factors. If participants skipped any of the Amazon links or proceeded too quickly, they were shown the product links again and on the third

time, they exited from the activity. We included this step to make sure that participants were familiar with the products before evaluating them.

Table 4.3: List of products

						
Product Name	Chef	Frielin	Madrid	Melbo	Brookl	Terra
Price	$14.99	$56.44	$35.00	$39.99	$29.99	$19.99

4.4.2 Collage Activity

After completing the pre-survey successfully, participants started the collage task to evaluate products. We asked them to drag and drop products on the collage and select features from a dropdown list to describe each product based on the sustainability criteria. Participants could select the features before or after dragging the product, and we did not provide additional guidance on what features to select. Moreover, participants were able to modify their product placements and feature selections up until they completed the collage activity. While participants may alter their feature selections to justify their product placement and vice versa, their justifications would still validate that they evaluated perceived sustainable features as more sustainable. We tested hypothesis 1 based on the placement of the selected features on the collage, and we tested hypotheses 2 based on the placement of the products on the collage.

4.4.2.1 Interface

We used the same collage tool interface from the webapp used in Liao et al. [72] but modified it to include a two-by-two grid, a set of products on the right side, and an

evaluation criteria button on the left side. The x-axis on the grid ranges from “Dislike” to “Like” while the y-axis depends on the version of the activity; ranging from “Not XX Sustainable” to “XX Sustainable,” where "XX" represents one of the three sustainability aspects: social, environmental, economic. We chose a two-by-two grid in this study for three reasons: first, it allows us to differentiate between what participants like about a product and what they determine is sustainable; second, it is a tested tool that has been used in literature for evaluating product perceptions [71,72], and third, it is an engaging way for participants to evaluate product sustainability. Figure 4.6 shows an example of the interface for the social sustainability version.

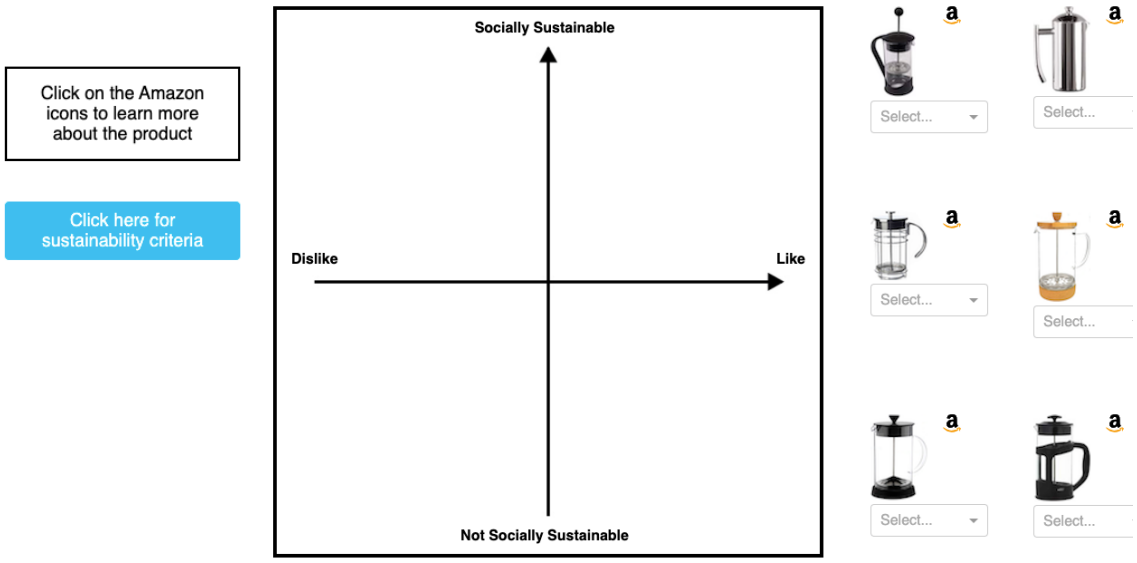

Figure 4.6: Collage tool interface for social sustainability

Clicking on the sustainability criteria button on the left opens a popup with information from the training section of the pre-survey. An example from the social sustainability version is shown in Fig. 4.7.

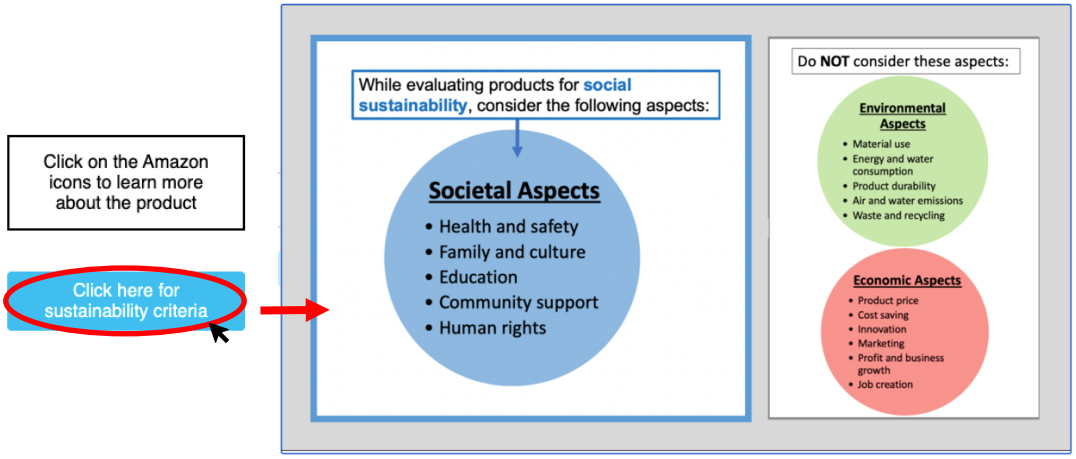

Figure 4.7: Evaluation criteria button for Social Sustainability

Each product image has an Amazon icon on the top right corner which links to a popup of the product's Amazon page when clicked on. An example is shown in Fig. 4.8. Similar to the pre-survey, Amazon popups in the collage opened akin to a browser in "incognito mode" so that past browsing history did not influence contents shown to participants. For both the sustainability criteria button and Amazon icon, participants could close the popup by clicking outside of the popup to return to the collage.

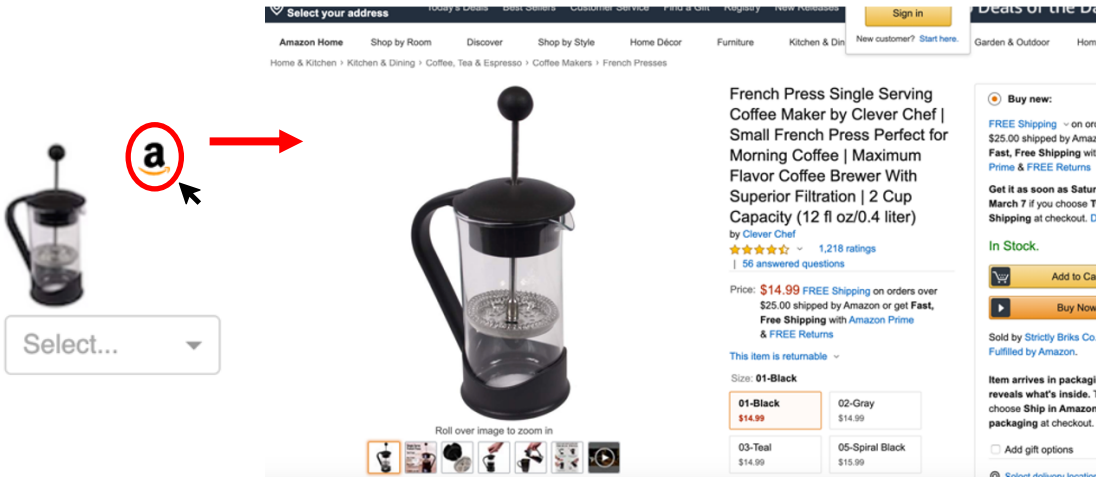

Figure 4.8: Amazon product page popup example

To evaluate the products, participants dragged and dropped the products onto the collage and then selected features from a dropdown list to describe each product based on the criteria (Fig. 4.9). See Section 4.4.2.2 for information on the features we provided. We asked participants to place all products on the grid and select at least one feature from the dropdown list for each product to proceed. The dropdown list was randomized between participants, and they could select as many features as they liked.

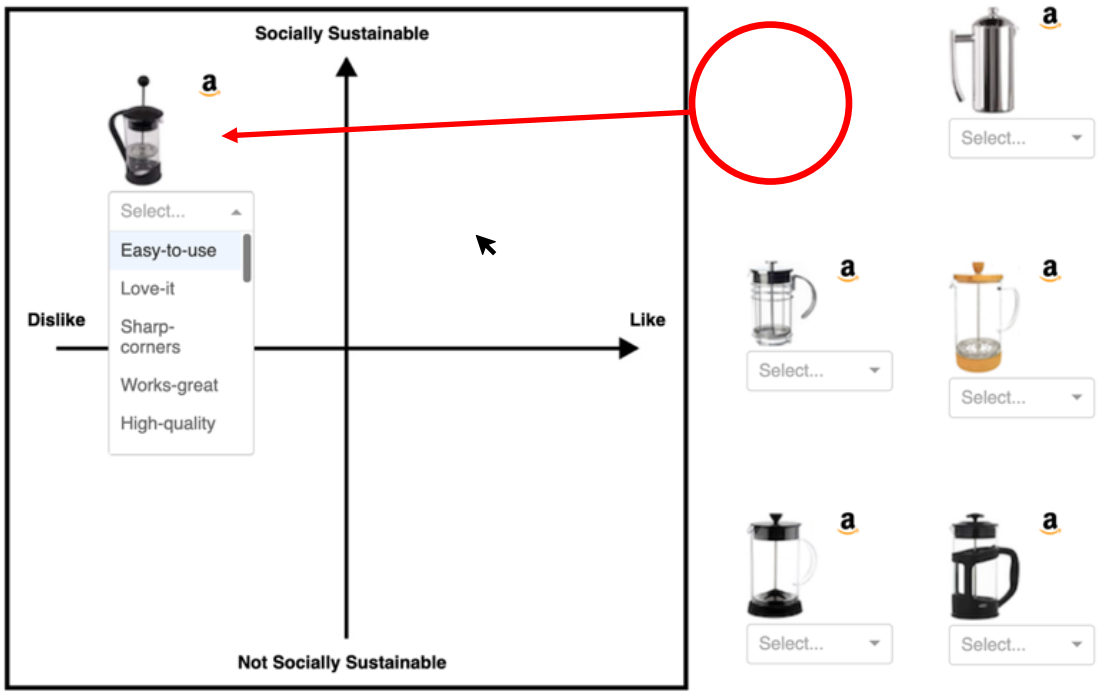

Figure 4.9: Dragging and dropping products on collage and selecting at least one feature to describe each product

Participants could also relocate the images on the collage until they proceeded to the next page. We recorded the location of the center of the product image as a float number. While reducing the location of the image to one point adds uncertainty, the impact is negligible since our focus is on the relative placement of products and features. On the next page, we asked participants to rate how relevant each of the

features they selected are to sustainability on a 5-point Likert scale. The scale labels included “Not at all related” for a 1 out of 5 rating, “Somewhat related” for a 3 out of 5 rating, and “Very related” for a 5 out of 5 rating. This gave us insight on which features were selected based solely on likeability versus perceived sustainability. After rating the features, participants were able to submit their evaluations and received a password to proceed with the post-survey.

4.4.2.2 Interface

Below we discuss two sets of features that we presented to participants: (1) positive and negative features perceived as sustainable, and (2) positive features perceived as sustainable and features not perceived as sustainable.

4.4.2.2.1 Positive and Negative Features Perceived as Sustainable

In the collage task we provided participants with a set of 20 features in a dropdown list to select from for each product. These features were extracted from online reviews of French presses in our previous work and are shown in Tables 4.1 and 4.2 for each of the sustainability aspects [45]. While our original study extracted 40 features for each sustainability aspect, we selected a subset of 20 features for this study to account for overlaps as well as to have a similar quantity of features as in previous collage experiments [72]. Our goal in this study is to validate that these features resonate with customers as sustainable using the collage approach. The selected features consist of the 10 most positive and 10 most negative features perceived as sustainable, according to review annotations and a machine learning algorithm. While the extracted features are perceived as sustainable, they may not actually contribute to

engineered sustainability. Each sustainability aspect has a corresponding set of 20 features. Participants actively elected to select at least one of these features for each product they placed on the collage and could select the features before or after placing the products. We did not provide additional guidance on which features to select. Out of the features participants selected, we analyzed the features that they rated as 3 out of 5 or higher on relevance to sustainability. Our goal was to filter out features that were selected solely based on liking the product since we were interested in the features that participants perceived as sustainable. We investigated how participants resonated with the features perceived as sustainable based on where the features were placed on the collage. Moreover, we investigated the relationship between the two axes of the collage, perceived sustainability and likeability of a product, based on the placement of the products on the collage.

4.4.2.2.2 Positive Features Perceived as Sustainable and Features not Perceived as Sustainable

We chose to test hypothesis 1 with a more challenging set of features to validate our findings. These features are more challenging because they are closer in sentiment. We tested how participants evaluate products on the collage when provided with 10 positive features perceived as sustainable and 10 features not perceived as sustainable. We opted to test these features on the environmental aspect of sustainability. For the 10 positive features, we used the same list as before for positive features of environmental sustainability in Table 4.1. For the 10 features not perceived as sustainable, we derived a list of features using data from our previous study where we asked participants to annotate parts of reviews that were relevant to sustainability [45].

We assumed that the unannotated parts of the reviews are not perceived as sustainable, combined them into one text, and identified adjectives and noun phrases from them using a part-of-speech tagger. We then randomly matched 10 adjectives to 10 noun phrases to generate the features in Table 4.4.

Table 4.4: Features not related to sustainability

Used for Environmental Aspects Only
Course ground coffee
Typical French press
Wonderful beverage
Single use
Daily coffee
Black product
Tight part
Light weight
Bitter coffee
French way
Moisture problem

Since this approach aims to identify features perceived as not sustainable, some of the features may contribute to engineered sustainability. For example, “single use” might contribute to engineered sustainability but was identified to not be perceived as sustainable using our approach. Moreover, our automated approach of identifying the features may result in noise. Our primary goal here was to automate how we select features perceived as not sustainable to limit potential biases from ourselves selecting the features.

4.4.3 Post-Survey

After completing the collage activity participants were directed to a post-survey where we asked them to rate the quality of product images, product descriptions, and

the overall quality of the products based on the respective Amazon product pages. We then asked participants about their purchasing behavior on Amazon to check if they are target customers who buy home and kitchen items on Amazon. Finally, we asked participants basic demographic questions to check for any other significant variables.

4.4.4 Participants

We recruited a total of 1200 participants from MTurk to complete the activity which took 20 minutes on average; participants were compensated $5 each for their time. We opted to recruit from MTurk over in-person participants so that we could quickly collect many responses. Moreover, the demographics of respondents on MTurk align closely with the online population [34], and therefore better fits a target Amazon customer. This is ideal for our study since participants were likely familiar with Amazon and comfortable to interact with the Amazon popup pages.

To ensure quality in the responses, we screened for participants on MTurk that have at least a 97% prior approval rating and are based in the United States. We set these as requirements on the MTurk platform and then validated participant location using screening questions in the survey. In accordance with literature, respondents in the United States consistently deliver better quality responses [33]. We also conducted the activities during weekday mornings Pacific Time as this was reported to help improve data quality [34]. Furthermore, the collage interface was compatible only with desktop devices with screens larger than 10”. We therefore screened participants for eligible devices. Participants self-reported their screen size.

Out of the 1200 participants that completed their task, we approved 935 based on two requirements: (1) completing the activity in time (t) that is within 1 standard deviation (σ) of the average time to complete the activity (μ) or longer (i.e. $t \geqslant \mu - \sigma$) and (2) correctly answering the check question, "Which sustainability criteria were you evaluating for?" which we asked in the post-survey. The first criteria aimed to filter for responses from participants that were going with their gut feeling rather than justifying their responses, and the second criteria served as an attention check. We excluded responses from the results if they did not meet one or both criteria. These requirements are similar to the ones used to approve responses in our previous study [45].

4.5 Analysis and Results

This section is split into three parts: in the first we analyze the participant demographic pool, in the second we analyze the placement of the features associated with testing hypothesis 1 and in the third we analyze the placement of the products associated with testing hypothesis 2.

4.5.1 Participant Demographics

Our demographic pool of 935 participants includes a broad representation of age groups, education levels, and incomes, with the gender distribution slightly skewed towards more male than female (Fig. 4.10). Most participants were young, white, educated and employed, with many having above-average incomes. This is in line with a prior demographics study of MTurk [33]. While the distributions are not representative of the general population, it is representative of the online population.

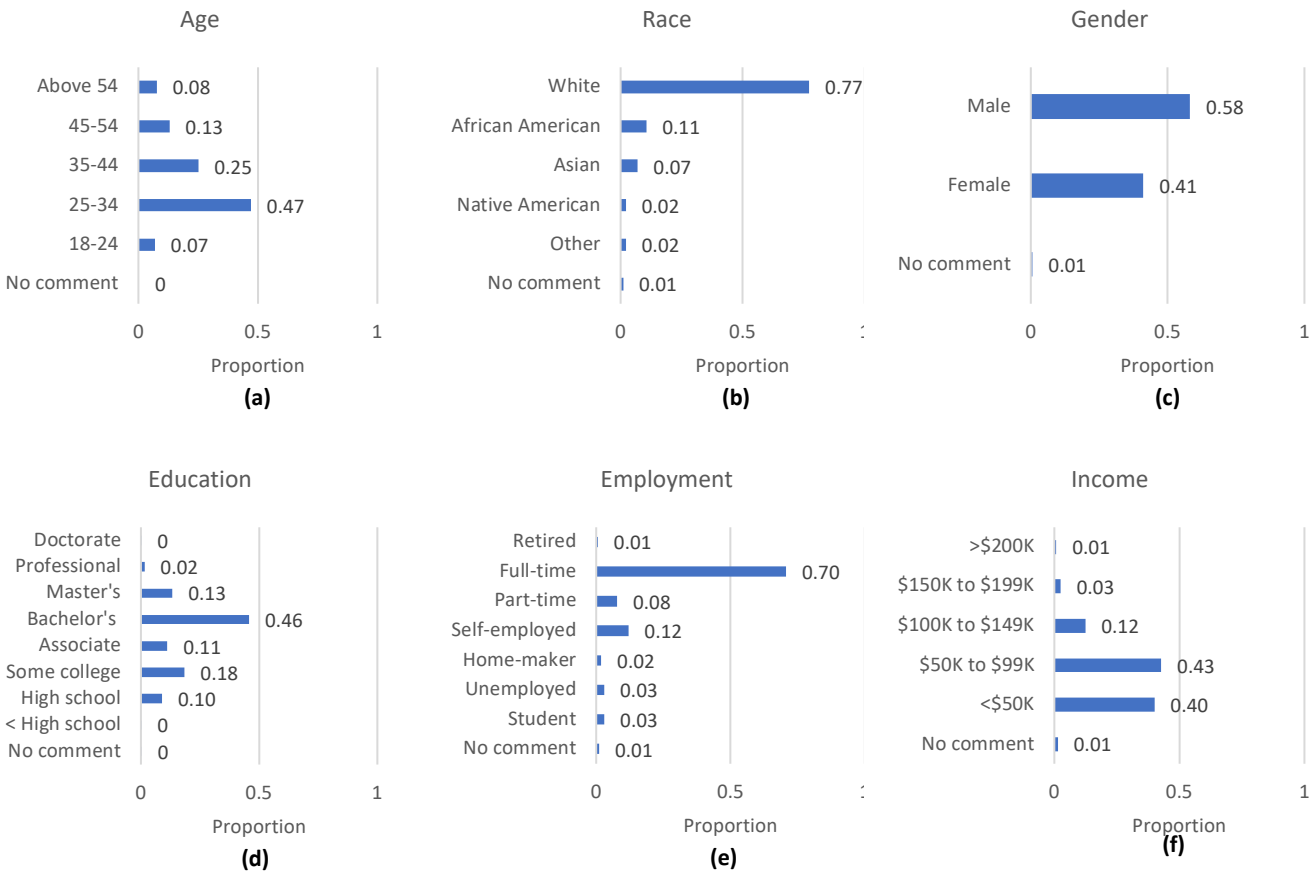

Figure 4.10: Participant demographics

The participants' purchasing habits show that over 90% of them have shopped on Amazon within the past year and that the majority are subscribed to Amazon Prime (Fig. 4.11). This indicates that participants are familiar with Amazon's user interface and were comfortable interacting with the Amazon popups in the study.

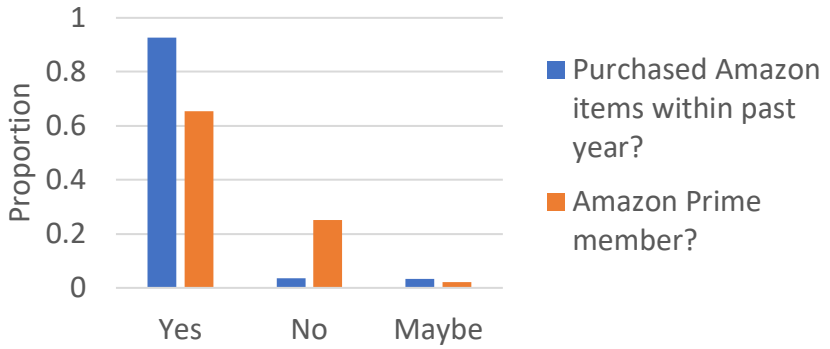

Figure 4.11: Distribution of participants that are Amazon customers

In addition, most participants purchase home and kitchen items from Amazon monthly or more (Fig. 4.12). This is ideal for our study because it indicates that participants can provide evaluations that are similar to a potential French press customer.

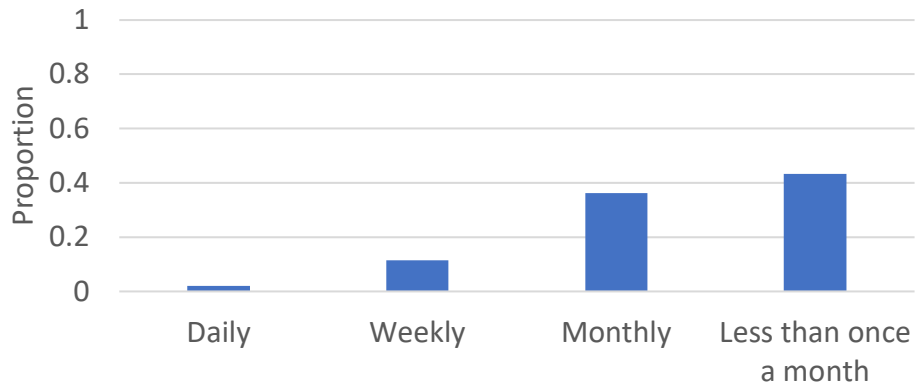

Figure 4.12: Distribution of participant purchase frequency from Amazon's home and kitchen department

4.5.2 Feature Analysis

Below we present our analysis for testing hypothesis 1 based on the placement of features in the collage task. Each of the 935 participants placed multiple features on the collage. After filtering for features that were rated less than 3 out of 5 as relevant to sustainability, we had a total of 7263 location points of features on the collage. We excluded an additional 373 data points that were not moved from their starting location when the collage activity launched (starting locations are shown in Fig. 4.6).

4.5.2.1 Positive and Negative Features Perceived as Sustainable

In this section we present the results and analyses for testing hypothesis 1 using positive and negative features perceived as sustainable. This hypothesis considers how

the perceived features resonate with participants when evaluating product sustainability. Table 4.5 summarizes the information on the features selected.

Table 4.5: Summary of features selected in collage

	Social Sustainability		Environmental Sustainability		Economic Sustainability		Combined	
	Positive Features	**Negative Features**	**Positive Features**	**Negative Features**	**Positive Features**	**Negative Features**	**Positive Features**	**Negative Features**
Number of participants	253		268		241		762	
Observations	1073	439	1328	613	1165	753	3566	1805
Average features per participant	4.24	1.74	4.96	2.29	4.83	3.12	4.68	2.37
Average features per product	178.83	73.17	221.33	102.17	194.17	125.5	594.33	300.83
Average features per product per participant	0.71	0.29	0.83	0.38	0.81	0.52	0.78	0.39
Most common feature selected	Good for my family	Looks flimsy	Well made	Too much plastic	Reasonable price	Too expens-ive	Reasonable price	Too much plastic

Participants selected positive features for products more often than negative in all activity versions. This was most apparent with social sustainability, followed by environmental sustainability, and then economic sustainability, and shows how participants resonate differently with features based on the sustainability criteria. The most used positive feature across all three sustainability criteria was “Reasonable price” while the most commonly used negative feature was “too much plastic”. This illustrates the features that resonated most with the participants and aligns with our previous findings on extracted French Press features perceived as sustainable from online reviews [45]. For social sustainability, “looks flimsy” is not intuitively relevant but is likely related to being perceived as unsafe for use.

To visualize the data, we plotted the average placement of features by the participants in the collage tasks. Figures 4.13 – 4.15 show the results for social, environmental, and economic criteria, respectively. Each plot is color-coded to differentiate between positive and negative features. In each of the figures we see distinct clusters between the average placement of positive features (green-shaded) and negative features (red-shaded) along the axes of the collage. The distinct clusters across the y-axis (measure of perceived sustainability) support hypothesis 1. Figure 4.15 for economic sustainability shows some overlap between positive and negative clusters which is likely attributed to machine learning model noise output from our previous study [45]. The machine learning model for economic sustainability had the weakest performance due to an imbalance between positive and negative annotations. Figure 4.16 shows the average placement of all features combined from the three sustainability criteria. We again see distinct clusters between positive and negative perceived sustainability features with a slight overlap attributed to the economic sustainability results. It is important to note that there is variance across both the x and y axes. While the location of features is represented as dots, the placement of products on the collage occupies a larger space and can introduce additional variance.

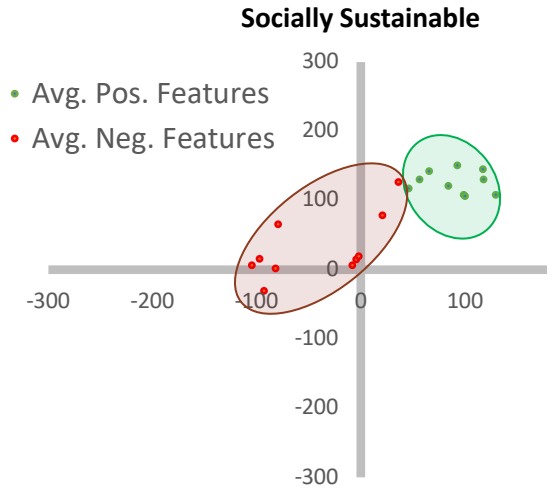

Figure 4.13: Average placement of positive and negative features perceived as socially sustainable on collage

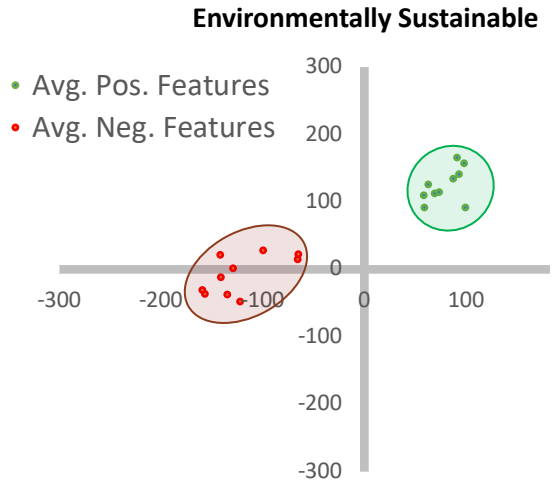

Figure 4.14: Average placement of positive and negative features perceived as environmentally sustainable on collage

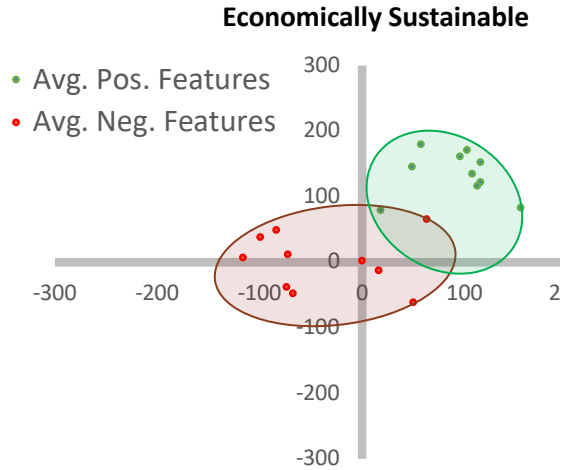

Figure 4.15: Average placement of positive and negative features perceived as economically sustainable on collage

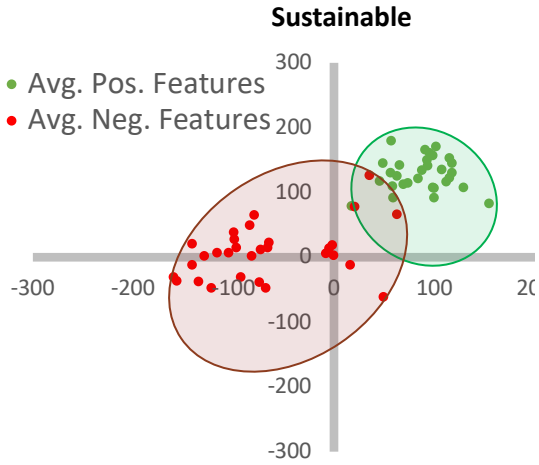

Figure 4.16: Average placement of positive and negative features perceived as sustainable for all criteria on collage

To test if each of the two groups of features are statistically different, we conducted a two-sample t-test assuming unequal variances using the y-coordinate values of features for each sustainability criteria. We assumed unequal variance based on Levene's test showing that the variances of the positive and negative feature locations are statistically different in each of the collage activities. Table 4.6 show the

results for social, environmental, economic, and combined criteria, respectively. In all cases we see that there is a significant difference along the y-axis (measure of perceived sustainability) where positive features are placed higher than negative features, supporting hypothesis 1.

Table 4.6: Two-sample t-test between positive and negative features perceived as sustainable

*: significant at p = 0.05, **: significant at p = 0.01, ***: significant at p = 0.001

	Social Sustainability		**Environmental Sustainability**		**Economic Sustainability**		**Combined**	
	Positive Features	**Negative Features**	**Positive Features**	**Negative Features**	**Positive Features**	**Negative Features**	**Positive Features**	**Negative Features**
Mean Y-Coordinate	129.02	24.65	132.08	-13.16	144.84	4.91	135.33	3.57
Observations	1073	439	1328	613	1165	753	3566	1805
P(T<=t) one-tail	<0.001***		<0.001***		<0.001***		<0.001***	
t Critical one-tail	1.65		1.65		1.65		1.65	

We further investigated if the statistical significance holds when considering repeated measures from participants. We conducted a multivariate analysis of variance (MANOVA) using the x and y coordinates of the features as the dependent variables and the participant demographics as the independent variables. Based on scatter plots showing non-linear patterns in our data, we chose to work with the Pillai criterion because it is the most powerful and robust statistic when assumptions of linearity and homogeneity of variances are not met [73]. The results are shown in Table 4.7 for social, environmental, economic, and combined criteria, respectively. In each case we see that the phrases are highly significant. We can therefore state with statistical significance that participants more often selected features perceived as sustainable when placing products higher on the sustainability axis (i.e., fail to reject hypothesis 1).

Table 4.7: MANOVA output with positive and negative features perceived as sustainable
*: significant at p = 0.05, **: significant at p = 0.01, ***: significant at p = 0.001

	Social Sustainability			**Environmental Sustainability**			**Economic Sustainability**			**Combined**		
	Pillai	~F	**Pr(>F)**	Pillai	~F	**Pr(>F)**	Pillai	~F	**Pr(>F)**	Pillai	~F	**Pr(>F)**
Product	0.095	12.02	<0.001***	0.357	82.1	<0.001***	0.173	28.18	<0.001***	0.155	78.0	<0.001***
Criteria	-	-	-	-	-	-	-	-	-	0.004	5.18	<0.001**
Feature	0.348	13.30	<0.001***	0.293	17.0	<0.001***	0.488	25.29	<0.001***	0.398	21.40	<0.001***
Age	0.007	0.85	0.581	0.003	0.58	0.832	0.004	0.65	0.768	0.001	0.54	0.869
Race	0.012	1.51	0.128	0.011	2.05	0.025*	0.008	1.13	0.338	0.003	1.44	0.156
Gender	0.005	1.57	0.179	0.001	0.44	0.779	0.006	2.31	0.056	0.001	1.32	0.260
Educ.	0.014	1.38	0.170	0.031	5.02	<0.001**	0.018	1.95	0.018*	0.008	2.21	0.004**
Emplo.	0.009	0.79	0.678	0.013	1.77	0.037*	0.009	0.98	0.472	0.004	1.34	0.177
Income	0.012	1.47	0.145	0.006	1.21	0.279	0.008	1.20	0.288	0.002	0.83	0.595

4.5.2.2 Demographic Interactions

In this section we present results related to the demographics of participants. We investigated the demographics as independent variables in the MANOVA analysis to gain more insight into the data. From the MANOVA results in Table 4.7 we see that certain demographics variables are significant for different cases. To understand how the variables influence each other we performed an analysis of variance (ANOVA) on each dependent variable separately. The results are shown in Table 4.8 for social, environmental, and economic, criteria respectively.

Table 4.8: ANOVA output for social, environmental, and economic sustainability
*: significant at p = 0.05, **: significant at p = 0.01, ***: significant at p = 0.001

	Social Sustainability				**Environmental Sustainability**				**Economic Sustainability**			
	Sustainability (y-axis)		Like (x-axis)		Sustainability (y-axis)		Like (x-axis)		Sustainability (y-axis)		Like (x-axis)	
	F	**Pr (>F)**	F	**Pr (>F)**	F	**Pr (>F)**	F	**Pr (>F)**	F	**Pr (>F)**	F	**Pr (>F)**
Product	5.19	<0.001***	21.9	<0.001***	101	<0.001***	112	<0.001***	27.4	<0.001***	28.8	<0.001***
Featre	8.70	<0.001***	27.3	<0.001***	14.3	<0.001***	31.0	<0.001***	24.7	<0.001***	42.0	<0.001***
Age	0.79	0.556	0.71	0.618	0.56	0.730	0.63	0.673	0.68	0.639	0.62	0.685
Race	2.22	0.049*	1.25	0.282	1.49	0.189	2.33	0.041*	0.80	0.553	1.48	0.195
Gendr	1.02	0.361	2.36	0.095	0.85	0.428	0.02	0.982	0.35	0.704	4.23	0.015*
Education	0.85	0.528	1.71	0.114	6.47	<0.001***	3.18	0.004**	1.54	0.160	2.30	0.025*
Emplo	1.02	0.416	0.53	0.810	2.08	0.043*	1.78	0.087	1.07	0.379	0.86	0.541
Incom	1.51	0.183	1.04	0.393	0.73	0.604	1.62	0.152	2.25	0.047*	0.14	0.982

The ANOVA results reveal how the demographics variables interact with the dependent variables. For social sustainability we see that race has a statistical significance for participants identifying product sustainability. In terms of how much participants like a product, demographics have no significance. For environmental sustainability, education and employment are significant for participants identifying product sustainability. In terms of how much participants like a product, race and education are significant. Moving on to economic sustainability, income is significant for participants identifying product sustainability. In terms of how much participants like a product, education and gender are significant.

While the results in Table 4.8 provide preliminary insights on demographic interactions with sustainability perceptions, the participant demographics are not representative. A deeper study on demographics is needed to identify stronger insights.

Table 4.9: ANOVA output for combined sustainability criteria
*: significant at p = 0.05, **: significant at p = 0.01, ***: significant at p = 0.001

	Sustainability (y-axis)		**Like (x-axis)**	
	F value	**Pr(>F)**	F value	**Pr(>F)**
Product	39.61	<0.001***	132.63	<0.001***
Criteria	6.91	<0.001***	4.81	0.008**
Feature	20.18	<0.001***	37.38	<0.001***
Age	0.80	0.550	0.24	0.947
Race	0.42	0.836	2.59	0.024*
Gender	0.81	0.446	2.08	0125
Education	1.39	0.194	3.04	0.002**
Employment	0.82	0.573	2.07	0.043*
Income	1.44	0.207	0.27	0.931

Finally, the results for when we combine the data from all three sustainability criteria are shown in Table 4.9. None of the demographics variables are significant for participants identifying product sustainability, although the criteria variable has a strong

significance. Race, education, and employment are significant variables for participants liking a product.

4.5.2.3 Positive Features Perceived as Sustainable and Features Not Perceived as Sustainable

In this section we present results for the collage activity with a more challenging set of features, including features perceived as sustainable and features not perceived as sustainable. These sets of features are closer in sentiment. Figure 4.17 shows the scatterplot for the average placement of features, color-coded by positive features and features not perceived as sustainable.

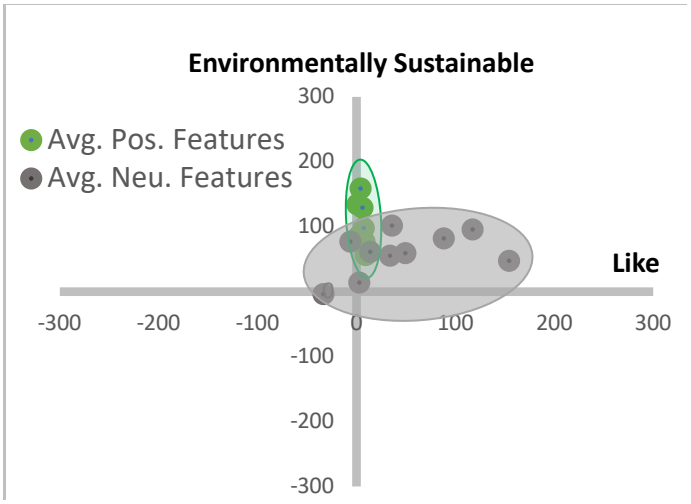

Figure 4.17: Average placement of positive features perceived as sustainable and features not perceived as sustainable

The distinct clusters are less prominent along the y-axis (measure of perceived sustainability), and we also see less of a horizontal spread in the average positive features likely due to the list of features having a closer range of sentiment in this activity. Table 4.10 shows the results for the two-sample t-test assuming unequal variances for the y-coordinate values features. We see that there is a significant

difference along the y-axis (measure of perceived sustainability) between the two groups of features, which supports our initial findings with hypothesis 1.

Table 4.10: Two-sample t-test between positive features perceived as environmentally sustainable and features not perceived as sustainable
*: significant at p = 0.05, **: significant at p = 0.01, ***: significant at p = 0.001

	Positive Features	**Features not perceived as sustainable**
Mean	100.00	43.46
Observations	1652	901
Number of participants	262	
Average features per participant	6.31	3.44
Average features per product	275.33	150.17
Average features per product per participant	1.05	0.57
P(T<=t) one-tail	<0.001***	
t Critical one-tail	1.65	
P(T<=t) two-tail	<0.001***	
t Critical two-tail	1.96	

We conducted a MANOVA shown in Table 4.11 and the results confirmed that features are highly significant. Therefore, even with the more challenging set of features we found that participants evaluating product sustainability on a collage selected features perceived as sustainable for products that they placed higher on the sustainability axis of collage (i.e., fail to reject hypothesis 1).

Table 4.11: MANOVA output with positive features perceived as sustainable and features not perceived as sustainable
*: significant at p = 0.05, **: significant at p = 0.01, ***: significant at p = 0.001

	Pillai	~F	num Df	den Df	Pr(>F)
Product	0.245	60.24	10	4318	<0.001***
Feature	0.083	4.92	38	4318	<0.001***
Age	0.010	2.10	10	4318	0.021*
Race	0.008	1.78	10	4318	0.059
Gender	0.004	2.39	4	4318	0.049*
Education	0.017	2.59	14	4318	<0.001***
Employment	0.013	2.01	14	4318	0.014*
Income	0.024	5.24	10	4318	<0.001***

4.5.3 Product Analysis

In this section we present the results and analyses for testing hypothesis 2. The hypothesis considers how perceived sustainability of a product and liking that product are related. Based on the MANOVA results in Table 4.7, we see that there is a significant difference in products across the y-axis (measure of perceived sustainability) and x-axis (measure of how much the product is liked) in each of the sustainability versions, which supports hypothesis 2. To investigate this further we ran a multiple linear regression model with the Like values of the product placement as the dependent variable and the Sustainability values of product placement and demographics as the independent variables. Table 4.12 shows the p-values from this model for each sustainability aspect; the Sustainability values were significant while the demographic variables were not.

Table 4.12: Multiple linear regression for liking the product versus perceived sustainability and demographics

*: significant at p = 0.05, **: significant at p = 0.01, ***: significant at p = 0.001

	Social	Environmental	Economic	Combined
Sustainability	<0.001***	<0.001***	<0.001***	<0.001***
Age	0.27	0.67	0.41	0.41
Race	0.69	0.28	0.39	0.36
Gender	0.58	0.81	0.12	0.46
Education	0.26	0.21	0.43	0.21
Employment	0.47	0.36	0.60	0.50
Income	0.38	0.47	0.47	0.43

We wanted to study the relationship between perceived sustainability and liking a product further, so we measured the correlation between the Sustainability and Like values of the product placements for the different sustainability aspects. We used a repeated measures correlation to determine the relationship between perceived

sustainability and likeability while controlling for between-participant variance [74]. We chose this measure because it considers that multiple data points on the collage can be attributed to the same participant. The results are shown in Table 4.13. The correlations range from 0.24 to 0.38. These correlations are low despite being significant, suggesting that perceived sustainability plays a small role in likeability. Based on the correlations, perceived sustainability of a product accounts for 24% to 38% to liking a product. The p-values are highly significant; therefore, a statistically significant relationship exists between the placement of a product on the sustainability and like axes of the collage (i.e., we fail to reject hypothesis 2).

Table 4.13: Repeated measures correlation between perceived sustainability of a product and liking the product

*: significant at p = 0.05, **: significant at p = 0.01, ***: significant at p = 0.001

	Social	Environmental	Economic	Combined
Repeated Measure Correlation	0.28	0.38	0.24	0.31
P-value	<0.001***	<0.001***	<0.001***	<0.001***

4.6 Discussion and Limitations

The patterns of evaluating product sustainability reveal essential insights for sustainable design. While it is important for designers to meet engineered sustainability criteria in products, the products need to also resonate with their intended customers. Therefore, designers must address both the perceptions and the engineered challenges of sustainability. Here, we detail the value of using features perceived as sustainable to resonate with customers and the effectiveness of the collage as an evaluation tool to validate sustainability perceptions.

First, we found that participants chose features perceived as sustainable over features that are not perceived as sustainable to describe products they identify as sustainable (Tables 4.7 and 4.11). To reiterate, perceived features used in this study may or may not contribute to engineered sustainability, yet they resonated with users as sustainable. Participants used positive features perceived as sustainable to describe products they identified as more sustainable and negative features to describe products they identified as less sustainable. Therefore, while perceived features might be different from engineered sustainability, they may also lead customers to learn more accurate information about a product.

Second, we measured a significant (yet low) correlation between participants liking a product and how sustainable they think the product is (Table 4.13). This relationship suggests that perceived sustainability plays a small role in how customers like a product, among other factors. When looking at sustainability as a whole, about 31% of why participants liked a product can be attributed to how sustainable they identified the product to be. The correlation was highest with environmental sustainability at 38% while it was lowest for economic sustainability at 24%. The low correlations indicate that likeability and sustainability of a product can be measured separately and demonstrate the effectiveness of the collage as a tool in this context. This contrasts with Ting et al.'s study on smart products where other attributes such as user delight and comfort had correlations of above 70% with likeability [72].

Third, we found that demographics can be a significant factor in how participants identify aspects of sustainability in a product (Table 4.8). For environmental

sustainability we found that education and employment had significant effects on how participants view sustainability. This suggests that education is an important factor in how participants perceive sustainability. For economic sustainability, income had a significant effect, while for social sustainability, race had a somewhat significant effect. These are both intuitive findings since race is an important social factor while income is an important economic factor. The effect of feature perceptions can therefore be enhanced by personalizing different sustainability aspects based on the target customer demographic. Interestingly, when looking at sustainability holistically we found that demographics did not have any significant effect (Table 4.9). This suggests that the effects from specific sustainability aspects average out when combined. These results reveal potentially impactful insights, but it is important to note that our participant demographic was skewed and that a deeper analysis is needed for meaningful conclusions on demographic interactions with sustainability perceptions. For example, gender or age may also have a significant effect on how participants perceive social sustainability although our results did not reflect this.

Our findings have direct real-world implications for designers. First, designers can use the collage method to validate features perceived as sustainable from online reviews. This enables sustainable designers to confidently consider perceived sustainable features in their products. Second, we showed that features perceived as sustainable that are extracted from online reviews resonate with participants as sustainable despite the features not contributing to engineered sustainability. Designers should therefore combine perceived and engineered sustainability features to

differentiate sustainable products with customers. We recommend that designers use the perceived sustainable features to improve how they communicate sustainability to customers in both their existing and next iteration products. For existing products, designers can adapt their product designs to include both features perceived as sustainable and engineered sustainability. For next iteration products, designers can use features perceived as sustainable to guide their design decisions in combination with engineered sustainability tools, for example, life cycle analyses. For example, they can determine the color or texture of the material in their product to align more closely with customer sustainability perceptions while also meeting engineered sustainability criteria. For niche products, we recommend that designers consider the demographics of their customers to refine the perceptions based on the different aspects of sustainability.

These insights have crucial implications for designers but come with a few limitations. First, we did not test for generalizability, therefore the findings might not apply to other products. The features used in this study were extracted from online reviews of French press products [45], and we tested how participants used the features to describe French presses when evaluating sustainability. Second, our participant demographics were skewed which may mean that our demographics findings are not repeatable for evaluating product sustainability. For example, certain demographic variables that we found significant in our study might not turn out significant, or certain variables that were not significant (for example, gender for perceiving social sustainability) may be significant in a repeated study. Therefore, while the results show

that demographics can have a significant role in sustainability perceptions, we recommend a deeper study to confirm how these variables interact with perceived sustainability. Third, while we demonstrated that there is a significant relationship between perceived sustainability and likeability of a product, this does not necessarily indicate purchase behavior. For example, the results showed perceptions of economic sustainability contribute just 24% to why participants like a product, but intuitively we know that price (an economic sustainability factor) plays a large role in a customer purchasing a product. In other words, a sustainable product may resonate with customers, but if the price point is too high it is unlikely that customers will purchase the product. This is supported by literature showing that intent to purchase a product does not equal making a purchase decision [75]. Other factors might also influence real online purchasing decisions that are not assessed here, for example the aesthetics of the product images or review ratings. We therefore recommend conducting an in-depth study on how features perceived as sustainable can influence purchase decisions of sustainable products.

4.7 Conclusion and Future Work

While perceived sustainability features may or may not contribute to engineered sustainability, this study validates that perceived sustainability features extracted from online reviews can help customers resonate more with product sustainability than features not related to sustainability. We designed a set of collage activities for participants to evaluate French press products on the three aspects of sustainability: social, environmental, and economic, and on how they like the products. We provided a

list of features to describe the products including positive and negative features perceived as sustainable as well as features not perceived as sustainable. The features used in this study were extracted from online reviews of French press products in a previous study using annotations and a natural language processing algorithm [45].

Our findings point to important directions for sustainable design. First, designers can more effectively communicate product sustainability to customers using features perceived as sustainable, even if the features may not contribute directly to engineered sustainability. Participants placed features perceived as sustainable higher on the perceived sustainability axis of the collage than features not related to sustainability. Second, we measured a significant (yet low) correlation between the collage axes, perceived sustainability and likeability. This demonstrates that perceived sustainability plays a small role in liking a product, and the low correlation demonstrates that perceived sustainability and likeability can be measured separately. Moreover, it highlights the value of the collage approach for validating customer perceptions. Third, designers can use demographics to identify relevant feature perceptions in niche markets, however our findings suggest that feature perceptions can be generalized across demographics.

For future we recommend closely investigating how perceived and engineered sustainable features interact with each other, the role of demographics on perceived sustainability, and how these factors can influence purchasing decisions to drive purchases for sustainable products.

5. CHAPTER 5
DIFFERENTIATING ONLINE PRODUCTS USING CUSTOMER PERCEPTIONS OF SUSTAINABILITY

Abstract

Customers make quick judgments when shopping online based on how they perceive product design features. These features can be visual such as material or can be descriptive such as “nice gift”. Relying on feature perceptions can save customers time but can also mislead them to make uninformed purchase decisions, for example, related to sustainability. In a previous study we developed a method to extract product design features perceived as sustainable from Amazon reviews, identifying that customer perceptions of product sustainability may differ from engineered sustainability. We previously crowdsourced annotations of French press reviews and used a natural language processing algorithm to extract the features. While these features may not contribute to engineered sustainability, customers identify the features as sustainable and enables them to make informed purchase decisions. In this study, we validate how our previously developed method can generalize by testing it with electric scooters and baby glass bottles. We first extracted features perceived as sustainable for both products and second, tested how participants interpret the features using a novel collage approach. Participants placed products on a set of two axes and selected features from a list. Based on our results we confirm that our proposed method is effective for identifying features perceived as sustainable, and that it can generalize for different products with limitations. We found that positively biased Amazon reviews can limit the natural language processing performance. We

recommend that designers carefully select products with balanced Amazon ratings and use our method to enable customers making informed purchasing decisions.

5.1 Introduction

The growth of e-commerce has changed the way customers make purchasing decisions. With an abundance of products available, customers rely on perceptions to make quick judgments between options. These perceptions are derived from prior experiences and available information, acting as mental shortcuts for customers to simplify decision making [3]. For example, customers may perceive a certain car model to be high quality because it was featured in a movie rather than because of its technical capabilities.

Gao et al. describe this process as “unconscious thought” and demonstrated that users can be more satisfied with their decision when relying on it [76]. The authors compared movie choice ratings between participants where one group was asked to carefully evaluate and justify their movie choice while another group was not. They found that participants who made decisions based on unconscious thought were more satisfied. Similarly, Zhang et al. describe this process using the term “heuristics” [77]. The authors demonstrated how perceived credibility and quantity of online reviews are important heuristics that influence customer purchasing decisions. The authors surveyed users of Dianping.com, a popular review site in China, and evaluated their ratings of informativeness and persuasiveness of reviews, among other factors.

While relying on perceptions can help customers simplify decisions, it can also mislead customers to make uninformed decisions. For example, Kordzadeh found that

physician ratings on healthcare websites were positively biased compared with third-party ratings [78]. Positive bias can lead new patients to form false impressions and make counterintuitive decisions. The author attributed the positive bias to long-term relationships formed between physicians and patients, among other factors, and recommended that healthcare providers account for this bias on their websites. Similarly, Lu et al. investigated customer browsing history and purchasing decisions with a large online travel agency in China [79]. The authors collected 10,000 clickstream and transaction data, finding that the complexity of information shown can help improve purchase probability up to a point. The same product in different contexts can therefore lead to different purchase decisions, making it crucial for sellers to understand the influence of customer perceptions on decision making.

A robust literature exists on customer perceptions of online website content, such as reviews and quantity of information (see section 5.2 for an overview). Limited research exists, however, on customer perceptions of product features. Product features can be visual, such as the shape of a product, or descriptive, for example, “nice gift”, and can communicate certain aspects about a product. For example, MacDonald et al. show how customers perceive the absorbency of paper towels based on the presence of quilted lines [65]. Simple design features can help communicate information about a product that can influence perceptions and decision making.

It is important to note that there can be a gap between customer perceptions of a product and engineered requirements. For example, customers may perceive a product to be expensive because it has shiny colors even though the materials chosen are

inexpensive. This gap is often seen with sustainable products where features perceived as sustainable may not contribute to engineered sustainability. Despite market research showing customers are willing to pay more for sustainable products [7], market success is limited because designers focus on engineered sustainability requirements while neglecting perceived sustainability [38]. Moreover, customers have grown skeptical of green marketing strategies, for example, eco-labels [80].

An alternative approach to create sustainable products is to design-in features perceived as sustainable by the customer. Such features are perceived by customers as sustainable but may not contribute to engineered sustainability. She and Macdonald demonstrate how perceived sustainable features led participants to prioritize engineered sustainability concerns in a decision scenario with toasters [2]. For example, an embossed leaf pattern on a toaster led participants to prioritize energy and shipping concerns of the product. While the embossed leaf pattern does not contribute to sustainability, it communicates information to customers that helps bridge the gap between perceived and engineered sustainability. In doing so, customers are better informed to align their intent with their purchase decisions.

Building on She and MacDonald's work, we previously developed a method to identify features perceived sustainable from online reviews using crowdsourced annotations and natural language processing [45] (refer to section 5.2.2.2 for a deeper overview). We used online reviews of French presses to validate the method and demonstrated a gap between features perceived as sustainable and engineered to be sustainable. For example, while energy and water consumption are critical engineered

sustainability requirements, they were not salient features perceived as sustainable. In a subsequent study, we confirmed that participants identified the extracted features as sustainable using a novel collage activity [81]. Participants placed products on a set of axes and selected features from a list. We found that they more often selected features perceived as sustainable when evaluating product sustainability on the collage. The results validate that participants identified the features as sustainable even if the features may not contribute directly to engineered sustainability.

In this study, we test the generalizability of our previous findings by recreating the methods using different products and assessing the similarities in results. Our goal is to provide designers a robust method to identify product feature perceptions from online reviews so that they may differentiate their products and drive purchase decisions. The rest of the paper is organized as follows: Section 5.2 presents a background on the role of customer perceptions in decision making, the research propositions and hypotheses are in Section 5.3, Section 5.4 presents our method, the results and analysis are in Section 5.5, Section 5.6 presents our discussion, and we conclude our paper in Section 5.7.

5.2 Related Work

As more purchases occur online, several papers have explored the changing context in which customers form perceptions, using tools such as machine learning and collage activities to extract perceptions from online reviews. We provide an overview of this work in this section. In addition, we provide details on our previous studies as we build heavily off them in this study.

5.2.1 Customer Perceptions in Online Decision Making

In this section we present literature on how customer perceptions shape online decision making. Wang et al. investigated the impact of online reviews embedded in product descriptions on purchasing decisions [82]. The authors simulated a shopping experience based on Taobao, a Chinese e-commerce website that automatically bundles online review fragments into descriptions for certain products. The authors recruited participants to explore the website while wearing an eye-tracking device and investigated how the participants interacted with pages that had and did not have online reviews in the descriptions. The results showed that product pages with online reviews in descriptions had longer fixation time on average, suggesting these descriptions aligned closer with customer perceptions. To determine the influence of purchase decisions, the authors then collected historical data from Taobao for two products, a shaving gel and an electric shaver. The data included sales, reputation, price, and whether the products had online reviews embedded in the descriptions. Using a hierarchical multiple regression model, the authors found that descriptions embedded in online reviews positively influenced purchase decisions. This finding demonstrates how perceptions of product features from online reviews can drive purchasing decisions.

Maslowska et al. studied the influence of product price and customer perceptions of reviews on online purchase decisions [83]. The authors used shopping data provided by two online retailers, one that sells unique and high-priced items while

the other sells health and beauty products. There were 2.5 – 3 million observations from each retailer. For each observation the authors had access to the number of reviews for a product, the average number of stars, whether the customer clicked on the "review tab", product price, and purchase decision. The authors used a logistic regression model with the purchase decision as a dependent variable and found that the product price plays an important role on how ratings and reviews influence purchase decision. For lower-priced products, average ratings can have a large influence with fewer reviews while for higher-priced products, more reviews are needed for the average rating to have an influence. These findings illustrate how price can influence of how customers perceive product reviews.

Helversen et al. investigated the relationship between customer age and the influence of perceptions of product attributes and reviews on purchasing decisions [84]. The authors designed three between-participant conjoint analysis surveys where they presented pairs of positively rated household products to participants. A mixture of highly positive and negative reviews was shown with a mixture of low and high ratings. The authors found that younger customers relied more on average ratings when product attributes were similar between paired choices, while older customers were quickly influenced by negative reviews. These results show the importance of factoring in age for how customers develop perceptions and make purchasing decisions.

Nysveen and Pederson explored how interactive features such as content personalization and customer communities influence perceptions of customer experience on a shopping website [85]. The authors designed six websites for two

made-up companies, an airline and a restaurant. Each company had one website with email functionality only, a second website with email and personalization features, and the third website with email and customer community features. Participants interacted with the websites and then responded to a survey on the ease of use, usefulness, and attitude towards the websites. The results showed that the interactive features had a moderate influence on perceptions of customer experience and emphasize the effect of the e-commerce platform to influence customer perceptions. This study demonstrates how website content not related to the product may still influence purchasing decisions.

Li et al. investigated how return policies influence customer perceptions of products and decision making depending on the market stage of a business [86]. The authors propose a multi-stage hidden Markov model which models randomly changing systems. They test it on 50,000 purchase records spanning three years from Taobao including returns, discounts, and total sales. The results showed that promotions and return policies had varying influence of repurchase behavior across different stages of market growth. For example, a company in the growth stage could benefit from flexible return policies and frequent promotion while a company in the introduction stage would not. Therefore, role of return policies on customer perceptions is crucial depending on the market stage of the seller.

5.2.2 Extracting Customer Perceptions from E-Commerce Websites

Literature discussed thus far looks at factors such as age, reviews, price, and website features to influence customer perceptions and decision making, but it neglects a key component which is how customers perceive features of the product itself. This

presents an opportunity for designers to determine how certain product features align with customer perceptions and can drive purchasing decisions. The development of e-commerce and social media provides a wealth of information that designers can tap in to online. In this section we present literature on methods to extract customer perceptions from online content including machine learning and collage approaches.

5.2.2.1 Machine Learning Approaches

Zhang et al. study the influence of self-descriptions of Airbnb host on customer trust and how they can influence booking behaviors [87]. The authors annotated 4179 host descriptions from Airbnb listings based on perceived trustworthiness of the hosts. The authors then used a deep learning model to predict perceived host trustworthiness for 75,000 host descriptions. Using this data, they extracted textual features including readability, sentiment intensity, and semantic content. Semantic content included personal information about the host such as family and work. From regression analyses the authors showed that readability of the self-description had a positive influence on perceived trust, while semantic intensity had a U-shaped relationship with trust. Moreover, semantic content had a positive influence on trust if the content was related to sociability. When looking at Airbnb booking decisions, the results showed that higher perceived trust of host led to more booking decisions. These results point to the importance of language when describing products to drive purchasing decisions.

Liu et al. use natural language processing to identify product competitive advantages from social media content [88]. The authors collected reviews of a Volkswagen Passat from two Chinese auto websites and identified competitors from the

reviews based on comparative language. An example review could include: “the sound system in the Passat sounds better than the one in my old Camry”. The authors first preprocessed the reviews by removing stop words and performing named-entity recognition. They then performed a sentiment analysis using logistic regression and a domain specific lexicon to assess customer sentiments towards features of the Volkswagen Passat compared to its competitors. With their method the authors demonstrated how customer perceptions can drive competitor analyses to inform design decisions for next iteration products.

Singh and Tucker develop a machine learning method to classify reviews based on categories that include form, function, behavior, service, and other [22]. The authors demonstrated the method using 900 reviews from three android phones. After preprocessing the reviews, the authors labeled each sentence according to one of the five categories resulting in 5741 labeled examples. The authors then used a decision tree algorithm achieving F1 scores of about 80%. After repeating the method with different product types, the authors achieved similar scores. The method enables the categorization of customer perceptions to enable efficient design and customer decision making.

Sun et al. investigate how perceived informativeness of reviews impact review helpfulness for both search and experience products [89]. The authors used cell phones, televisions, and laptops as search products, skin care, rice cookers, and running shoes as experience products. Based on a total of 13152 reviews from JD.com, the authors constructed a metric of informativeness using information about the reviewer and

review. Using a linear regression, the demonstrate the superiority of their method for predicting review helpfulness as perceived by the user.

5.2.2.2 Combining Machine Learning with Collage Approaches

Previous literature identified methods to extract customer perceptions from online content but did not identify specific product features that can inform design decisions. Moreover, previous literature does not test how users interpret product features in terms of liking and evaluating products. This gap is particularly crucial for sustainable products where designers often focus on engineered requirements while neglecting perceived requirements. Motivated by this gap, we previously conducted two studies where we first developed a natural language processing approach to extract product features perceived as sustainable from online reviews [45], and second developed a novel collage approach to test those features in terms of how users like and evaluate products [81]. We previously selected French presses as a case study. In this study we aim to validate the generalizability of our previous approaches by testing them with different products. We provide details on our previous work below since we build heavily off them for this study.

In the first study, we extracted features perceived as sustainable using crowdsourced annotations of online reviews and a natural language processing algorithm [45]. The approach combined research from identifying sustainability perceptions, rating design ideas, and natural language processing (Fig. 5.1) and is outlined in four steps (Fig. 5.2). We collected product reviews for a target product type from Amazon, annotated the reviews using a crowdsourcing platform based on criteria

related to the perceptions, modeled the reviews and annotations using natural language processing, and extracted features perceived as sustainable from the model.

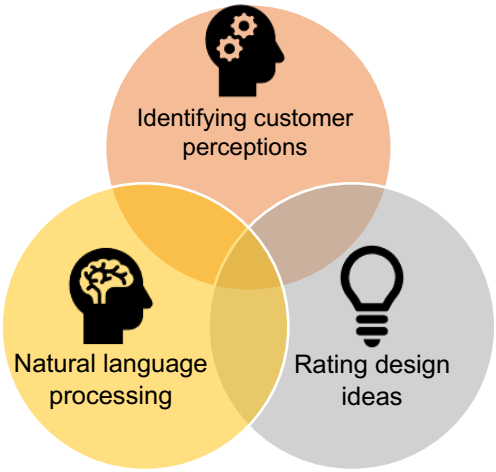

Figure 5.1: Interdisciplinary method flow

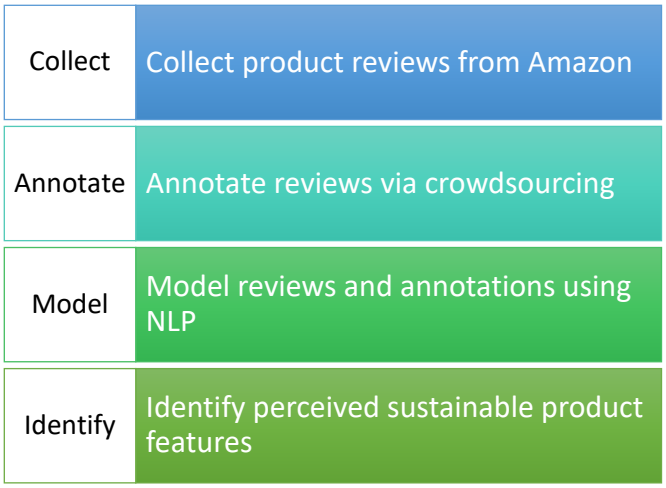

Figure 5.2: Extracting customer perceptions approach

We tested the method with 1474 reviews of French presses from Amazon and recruited 900 respondents from Amazon Mechanical Turk to annotate the reviews based on the three sustainability pillars: social, environmental, and economic. For a product to be truly sustainable it needs to account for each pillar. We previously conducted a pilot study that showed participants had more clarity when focusing on one pillar, so we assigned respondents to one of three versions of the survey to focus on each sustainability pillar separately and trained them on their assigned pillar (Fig. 5.3). Respondents then highlighted parts of reviews relevant to their pillar and rated the emotions in their highlights.

Figure 5.3: Sustainability pillar training

We modeled the annotations using a logistic classifier for each sustainability pillar and extracted French press features perceived as sustainable based on the beta parameters of the model. The precision, recall, and F1 scores for the model are shown in Table 5.1 (see section 5.4.1.3 for more on these metrics). With scores ranging from 0.83 to 0.95 for positive sentiment and 0.42 to 0.72 for negative sentiment, we were confident in the model performance while noting possibilities of noise for negative sentiment predictions. Since the noise may lead to false positives and negatives, we subsequently developed a collage approach to test the extracted features and validate that participants identify them as sustainable [81].

Table 5.1: Precision, recall and F1 scores for French press features perceived as sustainable

	Social sustainability			**Environmental sustainability**			**Economic sustainability**		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Positive Sentiment	0.85	0.87	0.86	0.83	0.86	0.85	0.85	0.95	0.90
Negative Sentiment	0.70	0.66	0.68	0.51	0.72	0.66	0.53	0.42	0.72

We identified salient positive features based on the largest positive beta parameters in the model and identified salient negative features based on the largest negative beta parameters in the model. For social aspects, positive features were related to family, work, and giving gifts while negative features were related to safety of the product such as the glass breaking. For environmental aspects, positive features were related to the material of the product, such as stainless steel and no plastic, while negative features related to the durability of the product. For economic aspects, features were mostly generic. For example, positive features included good price while negative features included false advertising. We then identified engineered sustainability requirements of a French press using a life cycle analysis and found that crucial engineered sustainability requirements, for example, energy and water consumption, were not salient perceived sustainable features. This demonstrated the gap between engineered and perceived sustainability and the importance for designers to account for both when creating sustainable products.

In the second study, we developed a novel collage approach to test the extracted features perceived as sustainable with users in terms of how they like products and evaluate sustainability [81]. Using the collage, we identified the relationship between features perceived as sustainable and user emotions in an engaging way without drawing attention to the features. We created a webapp collage activity with two axes: sustainability on the vertical axis (customized to one of the three pillars depending on the version of the collage) and likeability on the horizontal axis. An example of a social sustainability collage activity is shown in Fig. 5.4. We recruited 1200 participants from

Amazon Mechanical Turk, assigned them to one of three sustainability pillars, and asked them to evaluate six French press products on a collage. They placed images on the collage according to the two axes and selected product features from a dropdown list. The list included features perceived as sustainable that we extracted previously as well as features "not perceived as sustainable" that we identified for the collage study.

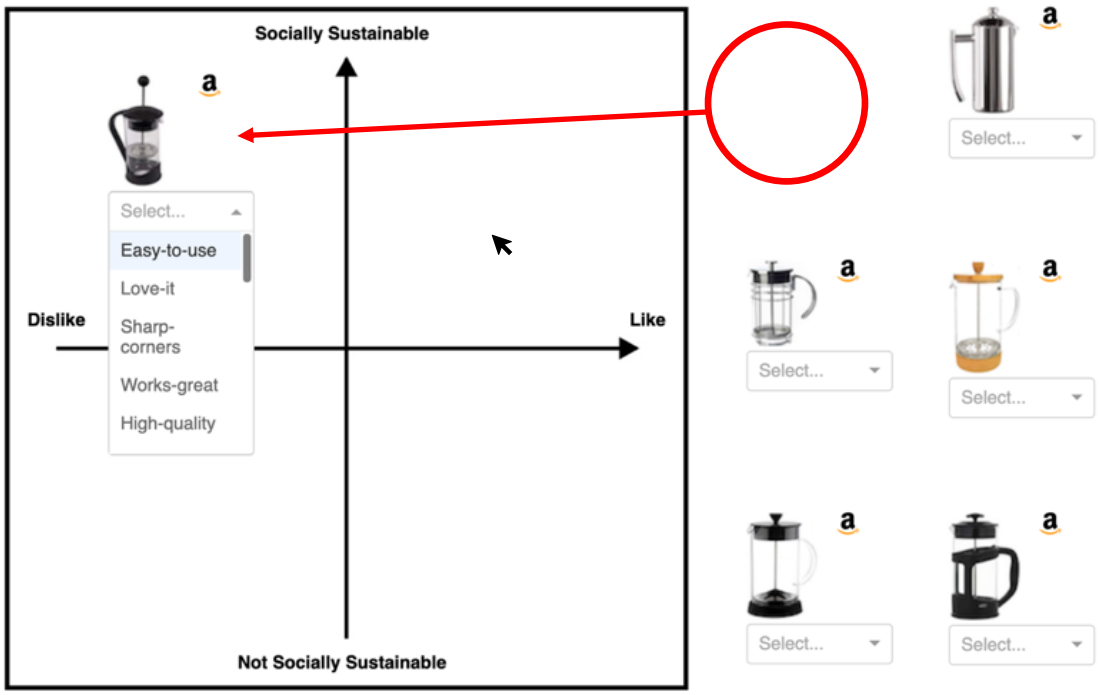

Figure 5.4: Dragging and dropping products on collage and selecting at least one phrase to describe each product. Example taken from a social sustainability collage activity

Based on participants' placement of the products and selection of the features on the collage, we showed that they actively chose features perceived as sustainable (as outlined by the method explained in Section 5.4.2.2) for products that they placed higher on the sustainability axis, indicating that these features stood out to them as sustainable despite not contributing to engineered sustainability. We also found a low correlation between perceived sustainability and likeability, validating that the collage is

an effective approach for measuring these two attributes separately. The results validated our previously extracted features perceived as sustainable as well as validated the collage tool as an effective tool to test features perceived as sustainable with users.

A limitation to the findings is that the approach has been tested on French presses only. We aim to address this limitation in this study by testing the generalizability of our approach across different product types.

5.3 Research Proposition and Hypotheses

This work aims to validate the generalizability of our previously developed approaches for extracting and testing features perceived as sustainable from online reviews. To validate our approaches, we extracted features perceived as sustainable for different products using annotations and a logistic classifier, and then used a collage tool to test the features with users in terms of how they like and evaluate the products. We asked participants to place products along the two axes of the collage, sustainability and likeability, and to label the products using a list of the extracted features from the logistic classifier. In previous studies we tested and confirmed the following propositions and hypotheses using French press products as a case study (Table 5.2).

Note that the hypotheses in Table 5.2 represent alternate hypotheses, contrasting with null hypotheses where no differences are expected to be found between stimuli. The results in Chapter 4 rejected the null hypotheses, therefore supporting the alternative hypotheses. Our goal for this study is to test if the same propositions and hypotheses hold when tested with multiple product types.

Table 5.2: Propositions and hypotheses from our previous studies

Proposition & Hypotheses	Status
P1: Phrases in product reviews perceived as sustainable contain semantic and syntactic characteristics that can be modeled [45]	Supported with French presses (Chapter 2)
P2: Designing-in perceptions can help customers create an alignment between perceived sustainability and sustainable products. Based on this, we propose that customers will evaluate perceived sustainable features as being sustainable [81] H1: participants evaluating product sustainability on a collage will select features perceived as sustainable for products that they place higher on the "sustainability" axis of the collage [81]	Supported with French presses (Chapter 4)
P3: Customers tend to like products that create cognitive alignment for them, and perceptions can help them achieve that. We therefore propose that perceptions of product sustainability contribute to how much customers like a sustainable product [81] H2: A statistically significant relationship exists between the placement of a product on the "sustainability" axis of the collage, and the "like axis of the collage [81]	Supported with French presses (Chapter 4)

5.4 Methods

The method in this paper is based on our work from two previous papers where we used a French press as the focal product [45,81]. In this study we validate how the method generalizes when applied to different product types. First, we extracted features perceived as sustainable from Amazon reviews for different product types, and second tested the features with users in terms of how they like and evaluate products (Fig. 5.5). These steps are explained below.

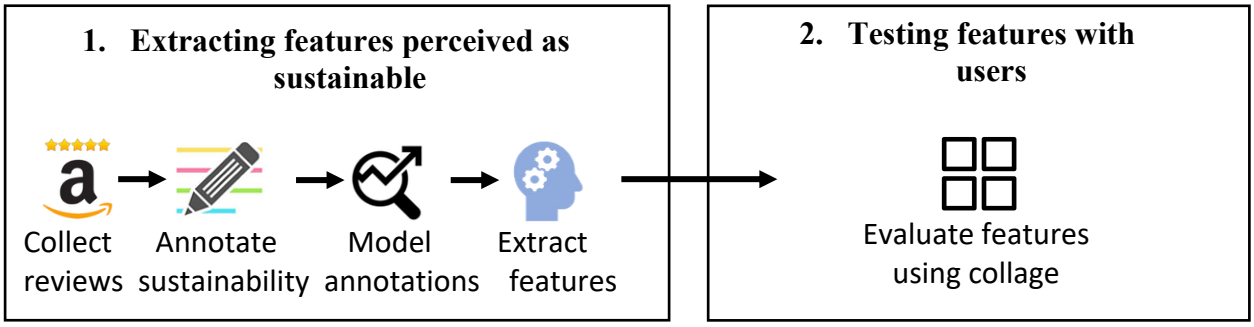

Figure 5.5: Method Overview

5.4.1 Extracting Features Perceived as Sustainable from Online Reviews

The methods outlined in this section aim to test proposition 1. We extracted features perceived as sustainable for electric scooters and baby glass bottles using the four-step process outlined in Fig. 5.2. Each step is explained below.

5.4.1.1 Collecting Reviews

We selected electric scooters and baby glass bottles as the focal products for this study because they (1) are different in design and function from the original French press product, (2) have varying aesthetic design features available, (3) regularly receive several hundred reviews on Amazon, and (4) likely have sustainability-related concerns for customers. We wanted to select products that are different from a French press to effectively evaluate how the method can be generalized. Products such as kettles or other coffee makers would have been too similar. We also selected products that have a large variety of features that reviewers can mention (paper plates, for example, would have been too simple) as well as products that have large amounts of reviews available for us to collect (there were limited reviews for electric bicycles, for example). Finally, since we are interested in extracting features that are perceived as sustainable, we wanted products where sustainability concerns are likely to be prominent in the

reviews. In the case of electric scooters, power consumption and battery life are likely important features for customers and are relevant to environmental sustainability, while for baby glass bottles health and safety are likely priority features which are relevant to social sustainability.

We scraped 1500 Amazon reviews from four electric scooters and 1444 Amazon reviews from eight baby glass bottles. We selected the four products and eight products, respectively, from Amazon so that they (1) have varying aesthetic features from one another, (2) are in a similar price range, (3) have less than 500 reviews per product, and (4) have at least 80% estimated authentic reviews according to a data analytics tool (fakespot.com) for each product. We selected products that have varying features to better test different features with users. Moreover, we selected products in a similar price range so that their quality and capabilities are similar to each other. Each product had less than 500 reviews so that we have a variety of products instead of one product dominating the reviews. The motivation was to have a variety of features to test. Finally, we selected products that are estimated to have a high number of authentic reviews so that we collect real customer opinions and perceptions. All the reviews scraped came from the United States to limit the number of reviews written in a foreign language. Moreover, we filtered reviews that were less than 10 words as they tended to be generic, for example, “this is a great product, I highly recommend it”.

5.4.1.2 Annotating Reviews

We recruited respondents from MTurk (referred to as “annotators”, see section 5.4.1.2.3) to annotate the scraped reviews based on sustainability criteria. These

annotations are then fed into a logistic classifier to extract features perceived as sustainable.

5.4.1.2.1 Survey Design

To guide the annotators, we designed a survey that trains and tests them on the sustainability criteria, and then shows them a set of 15 random reviews to annotate before answering a set of demographic questions (Fig. 5.6).

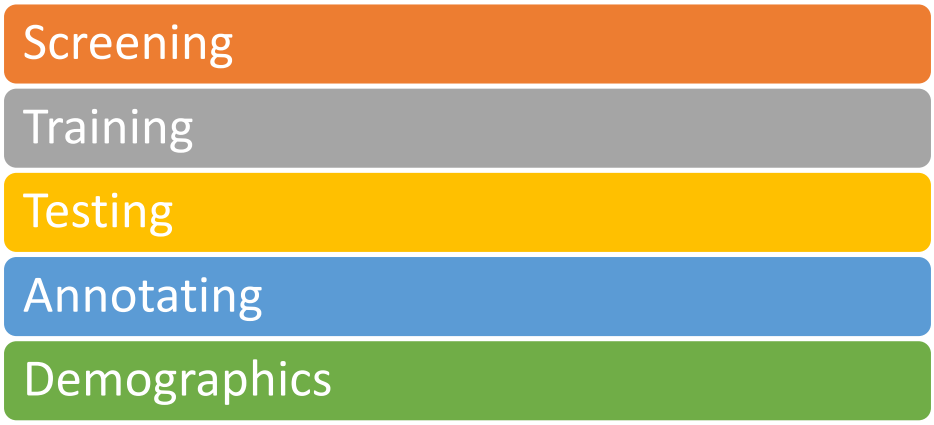

Figure 5.6: Annotation survey process

In total we had six different versions of the survey to account for the three sustainability aspects (social, environmental, and economic) and for each of the two product types (electric scooters and baby glass bottles), shown in Fig. 5.7. We assigned annotators to one of the six surveys so that they focus on one product type and one sustainability criteria. Previous pilot studies showed that this approach led to more clarity for the annotators and provided more usable responses [39].

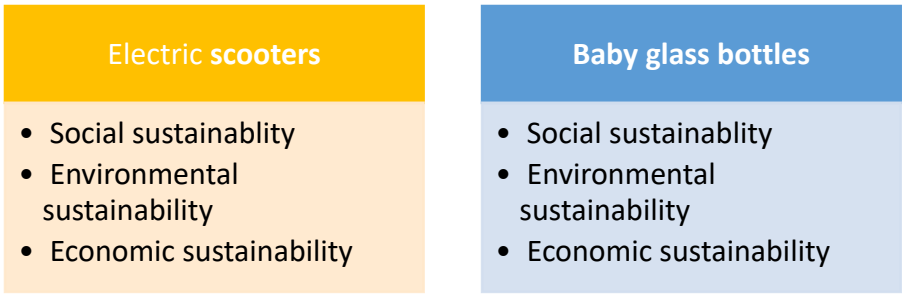

Figure 5.7: Three annotation survey versions per product

In the training portion of the survey, we displayed sustainability criteria to the annotators (Fig. 5.3) and showed them examples of annotated reviews according to their assigned sustainability aspect. We then tested them to confirm that they understood the training. After passing the test annotators began annotating the 15 reviews (see section 5.4.1.2.2) according to one of the sustainability aspects criteria.

5.4.1.2.2 Data Collection

To annotate the reviews scraped in section 5.4.1.1, we stored the reviews on a server so that they can be pulled live during the survey. We used a biased-random algorithm for selecting the reviews from the server to ensure that each review was presented to three different annotators. Having different annotators annotate the same review allowed us to capture different perspectives that can help the performance of the machine learning model [72]. Each participant saw 15 reviews in total, one at a time. For each review, we asked annotators to highlight up to five parts of the review that they found relevant to their assigned sustainability criteria. If they did not find any part of the review relevant, we asked them to highlight the entire review and label it as irrelevant to proceed to the next review. If they highlighted parts of the review as relevant, we asked annotators follow-up questions about each part before proceeding

to the next review. These included asking annotators to type-in the specific feature that is mentioned in the highlighted part, and to label the emotion on a 5-point Likert scale ranging from negative to positive.

5.4.1.2.3 Annotators

We recruited a total of 1800 annotators from Amazon Mechanical Turk to complete one of the six surveys. 900 annotators annotated reviews for electric scooters and 900 annotators annotated reviews for baby glass bottles. Within each product type, 300 annotators annotated the reviews for each of the three sustainability aspects. On average annotators took 20 minutes to complete their survey and we compensated them $4 each. Similar to the justifications used in [39], we recruited Amazon Mechanical Turk respondents instead of in-person annotators as it allowed us to collect many annotations in a short amount of time. Moreover, this online approach is timely due to the COVID-19 pandemic which is when we conducted this experiment.

To ensure high quality responses, we required respondents to have at least a 97% approval rating and to be based in the US. We set these requirements in the MTurk platform and confirmed them using screening questions in the survey. Moreover, we included a simple checkpoint question to gauge if annotators are paying attention. If annotators completed the survey faster than the average time by at least one standard deviation and incorrectly answered the checkpoint question, we assumed their response was low quality and did not include it in the analysis. These criteria are similar to what was used in our previous work [39]. Based on these criteria we approved 1702 out of the 1800 responses.

5.4.1.3 Machine Learning Model

We used a binary logistic classification model to extract features perceived as sustainable from the annotated reviews (represented in Eq. 2.4). We chose a logistic classification model because it has proven to be highly effective for natural language processing applications while remaining interpretable in terms of its beta parameters [90]. This enabled us to extract salient product features directly from the classifier, in contrast to deep learning approaches.

Our input "X" consisted of (1) phrases highlighted as "relevant" to sustainability and (2) product features typed in by the annotators, while the output "Y" is binary representing the emotion in each phrase. We binarized the output Y such that 0 represented negative or neutral emotion while 1 represented positive energy. We opted for a binary output instead of a multi-class model due to the limited explanatory power from our dataset for a multi-class model. The beta fitting parameters were optimized with a maximum likelihood shown in Eq. 2.5.

We first pre-processed the inputs to remove potential noise in the model. This included lowercasing all text, stemming words, removing stop words such as "and" or "is", and removing punctuation. We then processed these inputs so that they can be quantified in a matrix and fed into a classifier. For the highlighted phrases we used bag of words, bigrams, and trigrams. For the typed-in features we summarized them into a set number of "topics" using Latent Dirichlet Allocation (LDA) and then hot-encoded them for each highlighted phrase.

We split the data into a 70% training and 30% test set and implemented the logistic classifier model in Python using the Scikit package. We used five-fold cross validation on the training set and penalty terms to shrink fitting parameters based on Ridge regularization to address potential overfitting from high dimensionality. As an external validity check on the models, we used precision, recall, and F1 (Eqs. 2.6-2.8), respectively. These are often more robust measures than accuracy [90]. Using our binary classification model as an example, the precision of the model predicting a positive emotion would be the number of predicted phrases with actual positive emotion over the total number of phrases. Moreover, the recall of the model predicting a positive emotion would be number of actual positive features predicted over the total number of actual positive features in the dataset. The F1 of the model is a harmonic mean of precision and recall.

5.4.1.4 Extracting Perceptions

We extracted salient product features perceived as sustainable from the machine learning model that drove positive and negative sentiment. The magnitude of the beta parameters indicates the influence of a given feature on the model. The largest positive beta parameters in the model therefore point to the product features perceived as sustainable that drove positive sentiment while the largest negative beta parameters point to the product features perceived as sustainable that drove negative sentiment. These features come from reviews of multiple products to capture a variety of different features. After extracting the product features perceived as sustainable, we conducted a collage experiment to validate that participants identified these features as sustainable.

5.4.2 Testing Perceived Features Extracted from Online Reviews with Participants

The method outlined in this section aims to test hypotheses 1 and 2. We recruited 300 additional respondents (referred to as "participants" for this portion of the method) from MTurk to evaluate the products and features using the collage activity explained in Section 5.2.2.2. Based on the placement of products on the collage and the location of selected features we determined if participants identified the extracted features as sustainable. To guide participants through the activity, we designed three versions of a survey (accounting for each of the sustainability pillars) and assigned participants to one of the versions (see Fig. 5.8). Similar to Section 5.4.1.2.1, we asked participants to evaluate products for only one of the sustainability pillars based on pilot studies that demonstrated this led to more usable responses.

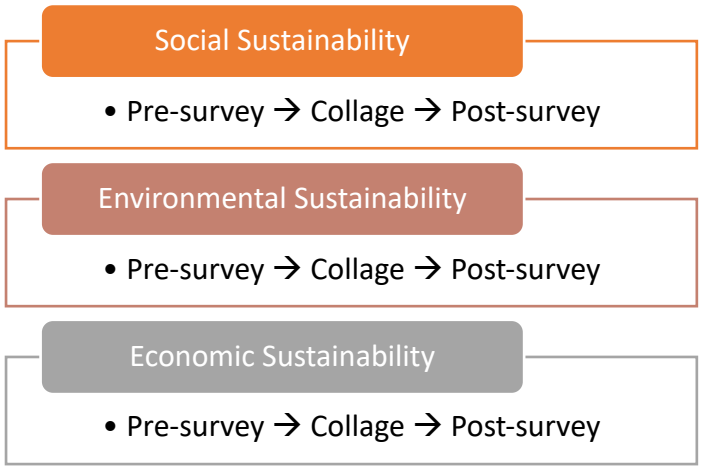

Figure 5.8: Three collage activity versions

5.4.2.1 Pre-survey

In the pre-survey we familiarized participants with their assigned sustainability criteria according to Fig. 5.3, as well as the products that they will evaluate (Table 5.3). We selected the products according to the criteria explained in Section 5.4.1.1.

Table 5.3: Products in Collage Activity

						
Product Name	Gotrax	Razor E300S	Mongoose	Razor EcoSmart	Segway	SKRT

During the pre-survey we trained participants on their assigned sustainability pillar and led them to Amazon pages of the products in Table 5.3 to familiarize themselves before evaluating. They had to open each of the Amazon pages and spend a certain amount of time on them to proceed with the activity. We required this to ensure that participants understood the characteristics of each product before evaluating them. Participants could also access the Amazon pages later when evaluating the products on the collage.

5.4.2.2 Collage Activity

After completing the pre-survey participants accessed a link to a collage webapp using the same interface shown in Fig. 5.4. Products were presented on the right side with buttons to access their Amazon pages for a refresher about each product if needed. On the left was a button to access the sustainability criteria for a given pillar from the pre-survey. The collage consisted of two axes ranging from “Not Sustainable” to “Sustainable” vertically and “Dislike” to “Like” horizontally. The sustainability axis was named social, environmental, or economic depending on the version. Participants dragged and dropped each product on the collage and then selected features from a dropdown menu for each product as shown in Fig. 5.4. We test hypothesis 1 based on the placement of the features on the collage, and tested hypothesis 2 based on the placement of products on the collage.

In the dropdown menu we provided the features we extracted from the machine learning models in Section 5.4.1.4. Each collage version included a list of 20 features that participants could select from. Ten of these features were the most positive salient features from the machine learning model and the other ten were the most negative salient features from the machine learning model. These features are derived from reviews of multiple products but are specific to a certain sustainability pillar. Each sustainability version of the collage had its own set of 20 features. The order of the features was randomized between participants. We present these features in Section 5.5 as part of the results.

To further test hypothesis 1, we conducted a fourth collage activity for environmental sustainability but with a more challenging set of features. These features included ten positive features perceived as sustainable from the original environmental collage activity and ten new features not perceived as sustainable. For the features not perceived as sustainable, we derived phrases from the unhighlighted parts of the annotated reviews collected using the method in Section 5.4.1.2.2. Since they were unhighlighted, we assumed that they were not perceived as sustainable. We combined the unhighlighted parts and identified ten random adjectives and ten nouns using named-entity recognition, and then randomly combined them to create descriptive features. The motivation was to identify these features in a fully automated way and avoid potential bias. This set of features is more challenging because the sentiments are closer together, and we cannot be sure if the perceptions are indeed not perceived as sustainable. The derived features are presented in Section 5.5. For this collage activity,

we recruited an additional 100 participants using the procedures outlined in Section 5.4.2.4.

After evaluating each product, participants rated each feature that they selected based on how relevant to sustainability they think it is using a 5-point Likert scale. We included this in the activity so that we can filter out from the participants' selection the features that they did not select due to sustainability. After rating the features participants completed a post-survey.

5.4.2.3 Post-survey

In the post-survey we asked participants to rate on a 5-point Likert scale the quality of images, product descriptions, and the overall product quality for each of the products they evaluated on the collage. Finally, we asked participants basic demographic questions.

5.4.2.4 Participants

We recruited 300 participants from MTurk to complete the collage activity using the same recruiting criteria as in Section 5.4.1.2.3, in addition to requiring participants to use a screen size of 10 inches or larger. This was to ensure compatibility with the collage interface. Participants self-reported their screen size in the screening question. They completed their task in 21 minutes on average and we compensated them $5 each. We did not analyze responses if they fell under one of the following: (1) participants completed the survey faster than the average time by at least one standard deviation or (2) they incorrectly answered a simple checkpoint question designed to gauge attention. Based on these criteria we analyzed 224 responses out of 300.

5.5 Results

We first present the results that test the generalizability of proposition 1 related to the features extracted from online reviews. Second, we present the results that test the generalizability of hypotheses 1 and 2 based on the placement of products and extracted features on the collage.

5.5.1 Features Perceived as Sustainable

This section presents the extracted features perceived as sustainable for electric scooters and baby glass bottles and tests the generalizability of proposition 1: Phrases in product reviews perceived as sustainable contain semantic and syntactic characteristics that can be modeled.

5.5.1.1 Electric Scooters

The results from the model evaluation and model output for electric scooters are shown below.

5.5.1.1.1 Model Evaluation

The precision, recall, and F1 scores for each of the sustainability pillars for electric scooters are shown in Table 5.4.

Table 5.4: Precision, recall and F1 scores for electric scooter features perceived as sustainable

	Social sustainability			**Environmental sustainability**			**Economic sustainability**		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Positive Sentiment	0.85	0.80	0.82	0.86	0.85	0.85	0.80	0.97	0.88
Negative Sentiment	0.33	0.41	0.36	0.51	0.52	0.51	0.60	0.16	0.25

The positive sentiment ranged between 0.80 and 0.97 across all three metrics and all three sustainability pillars, indicating that we can have high confidence in the quality of the model output for positive sentiment. The negative sentiment had a lower range however between 0.16 and 0.52, indicating that we are likely to see some level of noise in the model output and is important to keep in mind while analyzing the most salient negative features. These metrics are similar to the findings from our previous study with French presses (see table 5.1) which support the generalizability of proposition 1, although the negative sentiment scores fared worse with electric scooters here suggesting that there is a greater imbalance between positive and negative highlighted reviews.

5.5.1.1.2 Model Output

Figures 5.9-5.11 show the most salient 20 positive and negative features of electric scooters based on the parameters of the logistic classifier for social, environmental, and economic pillars, respectively. These features are derived from reviews of multiple products and have the largest positive and negative parameters in the model, indicating that they are the most salient features that annotators identified as sustainable. Note that the features shown in this graph are stemmed as part of preprocessing, which is why words such as “warranty” appear as “warranti”. The models were able to output specific product features perceived as sustainable for electric scooters, therefore supporting the generalizability of proposition 1.

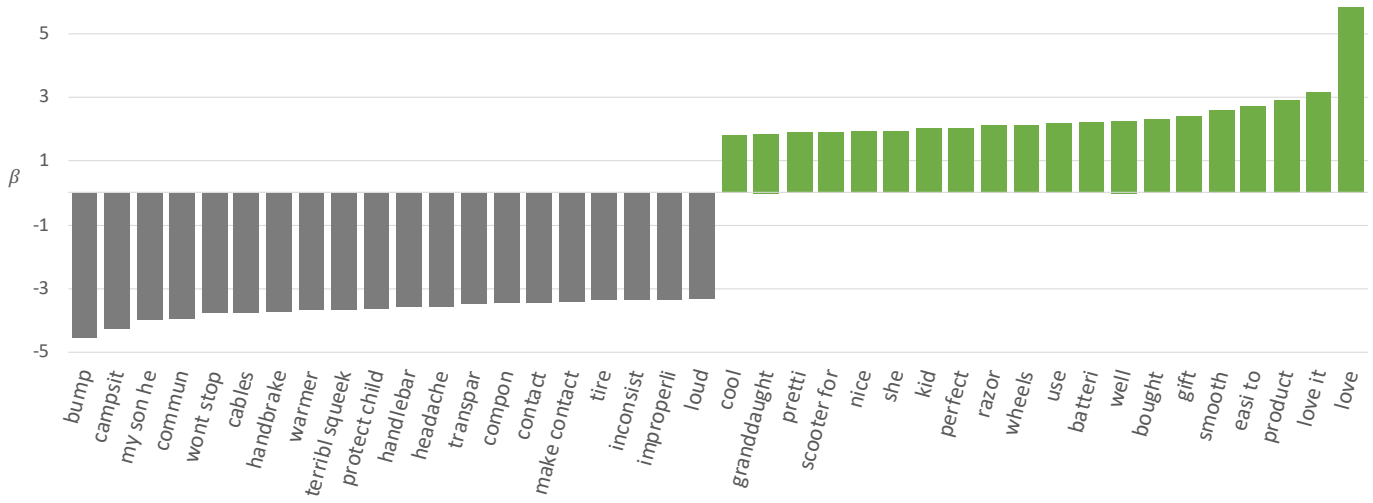

Figure 5.9: Most salient 20 positive and negative features of electric scooters perceived as sustainable for social sustainability

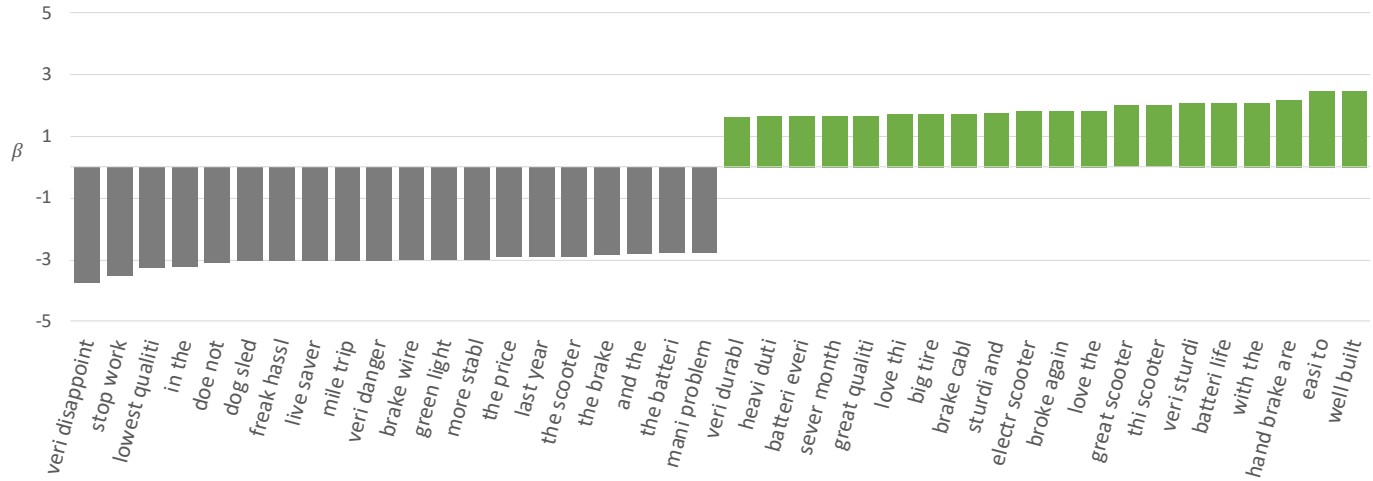

Figure 5.10: Most salient 20 positive and negative features of electric scooters perceived as sustainable for environmental sustainability

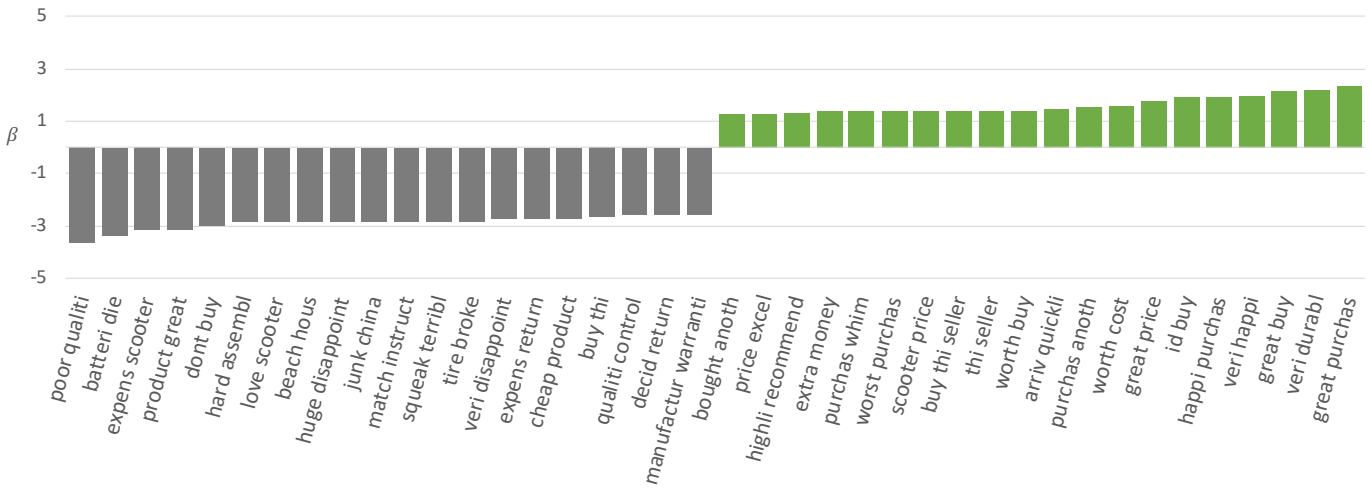

Figure 5.11: Most salient 20 positive and negative features of electric scooters perceived as sustainable for economic sustainability

5.5.1.2 Baby Glass Bottles

The results from the model evaluation and model output for baby glass bottles are shown below.

5.5.1.2.1 Model Evaluation

The precision, recall, and F1 scores for each of the sustainability pillars for baby glass bottles are shown in Table 5.5.

Table 5.5: Precision, recall and F1 scores for baby glass bottle features perceived as sustainable

	Social sustainability			**Environmental sustainability**			**Economic sustainability**		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Positive Sentiment	0.87	0.86	0.86	0.84	0.94	0.89	0.87	0.99	0.93
Negative Sentiment	0.28	0.29	0.28	0.36	0.16	0.22	0.47	0.06	0.11

Similar to our previous findings (Table 5.1), scores for positive sentiment are high ranging from 0.84 to 0.99. Scores for negative sentiment are exceptionally lower,

ranging from 0.06 to 0.29. This suggests that there may be considerable noise in the model output. This emphasizes the importance of data balance for using this approach to extract features. While Amazon reviews are on average more positive, certain products have exceptionally high reviews because they would not survive on Amazon otherwise, for example baby glass bottles. The findings thus far suggest that this approach may not be applicable to such products. Therefore, while proposition 1 may generalize for different products there are limitations in terms of selecting products with balanced reviews.

5.5.1.2.2 Model Output

Figures 5.12-5.14 show the most salient 20 positive and negative features of baby glass bottles based on the parameters of the logistic classifier for social, environmental, and economic pillars, respectively. There is less consistency in the extracted features for baby glass bottles. For example, many of the top negative features contain little meaning or are unintuitive, such as “bit” for social sustainability or “pretty durabl” for environmental sustainability. These are likely due to the low metrics identified in Table 5.5. Therefore, while proposition 1 generalized with electric scooters, it could not generalize with baby glass bottles due to the severe imbalance in product review sentiments.

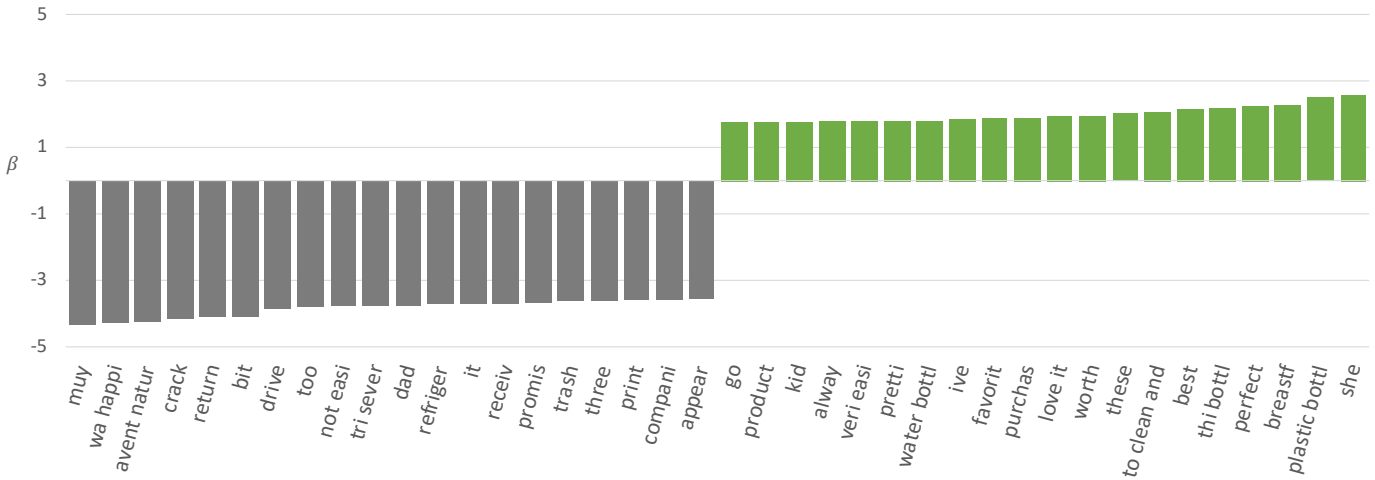

Figure 5.12: Most salient 20 positive and negative features of baby glass bottles perceived as sustainable for social sustainability

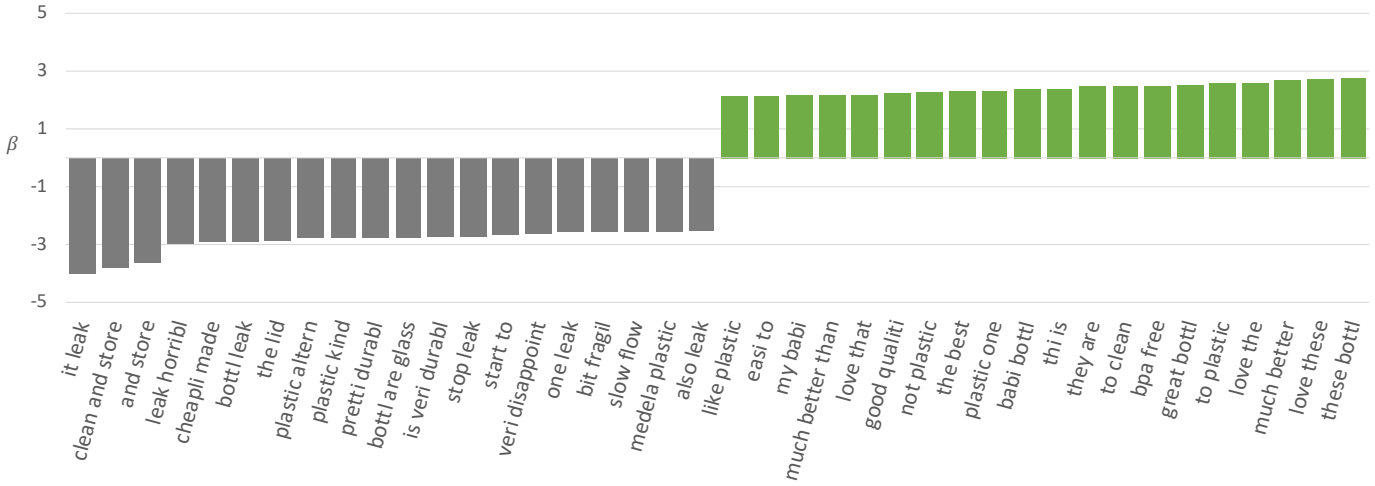

Figure 5.13: Most salient 20 positive and negative features of baby glass bottles perceived as sustainable for environmental sustainability

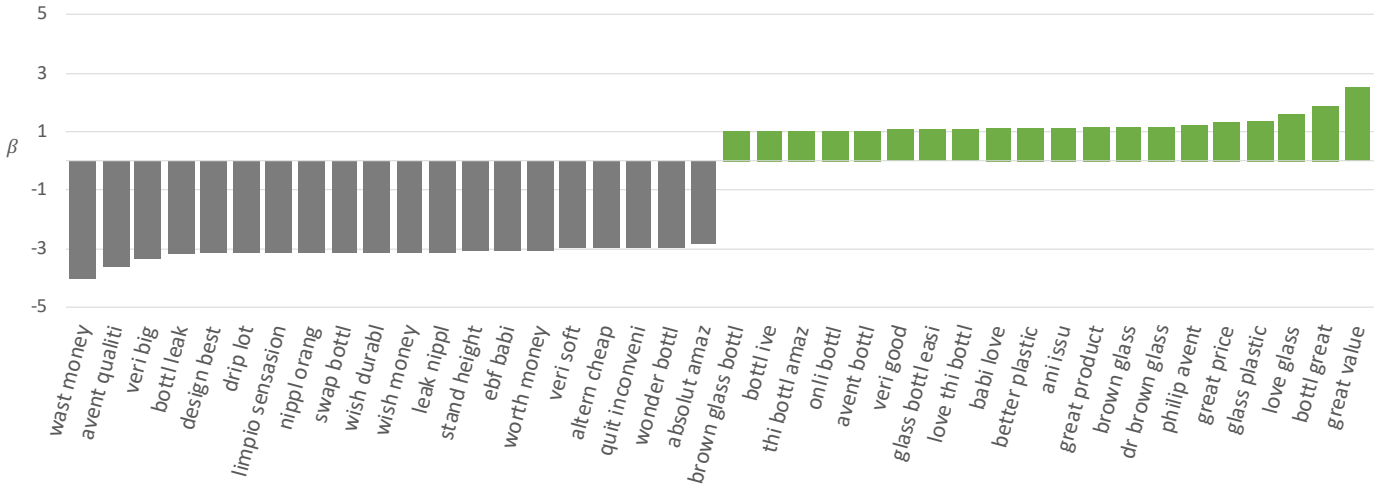

Figure 5.14: Most salient 20 positive and negative features of baby glass bottles perceived as sustainable for economic sustainability

Based on the findings from Section 5.5.1.2.1 and the low model evaluation scores for baby glass bottles, we opted to conduct the collage activity using the features extracted for the electric scooters only.

5.5.2 Collage Results

This section is split into two parts, first we analyze the location of electric scooter features on the collage which test hypothesis 1 and second, we analyze the placement of the products which test hypothesis 2. We excluded 294 datapoints for products that were not moved from their starting location (starting locations are outside the collage boundaries, see Fig. 5.4) from of a total of 1834 recorded datapoints.

5.5.2.1 Feature Analysis

The analysis below tests the generalizability of hypothesis 1: participants evaluating product sustainability on a collage will select features perceived as sustainable for products that they place higher on the “sustainability” axis of the collage.

5.5.2.1.1 Positive and Negative Features Perceived as Sustainable

Based on Figs. 5.12-5.14, we identified 10 positive and 10 negative features to provide to participants during the collage activity for each of the sustainability pillars. These features are shown in Tables 5.6 and 5.7.

Table 5.6: Positive perceptions of electric scooter sustainability

Social Aspects	Environmental Aspects	Economic Aspects
Love it	Well built	Great purchase
Easy to use	Easy to use	Very durable
Great gift	Electric-powered	Arrived quickly
Perfect for kids	Long battery range	Want more than one
Smooth ride	Very sturdy	Comprehensive warranty
Looks pretty	Heavy duty	Happy purchase
Looks cool	Very durable	Excellent price
Want this for my child	Quick charge	Highly recommend
Life saver	Big tires	Buy from this seller
Stable ride	Love this	Very durable

Table 5.7: Negative perceptions of electric scooter sustainability

Social Aspects	Environmental Aspects	Economic Aspects
Loud motor	Very disappointed	Decided to return
Inconsistent power	Stopped working	Useless warranty coverage
Terrible squeak	Difficult to assemble	Too expensive
Very dangerous	Battery died	Won't buy this
Not stable	Poor battery range	Huge disappointment
Difficult to use handbrake	Many problems	Expensive return
Poor ride quality	Brakes failed	Cheap product
Brake cables get tangled	Long charge time	Needs quality control
Tire broke	Low quality	Long wait time
Pure headache	Battery disposal	Piece of junk

Table 5.8 shows a summary of the features selected by the participants during the collage activity.

Table 5.8: Summary of features selected in collage

	Social Sustainability		Environmental Sustainability		Economic Sustainability		Combined	
	Positive Features	**Negative Features**	**Positive Features**	**Negative Features**	**Positive Features**	**Negative Features**	**Positive Features**	**Negative Features**
Number of participants	96		93		97		286	
Observations	355	222	384	154	371	242	1110	618
Average features per participant	3.70	2.31	4.13	1.66	3.82	2.49	3.88	2.16
Average features per product	59.17	37.00	64.00	25.67	61.83	40.33	185.00	103.00
Average features per product per participant	0.62	0.39	0.69	0.28	0.64	0.42	0.65	0.36
Most common feature selected	Great gift	Poor ride quality	Electric powered	Low quality	Excellent price	Too expens-ive	Electric powered	Poor ride quality

Similar to our previous findings with French presses, participants more often selected positive features than negative. This is demonstrated by the average number of features selected per participant. The most selected positive feature across the three sustainability criteria was “electric powered” while the most common negative feature was “poor ride quality”.

Figs. 5.15-5.18 show the average placement of features on the collage for each sustainability criteria, color-code by green for positive and red for negative.

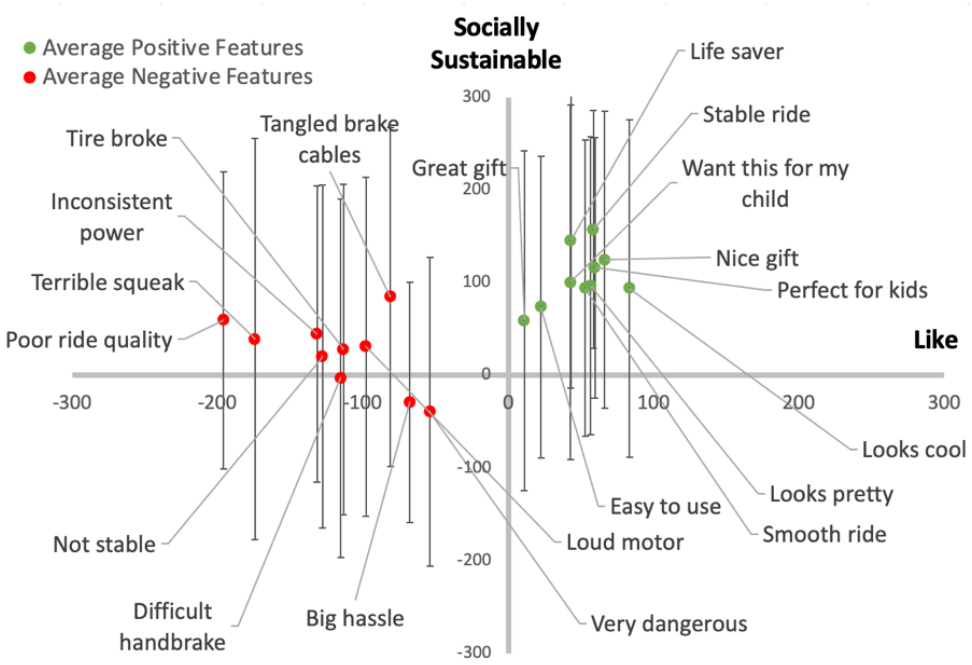

Figure 5.15: Average placement of positive and negative electric scooter features perceived as socially sustainable

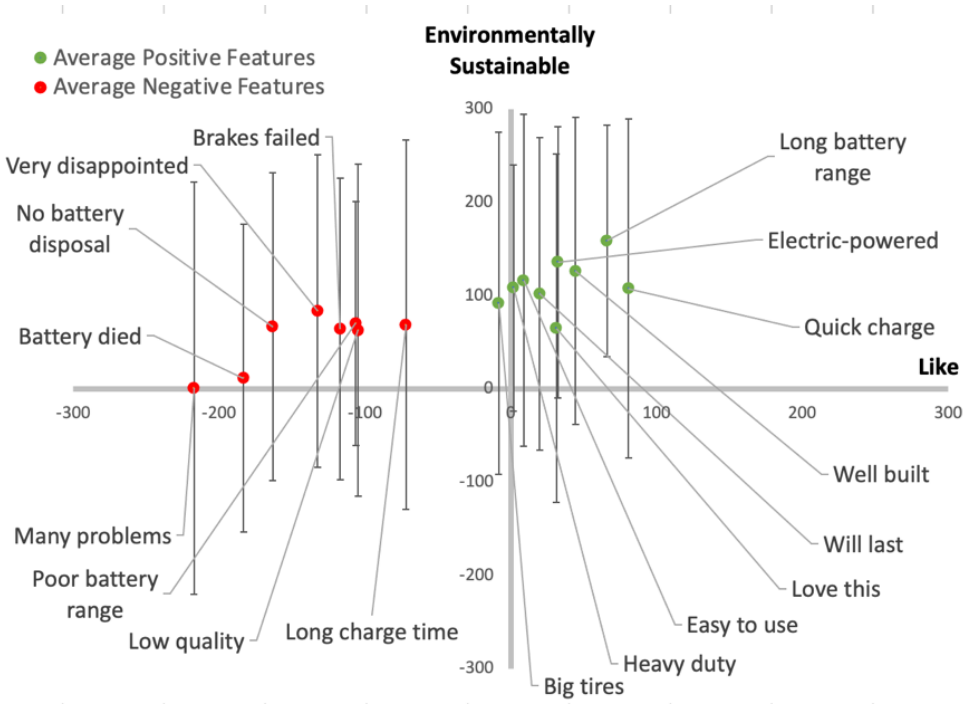

Figure 5.16: Average placement of positive and negative electric scooter features perceived as environmentally sustainable

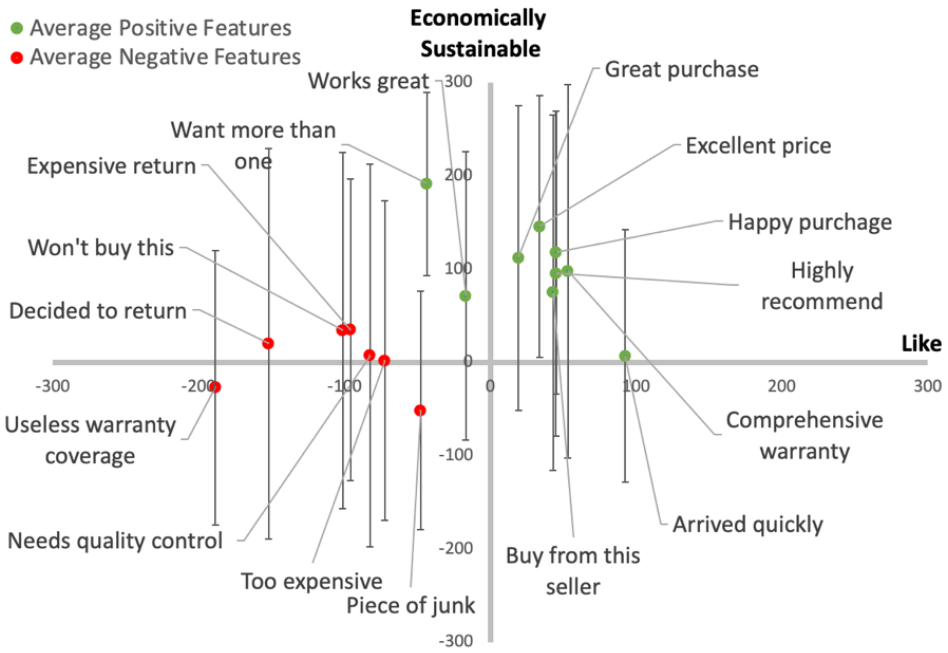

Figure 5.17: Average placement of positive and negative electric scooter features perceived as economically sustainable

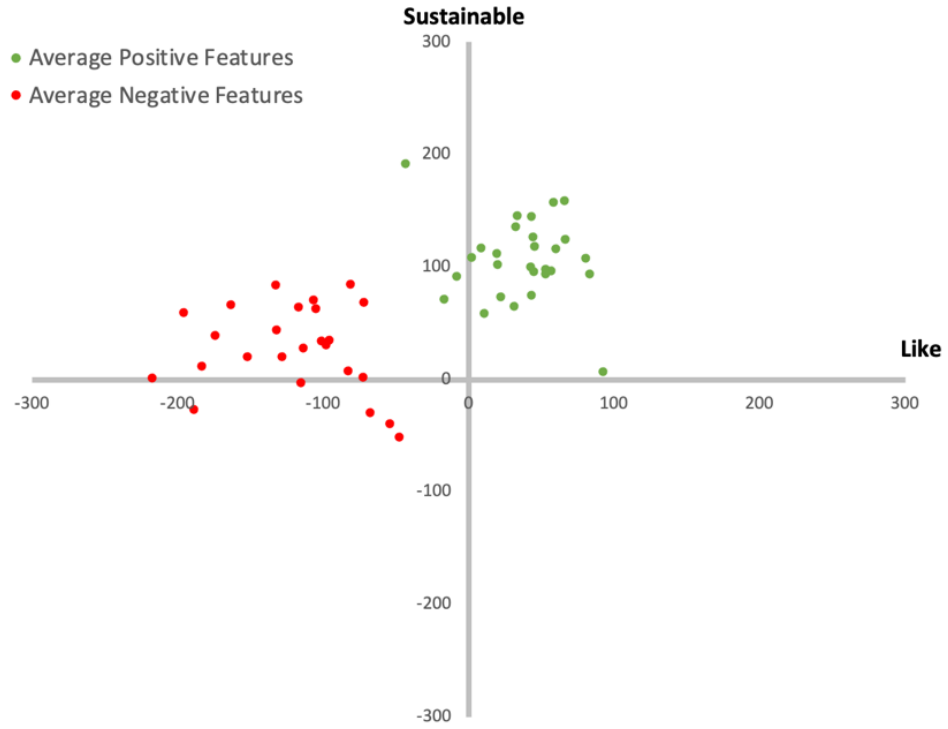

Figure 5.18: Average placement of positive and negative electric scooter features perceived as sustainable for all criteria

The figures show distinct clusters between the positive and negative features, which supports the generalizability of hypothesis 1. We performed a t-test on the y-coordinates between positive and negative clusters to determine if they are statistically different along the sustainability axis for each of the sustainability pillars (Table 5.9).

Table 5.9: Two-sample t-test between positive and negative features perceived as sustainable

*: significant at p = 0.05, **: significant at p = 0.01, ***: significant at p = 0.001

	Social sustainability		**Environmental sustainability**		**Economic sustainability**		**Combined sustainability**	
	Positive features	Negative features	Positive features	Negative features	Positive features	Negative features	Positive features	Negative features
Mean Y-coordinate	103	21	118	56	106	7	110	23.9
Observation	246	137	278	88	209	140	733	365
P(T<=t) one-tail	<0.001***		0.002**		<0.001***		<0.001***	
t Critical one-tail	1.65		1.66		1.65		1.65	
P(T<=t) two-tail	<0.001***		0.004**		<0.001***		<0.001***	
t Critical two-tail	1.97		1.98		1.97		1.96	

Similar to our previous findings with the French press, there was a significant difference along the vertical axis across all sustainability aspects which supports the generalizability of hypothesis 1.

For a more rigorous test that considers repeated measures, we performed a multivariate analysis (MANOVA) using the “x” and “y” coordinates from the collage as dependent variables, and the rest of available information as independent variables (Table 5.10). We chose the Pillai criterion for its robustness when linearity assumptions are not met [73]. Across all sustainability criteria the features were statistically significant, similar to our findings with features extracted for French presses. Thus, our

findings fail to reject hypothesis 1 for electric scooters and validates the generalizability of the hypothesis.

Table 5.10: MANOVA output with positive and negative features perceived as sustainable

*: significant at p = 0.05, **: significant at p = 0.01, ***: significant at p = 0.001

	Social Sustainability			**Environmental Sustainability**			**Economic Sustainability**			**Combined**		
	Pillai	~F	**Pr(>F)**	Pillai	~F	**Pr(>F)**	Pillai	~F	**Pr(>F)**	Pillai	~F	**Pr(>F)**
Product	0.139	5.35	<0.001 ***	0.206	8.51	<0.001 ***	0.230	8.36	<0.001 ***	0.120	14.0	<0.001 ***
Criteria	-	-	-	-	-	-	-	-	-	0.017	4.79	<0.001 ***
FeatureType	0.291	73.2	<0.001 ***	0.206	48.0	<0.001 ***	0.189	37.5	<0.001 ***	0.238	171	<0.001 ***
Age	0.019	1.17	0.323	0.017	0.78	0.618	0.025	1.02	0.421	0.009	1.18	0.308
Race	0.017	1.00	0.422	0.022	1.37	0.222	0.014	0.57	0.799	0.008	1.03	0.408
Gender	0.001	0.18	0.836	0.003	0.55	0.578	0.001***	0.16	0.852	0.001	0.61	0.546
Education	0.041	0.04	0.133	0.077	2.95	0.001	0.020	0.83	0.580	0.013	1.49	0.138
Employment	0.012	0.01	0.628	0.044	2.80	0.011*	0.019	1.06	0.389	0.003	0.47	0.881
Income	0.009	0.01	0.920	0.030	1.89	0.080	0.018	0.72	0.679	0.010	1.11	0.353

5.5.2.1.2 Demographic Interactions

From Table 5.10 we saw that certain demographic variables are significant for different sustainability aspects. To better understand the data, we performed an analysis of variance (ANOVA) to determine which variables are significant specifically with the sustainability axis (Table 5.11).

Table 5.11: ANOVA output for social, environmental, and economic sustainability

*: significant at p = 0.05, **: significant at p = 0.01, ***: significant at p = 0.001

	Social Sustainability				**Environmental Sustainability**				**Economic Sustainability**			
	Sustainability (y-axis)		Like (x-axis)		Sustainability (y-axis)		Like (x-axis)		Sustainability (y-axis)		Like (x-axis)	
	F	**Pr (>F)**	F	**Pr (>F)**	F	**Pr (>F)**	F	**Pr (>F)**	F	**Pr (>F)**	F	**Pr (>F)**
Product	5.30	<0.001 ***	5.30	<0.001 ***	7.36	<0.001 ***	9.71	<0.001 ***	14.5	<0.001 ***	3.04	0.01**
FeatureType	20.2	<0.001 ***	136	<0.001 ***	16.6	<0.001 ***	85.6	<0.001 ***	12.0	<0.001 ***	59.2	<0.001 ***
Age	0.82	0.485	1.33	0.517	0.81	0.730	0.71	0.589	1.16	0.326	0.77	0.543
Race	1.58	0.194	0.55	0.944	0.13	0.189	2.67	0.047*	0.81	0.517	0.29	0.884
Gender	0.11	0.740	0.21	0.395	0.73	0.428	0.29	0.592	0.14	0.707	0.16	0.691
Education	1.49	0.191	1.45	0.207	5.62	<0.001 ***	0.64	0.669	1.05	0.382	0.53	0.711
Emplymnt	0.14	0.938	1.27	0.284	4.66	0.003 **	0.86	0.460	1.29	0.277	0.74	0.526
Income	0.26	0.903	0.57	0.687	2.84	0.038*	0.72	0.542	0.47	0.760	0.91	0.460

Education and employment are both significant variables for how participants evaluated products on environmental sustainability, which agrees with our previous study on French presses. We found that income is also statistically significant for this pillar with electric scooters. None of the other sustainability pillars had significant demographic variables. This contrasts with our previous findings where race was statistically significant for social sustainability and income was significant for economic sustainability. The ANOVA results when combining data from all sustainability criteria are shown in Table 5.12.

Table 5.12: ANOVA output for combined sustainability criteria
*: significant at p = 0.05, **: significant at p = 0.01, ***: significant at p = 0.001

	Sustainability (y-axis)		**Like (x-axis)**	
	F value	**Pr(>F)**	F value	**Pr(>F)**
Product	21.3	<0.001***	7.07	<0.001***
Criteria	7.82	<0.001***	2.12	0.12
FeatureType	54.0	<0.001***	300	<0.001***
Age	1.48	0.206	0.88	0.476
Race	1.26	0.282	0.83	0.508
Gender	0.80	0.372	0.37	0.543
Education	2.23	0.053	0.75	0.586
Employment	0.43	0.786	0.50	0.735
Income	1.53	0.177	0.74	0.600

None of the demographics variables is significant when we combined all sustainability criteria together, which supports our previous findings with French presses that demographics are not significant when considering sustainability as a whole.

While there are similarities in the demographic interactions between our results for electric scooters and French presses, there are inconsistencies too. This is likely

attributed to our skewed demographic in both studies that can lead to inconsistent demographic interaction results.

5.5.2.1.3 Positive Features Perceived as Sustainable and Features Not Perceived as Sustainable

The features not perceived as sustainable that we derived from the unhighlighted parts of the Amazon reviews are shown in Table 5.13.

Table 5.13: Phrases not containing perceptions of electric scooter sustainability

Used for Environmental Aspects Only	
Great assembly	Significant cables
Cheap basket	Easy picture
Awesome brake	Typical work
Small fit	Good brand
Front product	Extra torque

Fig. 5.19 shows the average placement of the new set of features, including ten positive environmental features from Table 5.8 and the features not perceived as sustainable from Table 5.13.

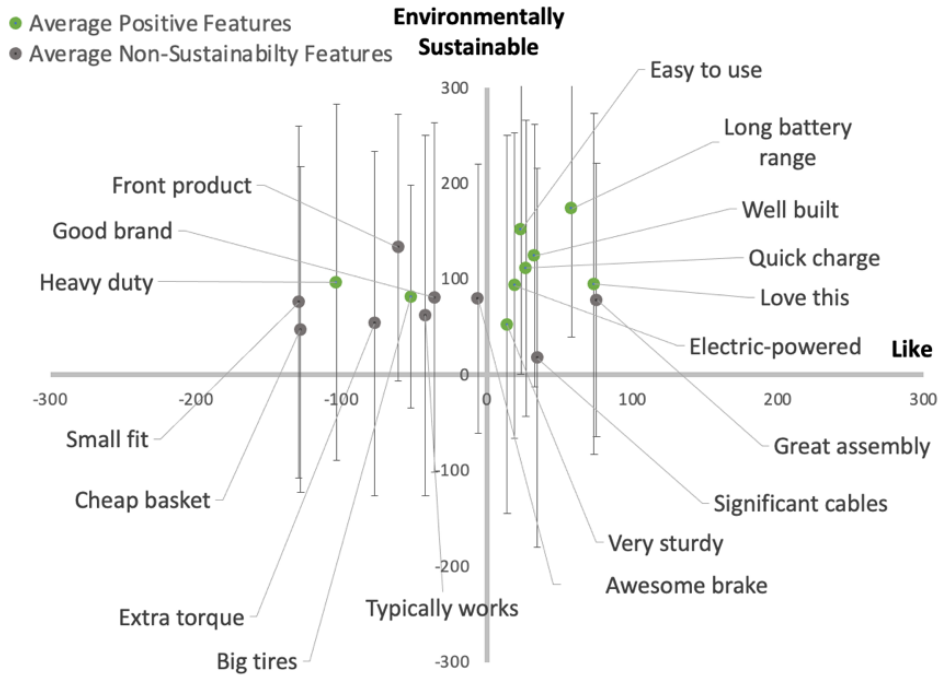

Figure 5.19: Average placement of positive features perceived as sustainable and features not related to sustainability

A t-test using the y-coordinates of the two sets of features is shown in Table 5.14.

Table 5.14: Two-sample t-test between positive features perceived as environmentally sustainable and features not related to sustainability

*: significant at p = 0.05, **: significant at p = 0.01, ***: significant at p = 0.001

	Positive Features	Features not related to sustainability
Mean	109	71
Observations	206	182
Number of participants	72	
Average features per participant	2.86	2.52
Average features per product	34.33	30.33
Average features per product per participant	0.48	0.42
P(T<=t) one-tail	0.013*	
t Critical one-tail	1.65	
P(T<=t) two-tail	0.026*	
t Critical two-tail	1.96	

The t-test shows a statistically significant difference, supporting the generalizability of hypothesis 1. A MANOVA analysis with repeated measures is shown in Table 5.15.

Table 5.15: MANOVA output with positive features perceived as sustainable and features not related to sustainability

*: significant at p = 0.05, **: significant at p = 0.01, ***: significant at p = 0.001

	Pillai	~F	num Df	den Df	Pr(>F)
Product	0.192	8.13	10	768	<0.001***
FeatureType	0.056	11.3	2	383	<0.001***
Age	0.028	1.08	10	768	0.377
Race	0.032	2.06	6	768	0.055
Gender	0.031	3.01	4	768	0.018*
Education	0.036	1.76	8	768	0.081
Employment	0.026	1.25	8	768	0.265
Income	0.009	0.58	6	768	0.746

The electric scooter features are statistically significant even with the more challenging list, similar to our previous findings with the French press. Thus, our findings support the generalizability of hypothesis 1 when using positive features perceived as sustainable and features not perceived as sustainable.

5.5.3 Product Analysis

In this section we present analyses for testing the generalizability of hypothesis 2: A statistically significant relationship exists between the placement of a product on the "sustainability" axis of the collage, and the "like axis of the collage. We used a repeated measured correlation to determine the relation between the "like" axis and "sustainability" axis based on where participants placed the products during the collage activity. The repeated measures correlation controls for between-participant variance [74]. The results are shown in Table 5.16.

Table 5.16: Repeated measures correlation between perceived sustainability of a product and liking the product

	Social sustainability	Environmental Sustainability	Economic Sustainability	Combined
Repeated Measure Correlation	**0.18**	**0.09**	**0.08**	**0.11**
P-value	**0.006**	**0.042**	**0.034**	**0.001**

There is a statistically significant relationship between liking a product and perceiving it as socially, environmentally, or economically sustainable. Moreover, there is a statistically significant relationship between liking a product and perceiving it as sustainable in general. These findings support the generalizability of hypothesis 2 and our previous findings when using a French press.

The correlations are low across the board, ranging from 0.08 to 0.18, suggesting that sustainability and liking a product can be measured separately and demonstrates the usefulness of the collage tool for assessing sustainability perceptions.

5.6 Discussion

The findings in this paper reveal important insights on shaping customer decisions for sustainable products in online markets. Designers often create sustainable products that meet engineering requirements while neglecting features perceived as sustainable by the customer. Including features perceived as sustainable in sustainable products can therefore align product information with customer perceptions and drive purchase decisions. It is crucial that designers meet both engineering requirements and customer perceptions. In this study, we validated the generalizability of our proposed method for designers to extract and test features perceived as sustainable.

Our findings support the generalizability of proposition 1 that phrases perceived as sustainable in reviews contain semantic and syntactic characteristics that can be modeled. Looking at the machine learning model metrics in tables 5.4 and 5.5 for electric scooters and baby glass bottles, respectively, we see that they are similar to the ones in our previous study with French presses (table 5.1). The metrics for negative sentiment faired poorer in this study, suggesting potential limitations on the generalizability (see below for more on limitations).

We found similarities and differences between features perceived as sustainable for electric scooters and baby glass bottles. For social sustainability in Fig. 5.9, many of the positive features for electric scooters were intangible, such as relating to family or gift-

giving. This is similar to what we found when we previously tested the method using French presses. For baby glass bottles in Fig. 5.12, the positive features focused more on the bottle itself. As for the negative features for social sustainability, electric scooters had mainly tangible features relating to convenience, safety, and comfort while baby glass bottles had both intangible features such as “dad” or “promise”, and tangible features such as “crack”.

For environmental sustainability, the positive features for both electric scooters and baby glass bottles are mainly tangible. In the case of baby glass bottles in Fig. 5.13 this mainly related to the material such as “not plastic” and “bpa free”. Our previous results with French presses also showed that positive features for environmental sustainability focused on material. For electric scooters in Fig. 5.10, the positive features included many components such as the battery life, brakes, and tires. The same pattern appeared with negative features where baby glass bottles mainly focused on material while electric scooters included a range of features.

For economic sustainability, features for electric scooter in Fig. 5.11 related to how great of a value the product is, such as “great purchas” or “poor qualiti”. This is similar to what we found with our previous study with French presses. For baby glass bottles in Fig. 5.14, positive features included different brands, as well as tangible features, for example, “plastic” while for negative features they included tangible features such as “bottl leak”.

An important insight from these findings is that the gap between engineered and perceived sustainability can be small or large depending on the purpose and use of the

product. Customers therefore perceive sustainability differently for different products, and perceptions may or may not align with engineering sustainability. For example, while an important sustainability feature such as energy consumption energy was not salient with French presses, it was for electric scooters in terms of battery life. A likely reason for this is that battery life has a direct impact on the customer experience of using an electric scooter. Designers should therefore be clear on and identify the sustainable features that may have a large gap between engineered and perceived sustainability to better align with customer needs. We also saw annotators perceive sustainability differently depending on if a product might be considered a "necessity", such as baby glass bottles, or "nice to have", such as electric scooters and French presses. For social and economic sustainability, tangible features were more apparent for "necessity" products than "nice to have" products. For example, many of the positive social sustainability features for baby glass bottles referred to the bottle, rather than focusing on family or gift-giving such as with electric scooters or French presses. Moreover, brand played an important role for positive economic sustainability features of baby glass bottles while it did not for the other products.

The results from the collage activity demonstrate a strong support for the generalizability of hypotheses 1 and 2. These hypotheses test that participants identify the features we extracted as sustainable and that there is a significant relationship between participants evaluating sustainability and likeability of products. Tables 5.10 and 15 show that participants consistently placed positive electric scooter features perceived as sustainable higher on the sustainability axis than they did for other

features across all sustainability pillars. This is similar to our results with French presses and supports the generalizability of hypothesis 1. It demonstrates that the method in this paper can be generalized to different products to extract perceived sustainable features that customers identify as sustainable. Designers can therefore use this method to identify and include features in sustainable products that align with customer perceptions of sustainability.

We also identified significant relationships between participants perceiving sustainability and liking a product in Table 5.16. The results indicate that the way customers perceive sustainability in products plays a role in how they like products, similar to our previous findings when we used French press products. Our findings therefore support the generalizability of hypothesis 2 that a significant relationship exists between evaluating sustainability and likeability of different products. Moreover, Table 5.16 shows low correlations for the different sustainability criteria, as we found in our previous work. The correlations were lower with electric scooters however, with social sustainability having the highest correlation at 0.18. This contrasts our results with French presses where environmental sustainability had the highest correlation at 38%. This suggests that the role that perceived sustainability plays in customers liking a product can differ between product types. The low correlations for both electric scooters and French presses, however, support that perceived sustainability can be measured separately from liking a product and demonstrate the effectiveness of the collage tool for designers to test perceived sustainable features with participants.

When looking at how demographics play a role with perceiving sustainability, our findings in Table 5.16 are similar to what we found in our previous study. Namely, with both electric scooters and French presses we found that education and employment have a significant effect on participants perceiving environmental sustainability in products. Some differences from last time include that we found that income has a significant effect on environmental sustainability, while demographics did not have any effect for social and economic sustainability. We did find however that when sustainability pillars are combined, the effects of all types of demographics become negligible on perceptions similar to when we used French presses. The inconsistencies between our two studies are likely due to the skewed demographic base that is not highly representative.

Our findings reveal crucial implications for guiding customers to making informed purchase decisions. We recommend that designers use the method in this study so that they may bridge the gap between perceived and engineered sustainability in their products and drive purchase decisions. We showed how the method can derive insights for different products and confirmed that participants identified the features extracted from online reviews as sustainable using the collage tool. Although features perceived as sustainable may not directly influence the sustainability of the product, it communicates the sustainability of the product to the customer in a way that matters to them and can influence their decisions towards informed purchases. Moreover, we recommend that designers use the collage tool to test perceived features with customers, including how this may differ across different demographics.

The findings in this paper do come with limitations. Amazon products tend to have more positive than negative reviews by design, as they would not thrive on the platform if it were the other way around. One limitation therefore is that the machine learning model performance may suffer due to an imbalance in the dataset (Tables 5.4 and 5.5). We saw this in our results, for example, “love scooter” appears as a salient negative economic sustainability feature for electric scooters in Fig. 5.11. We found a greater imbalance in the annotated reviews for baby glass bottles, possibly because reviews for them tend to be exceptionally high to survive on Amazon. Designers, therefore, need to carefully assess potential data imbalance before using this method. Possible workarounds include collecting enough negative reviews to have a neutral overall rating score when annotating, or to collect reviews from a different platform that may have more balanced ratings.

5.7 Conclusion

This study validates the generalizability of our previously developed method to extract and test features perceived as sustainable from online reviews for different products. The insights from our results can help shape customer decisions to make informed purchases. To demonstrate this, we used two focal products, electric scooters and baby glass bottles, and recreated our previous work where we used French presses [45,81]. We collected Amazon reviews for the focal products and recruited Amazon Mechanical Turks (MTurks) to annotate fragments of the reviews that are relevant to one of the sustainability pillars – social, environmental, and economic. We then modeled the annotations using a logistic classifier and extracted features perceived as

sustainable based on the parameters of the model. We confirmed that the perceived features were identified as sustainable using a novel collage activity. We tasked participants with placing products along the two axes of the collage, sustainability and like, and to select from a dropdown menu features that we extracted from the machine learning model.

Based on the results we found that our previously proposed method does generalize with limitations. We shared crucial design insights that can help customers make informed sustainability purchase decisions. Designers can use the method in this study on different products to identify the gaps between perceived and engineered sustainability and create sustainable products that can drive purchasing decisions. We demonstrated that this gap can be small or large for different products and confirmed that this method can be applied to identify salient sustainable features based on customer perceptions. We recommend that designers use the collage tool to test and better understand customer perceptions of sustainability. We demonstrated how perceived sustainability and liking a product can be measured separately based on their low correlation. Moreover, we recommend that designers consider the influence of demographics when focusing on perceptions of environmental sustainability. The effects of demographics become negligible when considering all three sustainability pillars together.

A limitation to our findings is that the method can be ineffective if there is an imbalance of positive and negative annotations from the reviews. Products should therefore be carefully selected to ensure a balanced dataset. Moreover, we recommend

conducting a more thorough demographic analysis to better understand the relationship between demographics and customer perceptions. Finally, our analyses do not include real purchase decisions. For next steps, we aim to address the limitations in this study by exploring modifications to the method for imbalanced data and investigating how features perceived as sustainable in products influence purchase decisions.

6. CHAPTER 6

A Test for Product Design Features Perceived as Sustainable to Drive Online Purchase Decisions

Abstract

Designers are challenged to create sustainable products that succeed in the marketplace, often relying on life cycle analyses to identify engineered sustainable features while neglecting perceived-as-sustainable (PAS) features. PAS features may not contribute to engineered sustainability but are identified by customers as sustainable. In previous papers we proposed methods for extracting PAS features from online reviews using machine learning techniques and validating them using collage placement techniques. We demonstrated our methods using French presses (and other products). In this paper, we combined design and marketing approaches to test our previously extracted PAS features in terms of valuing and purchasing products that include PAS features, as compared to others that do not. We built a simulated Amazon shopping experience with incentive alignment and constructed a within-subject, fractional-factorial design that included a variety of product features and physical appearances. We collected data on purchase intent, willingness to pay, and sustainability rating. We found that participants opted to purchase products with PAS features more often than products with dummy features. Participants also indicated they were willing to pay more for products with PAS features and rated those products as more sustainable, despite the features not contributing to engineered sustainability. Our findings demonstrate the potential value of identifying and including PAS features in sustainable products and a new application for shopping simulation experiments in design research.

We recommend that sustainable designers include both engineered (real) and PAS features in sustainable products to align with customer needs, drive purchasing decisions, and potentially increase profitability.

6.1 Introduction

Research on customer purchase intent versus actual purchases has revealed gaps, for examples see [2,91]. This is especially true for sustainable products, where factors such as social desirability bias can influence what customers share about their intentions. In a paper towel survey for example, MacDonald et al. found that 87% of participants stated they would not buy non-recycled paper, but also indicated they bought from brands with non-recycled paper the last time they went shopping [92]. This mismatch challenges designers to create successful sustainable products in a market with an apparent growing demand from customers [7], but unfortunately a lack of sustainability information and/or knowledge among customers.

To address this challenge, designers have created methods for identifying design cues that can help customers form product perceptions and boost their intentions to purchase sustainable products. MacDonald et al. demonstrated that customer preferences are constructed, in-part, based on the context of the purchase decision and are not necessarily innate in people. The authors provided participants with slightly modified versions of discrete choice surveys for paper towels and found inconsistent preferences between them [65]. She and MacDonald built on this by demonstrating how visual design features termed “sustainability triggers” led participants to prioritize hidden sustainability features in a realistic decision scenario for toaster prototypes [2].

For example, a sustainability trigger such as an embossed leaf pattern correlated with a prioritization of engineered sustainability features such as energy usage in survey questions on purchase intention. Seemingly superfluous features can therefore help customers value engineered (real) sustainability information about a product that aligns with their perceptions.

Designers typically focus on the hard facts of sustainability when designing sustainable products. They use tools such as Life Cycle Analysis (LCA) to prioritize design goals, such as energy usage and recyclability [4]. Unfortunately, much of this hard work is hidden within the final product, and unless customers know the right questions to ask, think to ask these questions, and know where to find the relevant information, many sustainable design efforts are never known to the customer. Much of what the customer perceives as related to sustainability is what they can see on the surface of the product.

In a previous paper, we proposed a method for designers to extract perceived-as-sustainable (PAS) features from online reviews using a natural language processing machine learning algorithm combined with human annotators [45]. While these features may not contribute to engineered sustainability, meaning they do not decrease the life-cycle impact of the product, the features aid in communicating the purpose of the product to the customer. The inclusion of PAS features supports existing sustainable

[4] http://www.sustainableminds.com/

design methods, for example, LCAs, in that designers can create sustainable products that meet both engineered and perceived sustainability requirements.

As a case-study for our previously proposed method, we extracted salient PAS features from French press online reviews that drove positive and negative sentiment. We demonstrated that there is a gap between engineered and PAS features, highlighting the importance of accounting for both in design. In a subsequent paper, we used a novel collage approach to validate that users identified the PAS features as sustainable despite these features not necessarily contributing to engineered sustainability [81].

In this paper, we conduct a strong test of PAS features by investigating how they can drive purchasing decisions for sustainable products in a simulated Amazon shopping experience (Fig. 6.1). Our goal is to provide a validated method for designers to create sustainable products that resonate with customers and drive purchasing decisions that are based, in-part, on valuing sustainability. We test the French press features extracted from our previous paper using a within-subject fractional factorial experiment design to demonstrate how PAS features influence purchase decisions. The rest of the paper is organized as follows: the background is presented in Section 6.2, followed by an overview of the propositions and hypotheses in Section 6.3. In Section 6.4 we describe the methods, in Section 6.5 we present the results, and in Section 6.6 we discuss the findings. Finally, we conclude the paper in Section 6.7.

Figure 6.1: Current paper builds off work from previous papers

6.2 Background

There is an extensive body of research from design and marketing on investigating customer choices and purchase decisions. Our work utilizes approaches from both research areas to identify how PAS features may influence online purchasing decisions. Common design approaches include conjoint and discrete choice analyses to tease-out preference of product features presented in different combinations of options. Designers can model and predict which product configurations are the most valued by customers based on these preferences [93]. Marketing approaches typically rely on historical data to model factors that influence purchase decisions. This section provides an overview of customer preference modeling from design and marketing and provides an overview of our previous papers that we build off for this paper.

6.2.1 Customer Preference Modeling in Design

This section presents an overview of recent customer preference modeling research in design. Suryadi and Kim proposed an automated method to construct choice sets using online product information and customer reviews [94]. The authors mined Amazon product data for 84 laptops and 46,194 verified customer reviews. From the data they clustered products using X-means on the attributes, clustered customers using vector representation similarity of the reviews, and then constructed choice sets using a multinomial logit model. Using KL divergence, the authors showed that the generated choice sets have higher preference predictive ability compared to a baseline random constructed choice set.

Goucher-Lambert et al. used functional magnetic resonance imaging (fMRI) to investigate how customers make multi-attribute product decisions when considering sustainability [95]. The authors recruited participants to complete a within-subject conjoint analysis inside an fMRI. Participants were presented with two product options at a time with information on their form, function, and price. In the control condition participants were also shown the Poisson's ratio while in the test condition participants were shown information on the environmental impact. Using empirical fMRI results, the authors found that participants prioritized function while deprioritizing visual appearance when given environmental information about products. This work validated findings of a previous conjoint analysis study [96].

Tovares et al. proposed a method for incorporating experiences into consumer preference modeling [97]. The authors used virtual reality (VR) to provide participants

with the ability to interact with products before indicating their preferences. Two within-subject experiments were conducted: one explored layout preference using a truck cab dashboard and the other explored form preference using mugs. For each experiment, participants completed a conjoint analysis with an experiential setup and a non-experiential (standard) setup. In addition, for the mug experiment, participants completed a "real" conjoint analysis where they interacted with real mugs before indicating preferences. The authors found that the experiential conjoint analyses did not provide better preference predictive capabilities than the visual conjoint analysis, although the results from the experiential and real mug conjoint analyses were statistically similar.

Maccioni et al. investigated preferences for sustainable products using a combination of stated preferences and biological measurements [67]. The authors recruited 43 participants to evaluate 20 baseline products and 20 eco-friendly products. Participants wore eye-tracking equipment and a device that measures galvanic skin response while evaluating products. The authors found that participants perceived eco-friendly products as more innovative while they perceived the baseline products as more functional and reliable. No significant results were found from biological measurements.

6.2.2 Customer Preference Modeling in Marketing using Online Reviews

This section presents some relevant papers from marketing research that focus on customer preference modeling using online reviews.

Chevalier and Mayzlin studied the impact of online reviews on sales using data from Amazon and Barnes and Nobles [98]. For a sample of books, the authors compared differences in the number of reviews and their ratings over three time points from both websites and determined their relationship with relative sales rank. Using a linear model, the authors found that positive reviews on one site correspond to higher sales relative to the other site. Moreover, they found that the decrease in sales from a negative review is greater than the increase in sales from a positive review.

Chen et al. disaggregated the impacts of online reviews and recommendations on online sales rank [99]. The authors used digital cameras as a case-study and collected information on number of reviews, ratings, recommended cameras in terms of purchase percentage, and sales rank. Using a linear model, the authors found that a negative review had a much greater impact on sales than a positive one. Moreover, they found that positive recommendations (high purchase percentages, for example “86% of users ultimately purchase this product.”) have a positive effect on sales while negative recommendations (low purchase percentages) have no effect on sales.

Liu studied the impact of Yahoo movie reviews on Box Office revenues [100]. The author found that reviews are most active during the prerelease of a movie and more critical after the release. Moreover, using a linear model Liu found that the volume of reviews around the time of release correlates with Box office revenues and not the valence of the reviews. Dhar and Chang built on this by studying the impact of blog posts and social networking sites on music sales [101]. The authors collected the volume of blog posts for an album, the number of friends an artist has on Myspace (a social

media platform), and the number and ratings of online album reviews. The authors used a linear model to study the impact of the data on music sales four weeks before and after the album release. Data for 108 albums were collected. Album sales were computed based on sales ranks from Amazon.com. The authors found a positive correlation between the volume of blog posts with future album sales.

A limitation of the above approaches is that they do not study how specific product features may be driving online sales. Our work utilizes approaches from design and marketing to test PAS design features and provide actionable insights for designers on driving purchasing decisions for sustainable products. An overview of our previous work is included below to provide context on PAS design features.

6.2.3 Extracting and Testing Features Perceived as Sustainable from Online Reviews

We briefly summarize two previous papers here, as this paper builds off them. The first paper developed a semi-automated approach to extract features perceived-as-sustainable (PAS) from online reviews using crowdsourced annotations of online reviews and a natural language processing algorithm [45]. As a case-study, we used French presses and collected 1474 reviews to extract PAS features from. We recruited 900 annotators from Amazon Mechanical Turk (MTurk) to highlight phrases in reviews they perceived as sustainable. Annotators were trained and assigned to one of three sustainability pillars: social, environmental, and economic. Using a logistic classifier model and for each sustainability pillar, we then extracted salient PAS features with positive and negative sentiment based on the beta parameters of the model. Table 6.1

shows the positive salient features extracted. A subset of these features is selected for this paper (see section 6.4.2.1).

Table 6.1: Positive features of French presses perceived as sustainable [45]

Social Aspects	**Environmental Aspects**	**Economic Aspects**
Easy to use	Well made	Easy to clean
Love it	Easy to use	Great quality
Nice gift	Strong glass	Want more than one
Good for my family	Easy to clean	Reasonable price
Perfect for two	Solid design	Works great
Use with my spouse	Will last	Worth the price
Take to work	Stainless steel	Good customer service
Easy to clean	No plastic	Great value
High quality	Metal frame	Best price
Works great	Sturdy	Hard to beat

The features in Table 6.1 include a combination of visual and descriptive, and tangible and intangible features. Notably, energy and water consumption were not identified as salient environmental PAS features although they are important engineered sustainability requirements for French presses. To investigate this further we conducted a life-cycle analysis and found the use phase (where energy and water consumption contribute) had one of the largest environmental impacts (Fig. 6.2). This gap between engineered and PAS features highlights the importance for designers to consider both in sustainable design.

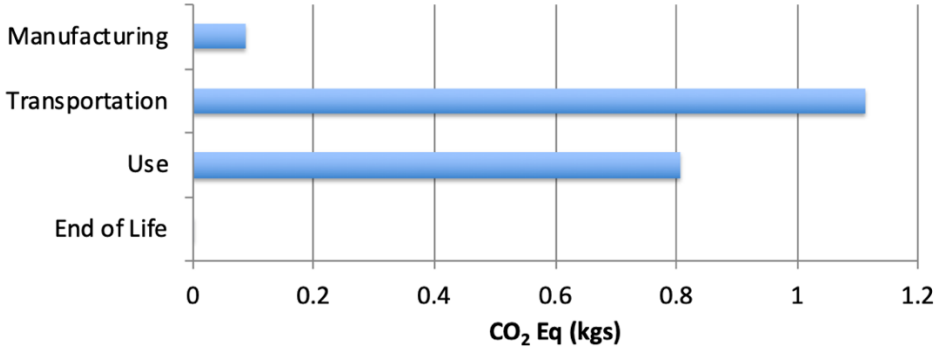

Figure 6.2: Life Cycle Analysis of French Press

The second paper tested the extracted PAS features to determine if users would identify them as sustainable and how the features might relate to users liking a product [81]. We designed a novel collage approach where participants dragged and dropped products on a set of two axes, sustainability and likeability, and selected features from a dropdown menu. Figure 6.3 shows an example of a product being placed on the collage with features to select from. The placement of products and features on the collage validated that participants identified PAS features as sustainable and that the collage is an effective tool for testing customer perceptions.

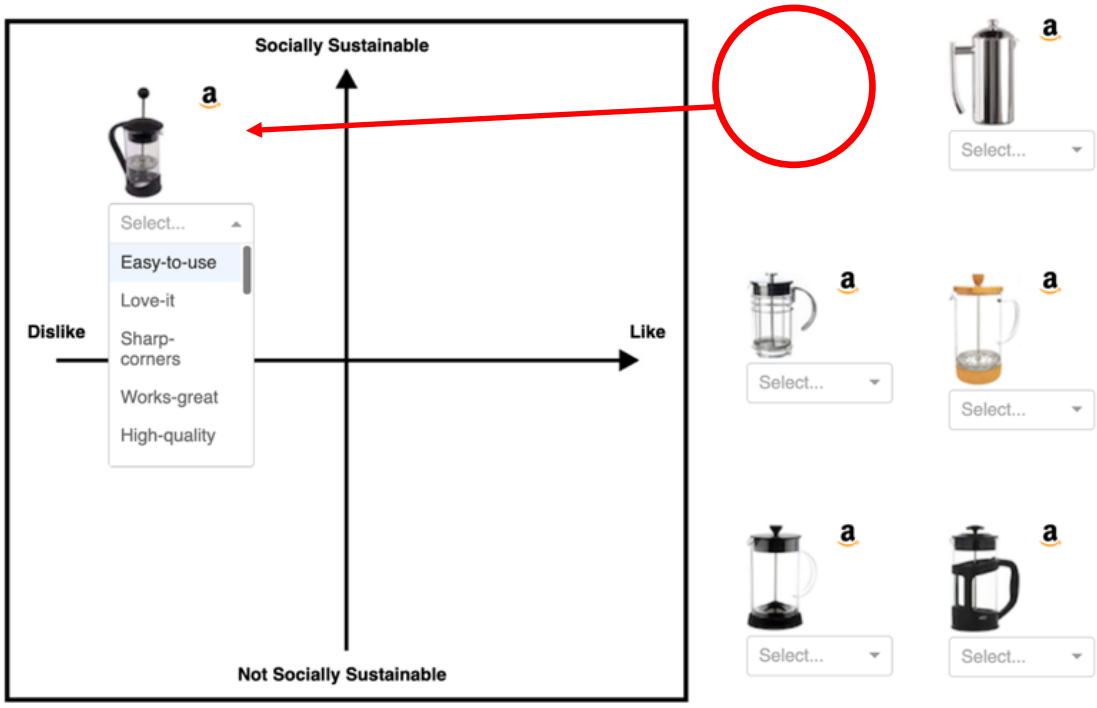

Figure 6.3: Dragging and dropping products on collage and selecting at least one feature to describe each product

6.3 Research Propositions and Hypotheses

In this research, we leverage perceived-as-sustainable (PAS) French press features identified in previous work. These features do not contribute to real engineered sustainability [45]. We test these features in a simulated Amazon shopping experience where we modified images, descriptions, and reviews according to PAS features as well as "dummy" features. We tested the influence of the PAS features on purchase decisions in a within-subject study. The following propositions and hypotheses are tested.

PROPOSITION 1: Online customers rely on product descriptions to guide their purchasing decisions. Based on this, we propose that designers can modify descriptions to drive purchasing decisions for sustainable products.

Hypothesis 1: Participants are **more likely** to opt to purchase a product when the description is combined with features extracted from online reviews that are perceived-as-sustainable versus dummy features.

PROPOSITION 2: Online customers rely on product descriptions to learn about products. Based on this, we propose that designers can modify product descriptions so that customers resonate more with sustainable products.

Hypothesis 2a: Participants will rate a product as **more desirable** when the description is combined with perceived-as-sustainable features extracted versus dummy features.

Hypothesis 2b: Participants will rate a product as **more sustainable** when the description is combined with perceived-as-sustainable features versus dummy features.

6.4 Method

To test the hypotheses, we designed and conducted a simulated Amazon shopping experience for 200 Amazon Mechanical Turk workers (referred to as participants, see section 6.4.4 for more information). Participants browsed between product options based on fractional factorial design and selected to "purchase" a product as if they were making real purchase decisions. Participants also completed a post-experiment-survey, in which they rated the products based on desirability and sustainability. In the following sections we provide an overview of the experiment design and the products and features used in the shopping simulation. The experiment contained base and dummy features that we created for this experiment, and PAS features extracted from a previous paper [45]. Then we discuss how we tested the products and features in a simulated Amazon shopping experience. Finally, we provide an overview of the post-survey and the participants in the experiment.

6.4.1 Experiment Design Overview

The experiment compared how participants made purchasing decisions when given products with dummy features in a control condition versus products with PAS features in a test condition (see Fig. 6.4). The stimuli included base products to create a reference point between both conditions. We used a within-subject experiment design to assess how PAS features can influence purchasing decisions.

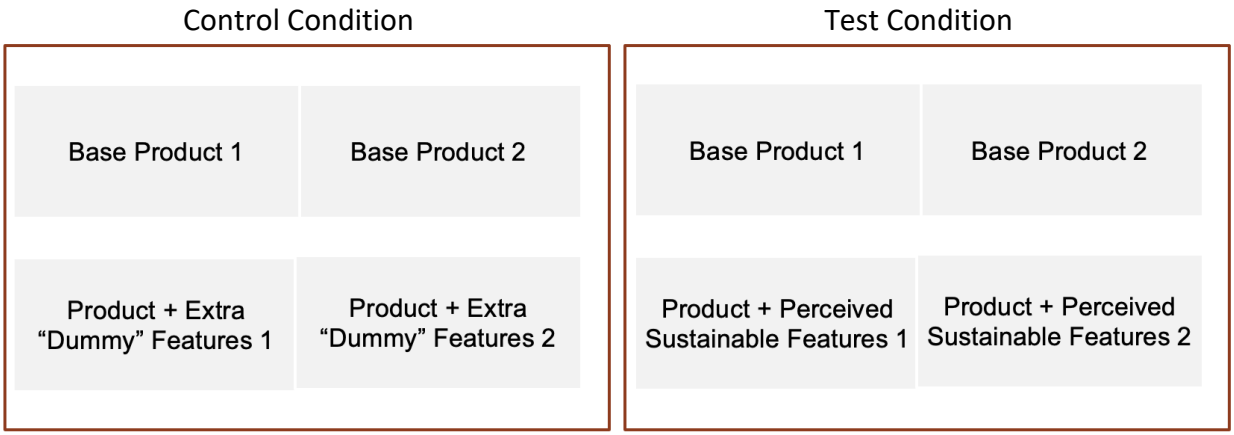

Figure 6.4: Within-subject experiment design

The experiment was conducted via a Qualtrics survey with instructions about the activity, a test to make sure participants understood the task, and links to the shopping simulations. The test questions asked participants to recall from the instructions how many product options were included in each shopping simulation, the type of product they were shopping for, and the number of shopping simulations they were completing.

After passing the test, participants received links that led them to the control and test condition shopping simulations. Participants always completed the control condition first to limit the chance of social desirability bias influencing participants' choices in the following condition. The goal was to provide the least advantage for PAS features to rigorously test their ability to drive purchase decisions in the test conditions. Each shopping simulation condition displayed four products to browse from. In the control condition two base products and two products with dummy features were displayed. In the test condition the same two base products, and two products with PAS features were displayed. Participants had to spend a minimum of five minutes on each simulation before they could proceed to the next one. To incentivize participants to

evaluate products carefully, we used incentive alignment as part of their reward (see section 6.4.4).

Following the completion of the shopping simulation, participants received a password to a Qualtrics post-survey. Participants entered which product they selected for purchase and rated all products on a 5-point Likert scale based on their desirability and based on their sustainability. As a proxy for desirability, the survey asked participants to rate their willingness to purchase a product. Participants also selected from a list the main driving factor in their shopping decisions online. Lastly, participants entered demographic information before the completion of the survey.

6.4.2 Products

The experiment focused on French press coffee brewers, building off our findings from previous papers [45,81], because they are popularly reviewed products, sold with various aesthetic and practical design features. Additionally, the sustainability concerns associated with French presses present an opportunity to study consumers' perceptions of sustainability, such as how quickly a glass exterior might break or how much energy can be saved with a heat-insulated press. The experiment showed four presses that had a variety of features in each condition to simulate a realistic shopping experience. In this section we describe the features, images, descriptions, and reviews that we used in the paper.

6.4.2.1 Product Features

The presses had three feature categories: base features, dummy features, and PAS features (Table 6.2). Each category had two levels that could be displayed. Note

that French presses had dummy features in the control condition, and PAS features in the test condition, never both.

Table 6.2: Breakdown of product features

Feature Category	Feature Name	Level 1	Level 2
Base	Handle shape	Circular	Rectangular
Base	Spout	Filter	Easy-pour
Base	Lid	Button	Lift
Dummy	Hourglass timer	Present	Not present
Dummy	Ventilated lid	Present	Not present
PAS	Material	Stainless Steel	Plastic
PAS	Glass	Strong glass	Not present
PAS	Clean	Easy to clean	Not present
PAS	Quality	High quality	Not present
PAS	Gift	Perfect gift	Not present

All products shown included all three base features at one level, consisting of core functional features of a French press, including a handle, spout, and lid. For the handle, the varieties included either a circular or rectangular shape; for the spout, the varieties included filtered or easy-pour; and for the lid these included a button or lift mechanism. The purpose of these features was to provide a baseline to compare purchasing decisions between the two experimental conditions. The base features pilot tested as neutral and did not impact customer sentiment significantly, see section 6.4.3.2 for details.

Dummy features are intended to appeal to customers for their functionality but are not strongly related to either engineered or perceived sustainability. These included a built-in handle hourglass timer for proper brewing time, and a ventilated lid to help steam escape. The dummy features were included in the control conditions only. The goal of the dummy features was to challenge and assess the popularity of products with

PAS features. Pilot-testing aided in selecting dummy features that were on par with the PAS features in terms of desirability (see section 6.4.3.2).

A previous paper developed a method to extract PAS features from online reviews, demonstrating the method using French presses and identifying a gap between engineered and PAS features (Table 6.1) [45]. We selected a subset of PAS features for this experiment, shown in Table 6.2. We selected this subset because it includes both visual and descriptive features and is representative of perceptions from the three sustainability pillars: social, environmental, and economic.

For the material feature, products were made of plastic or stainless steel (with steel being a PAS feature). For the remaining PAS features, the products either had or did not have them, e.g.: strong glass, easy to clean, high quality, and perfect gift—all PAS. The PAS features were included in only the test condition.

6.4.2.2 Fractional Factorial Experiment Design

With the available features in Table 6.2, we used a fractional factorial experiment design to account for different combinations and created twelve different products—each participant saw eight of these products. The features per product are shown in Table 6.3. For each product we created images, descriptions, and reviews to include in the Amazon shopping simulation as described below.

Table 6.3: Features per product

French Press #	Condition	Base Features						Perceieved Sustainable Features					Extra Features	
		Handle		Spout		Lid		Stainless steel	Strong glass	Easy to clean	High quality construction	Perfect gift	Hourglass timer	Ventilated lid
		Circular	Rectangular	Filter	Easy-pour	Button	Lift							
1	C&T	■		■		■								
2	C&T		■		■	■								
3	C		■	■			■						■	
4	C	■			■		■							■
5	T	■		■		■		■						
6	T		■		■	■			■					
7	T		■	■			■			■				
8	T	■			■		■				■			
9	T	■		■		■						■		
10	T		■		■	■		■	■					
11	T		■	■			■	■		■				
12	T	■			■		■	■			■			

6.4.2.2.1 Product Images

We rendered images of the 12 products for this study using the computer-aided design software Fusion 360 (Fig. 6.5). Products 1 and 2 include base features only and are shown in every experiment condition. Products 3 and 4 have dummy features and are only shown in the control condition. Products 5 to 12 have PAS features; we randomly created five random pairs and assigned participants to a test condition with one of five pairs. The other two products in the test condition were the base products.

Each French press was designed to closely resemble the others, as well as those on the market. All products were shown on a white background from the same angle. Additionally, each product had a close-up image of the handle and top.

Pilot-testing ensured that the rendered product images were equally aesthetically pleasing (see section 6.4.3.2), aiming to minimize the effect of other potential variables on purchasing decisions. It is important to note that some features in Table 6.2 cannot be shown visually, for example “easy to clean”. These features are included in descriptions on the Amazon simulated product page.

Figure 6.5: Product image renderings

6.4.2.2.2 Product Descriptions

Each product had a corresponding description that outlined the product's features in a bulleted list. Descriptions were written with the goals of brevity, maintaining Amazon's organizational structure, and emphasizing the feature. Each description ranged between 25 and 30 words, and the feature was described at the beginning of each list item (see Fig. 6.8 in section 6.4.3.1 for an example).

Products were also titled using their features, a technique that is commonly used on Amazon. For example, Product 1 was named "French press Coffee Maker with button for lid removal, filtered spout, and circular grip handle." These descriptions made the product features noticeable. Pilot-testing ensured that the descriptions were equally readable and understandable (see section 6.4.3.2).

6.4.2.2.3 Product Reviews

Each product was shown with five unique reviews: three 5-star reviews and two 3-star reviews. All products were rated as 4.2 stars overall. Having products with equally positive reviews mitigated the influence of reviews on purchasing decisions. Additionally, the reviews did not mention any of the features in the experiment and did not mention sustainability. We created the review text as follows. First, we selected initial candidate reviews from Amazon listings of French presses and then removed specific details so that the reviews were applicable to any French press. Each review had only two to three sentences total. Pilot-testing ensured that the five reviews for each product were equally positive (see Section 6.4.3.2).

6.4.3 Amazon Shopping Experience

In this section we present the shopping experience flow as well as measures taken to normalize web content between products.

6.4.3.1 Simulation Flow

Participants were able to freely click and browse between three types of pages: the Product Search Page, the Product Pages, and the Checkout Page (Fig. 6.6). The experience was a controlled simulation without distracting content, such as advertisements and hyperlinks to other web pages on Amazon.com.

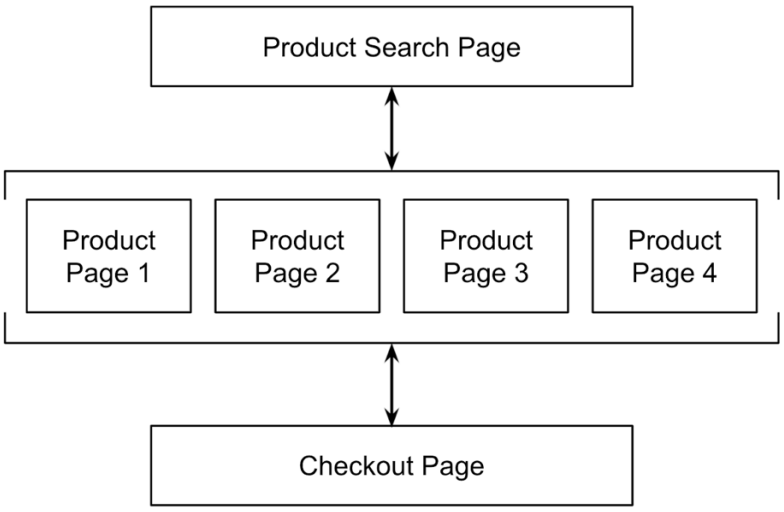

Figure 6.6: Simulated Amazon flow

The Product Search Page showcased four French presses on the participant's screen (Fig. 6.7). This page was intended to replicate the results of a consumer searching "French presses" on Amazon.com. The products shown on this page depended on which experiment condition the participants were taking (see section 6.4.2).

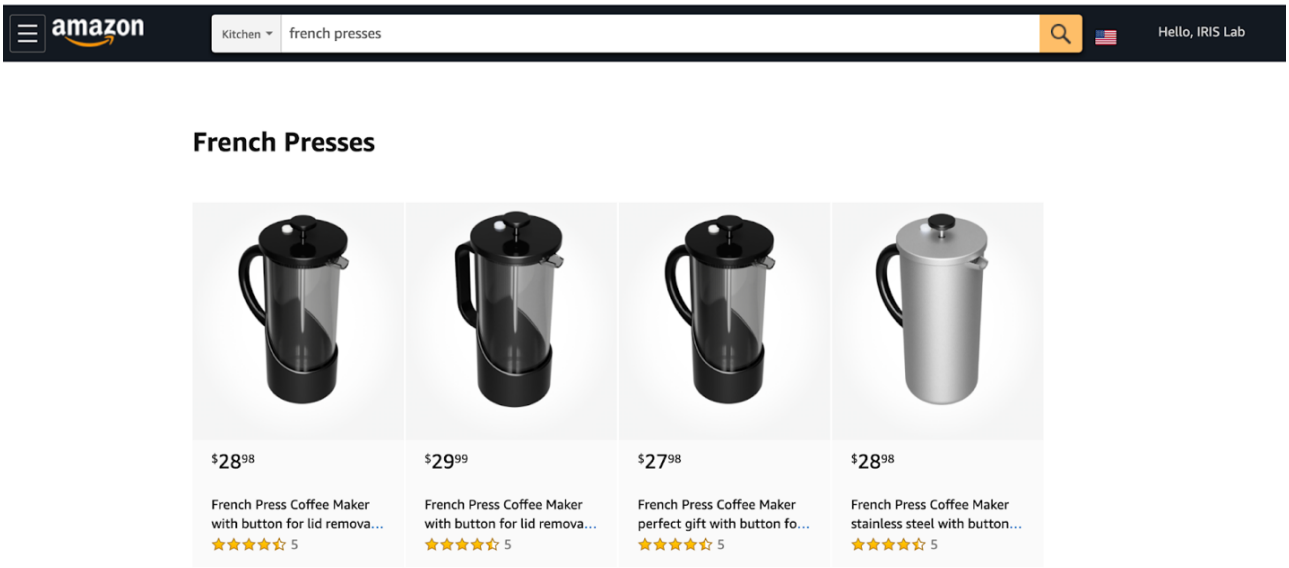

Figure 6.7: Product search page

From the Product Search Page, participants can click on the product's image, price, title, or reviews to access the Product Information Page (Fig. 6.8).

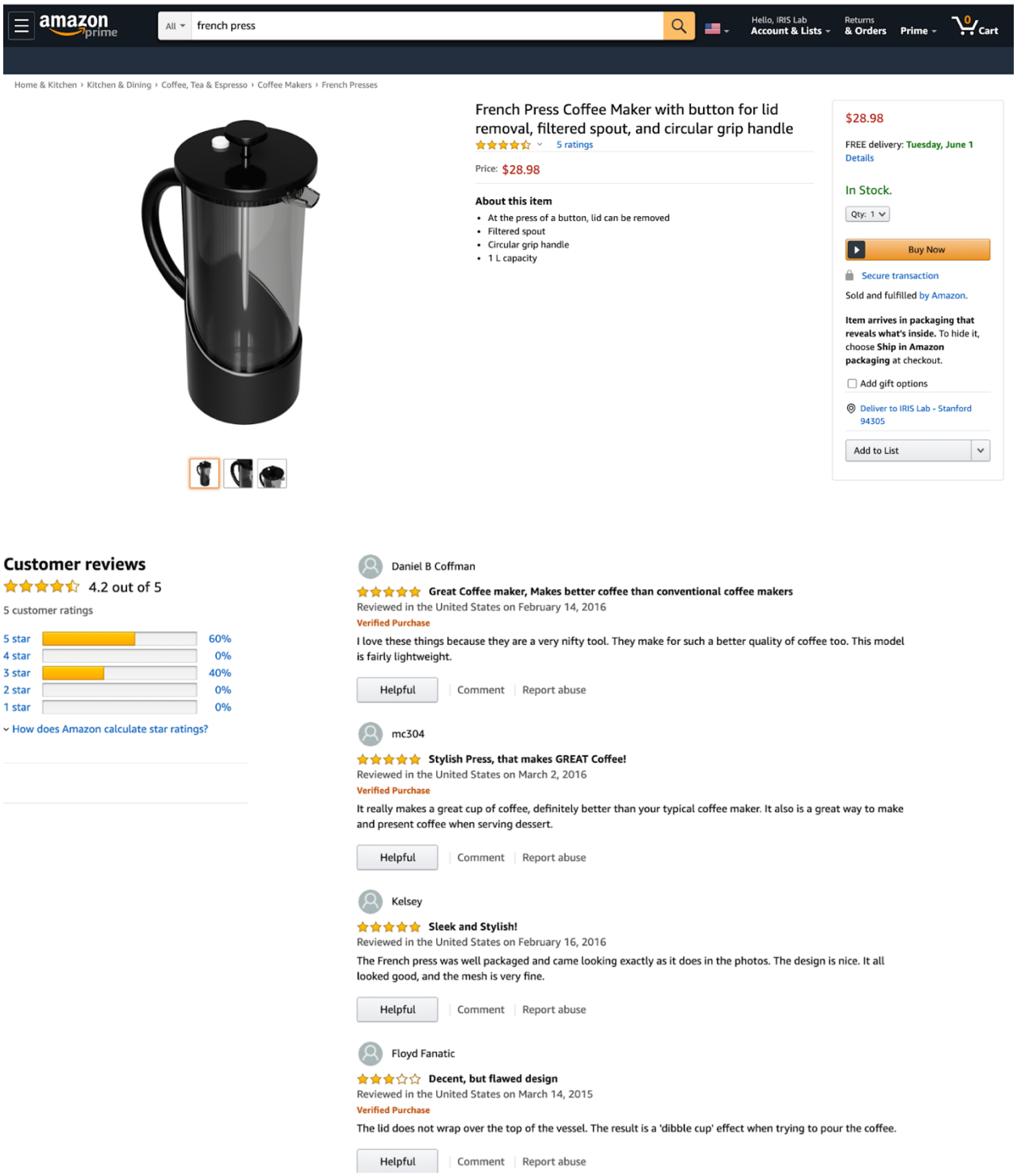

Figure 6.8: Product information page

The product information page provides details on the product's features, five reviews from past consumers, and three images of the French press product. The

participants can zoom in on the image for a closer look at the French press product. All links to external pages were deactivated to prevent the participant from navigating away from our survey. From the Product Page, participants can click on the orange “Buy Now” button to access the Checkout Page or go back to the Product Search Page to read about another product.

The Checkout Page was intended to model the experience of officially purchasing a product on Amazon.com **(**Fig. 6.9**)**. All data entry queries were removed so that the user did not enter any personal information to proceed with buying the French press of their choice. Clicking the “Place your order” button ended the shopping experience.

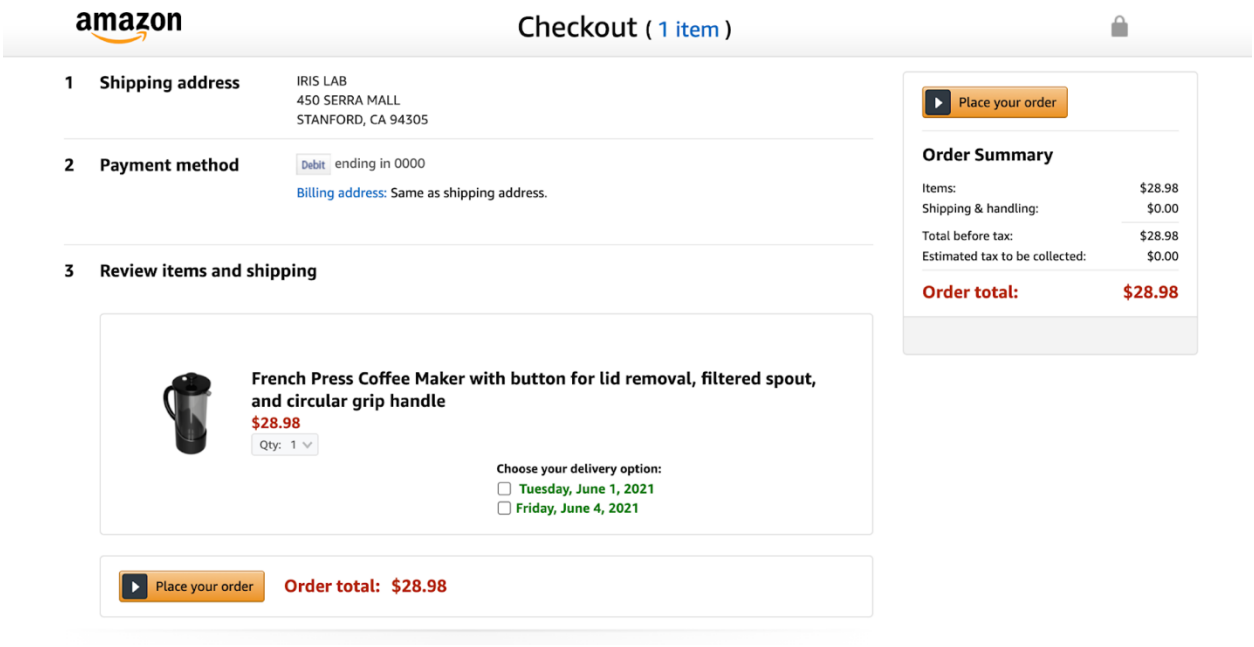

Figure 6.9: Checkout page

6.4.3.2 Normalizing Web Content

We took careful steps to normalize web content between products and control the influence of external variables on purchasing behavior. External variables include

brand, price, number of reviews, review ratings, review content, description content, number of images, and image quality.

Prior to launching the full shopping simulation, we conducted a pilot study to measure the equality of images, descriptions, and reviews between products used in the study. The goal was to control variables so that only product features had a significant difference between products. The pilot study asked participants to indicate their level of agreement on a range of statements using a 5-point Likert scale. For the images, the pilot study asked participants to rate how aesthetically pleasing the images are and their level of quality. For the descriptions, the pilot study asked participants to rate how easy they are to read and understand. For the reviews, the pilot study asked participants to rate their clarity and sentiment. Participants rated images, descriptions, and reviews separately. We modified product information and ran several rounds of the pilot study to achieve no statistically significant differences in ratings between products across the board.

Additional measures we took include controlling for branding and prices. To prevent branding or previous knowledge of a brand from impacting purchasing decisions, we removed brand names from the product titles, descriptions, and images. Product prices had a $2 range between $27.98 and $29.99. We decided on this range so that the different prices can portray a realistic shopping experience, but at the same time have a negligible effect on purchase behavior.

6.4.4 Participants

We recruited a total of 200 participants from MTurk to complete the shopping experience which took 25 minutes on average; participants were compensated $6 each for their time. To create an incentive alignment, we also entered participants into a lottery for a product of similar or less value to the product they chose for purchase in the experiment. We opted to recruit from MTurk instead of in-person participants to accommodate for COVID-19 restrictions and to quickly collect many responses. Moreover, MTurk demographics are likely to match online users and therefore online shopper demographics more closely [34].

To increase the quality of data collected, we screened for participants on MTurk using features on the MTurk platform as well as screening questions in Qualtrics. We required that participants should have a 97% prior approval rating and are based in the United States; literature shows that respondents in the United States tend to deliver better quality responses [33].

Out of the 200 participants that completed their task, we approved 162 based on two requirements: (1) completing the activity in time (t) that is within 1 standard deviation(s) of the average time to complete the activity (μ) or longer (i.e., $t \geqslant \mu - s$) and (2) correctly answering the check question, “What is the capital city of the United States?” which we asked in the post-survey. We did not analyze results if they did not meet one or both criteria. We used similar approval criteria in previous papers [45,81].

6.5 Analysis and Results

This section is split into two parts: first, we present participant data and demographics, and second, we present the shopping experience results that test the hypotheses from section 6.3.

6.5.1 Demographics

The demographics of the 162 approved participants are summarized in Fig. 6.10. Our participants were mostly young, white, educated, working full-time, about 60% male, and making about an average US income. The demographics of our participants are similar to those of the Amazon Mechanical Turk respondents in our previous paper [81] and are in-line with demographic analyses of Amazon Mechanical Turk respondents from literature [33]. While this demographic is not representative of the general US population, it is closer to typical online users and is ideal for studying online purchasing decisions [81].

Figure 6.11 shows the most important factor participants reported for making purchases on Amazon. Reviews, brand, and price were the highest three factors, followed by product description. In our shopping simulation, we normalized all factors besides product description to isolate the influence of different product features on purchasing decisions.

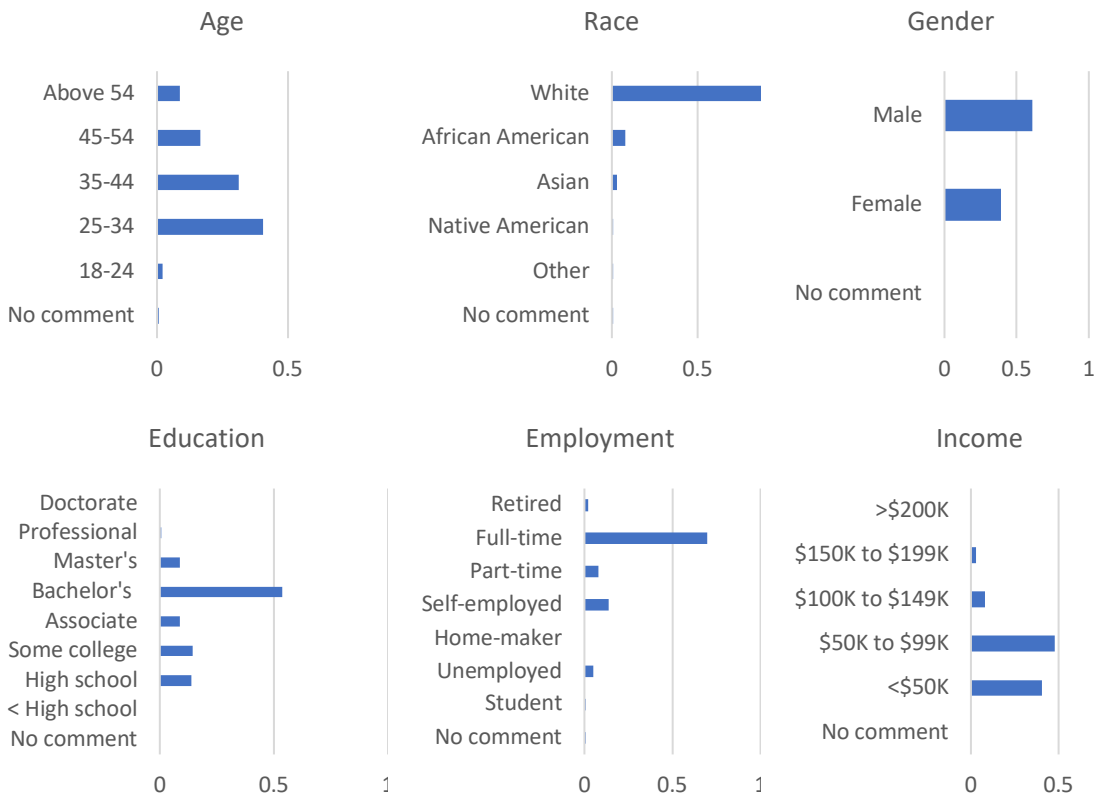

Figure 6.10: Participant demographics

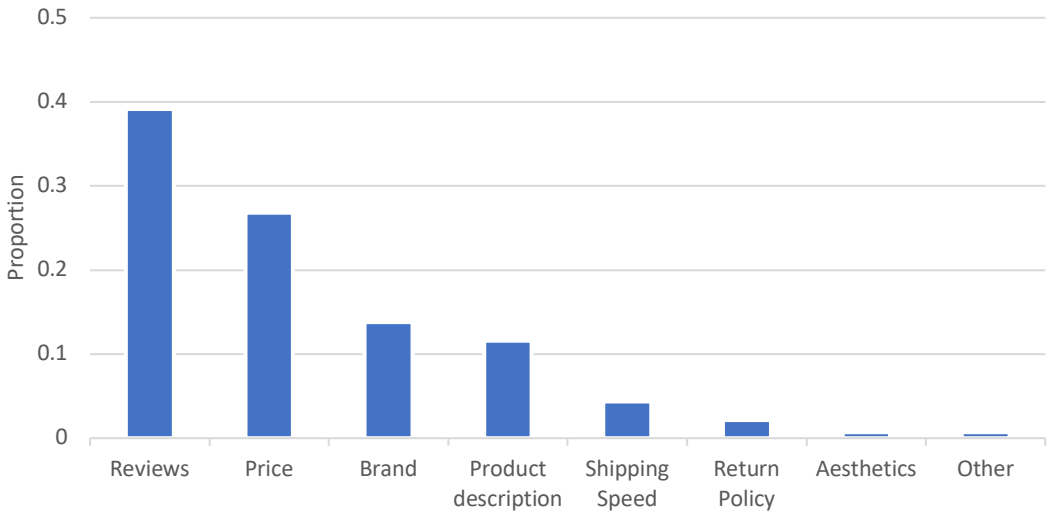

Figure 6.11: Self-reported important factors for purchasing on Amazon by participants

6.5.2 Shopping Simulation

This section is split into three parts: first, we present the results on purchasing decisions that test hypothesis 1 (products with PAS features are more likely to be

purchased than those with dummy features), second, we present the results on desirability ratings that test hypothesis 2a (products with PAS features are rated as more desirable than those with dummy features), and third, we present the results on sustainability ratings that test hypothesis 2b (products with PAS features are rated as more sustainable than those with dummy features).

6.5.2.1 Products Selected for Purchase

Fig. 6.12 shows the raw counts of products selected for purchase in the control and test conditions. More participants selected to purchase products with PAS features in the test condition than products with dummy features in the control condition, suggesting that products with PAS features can drive purchasing decisions.

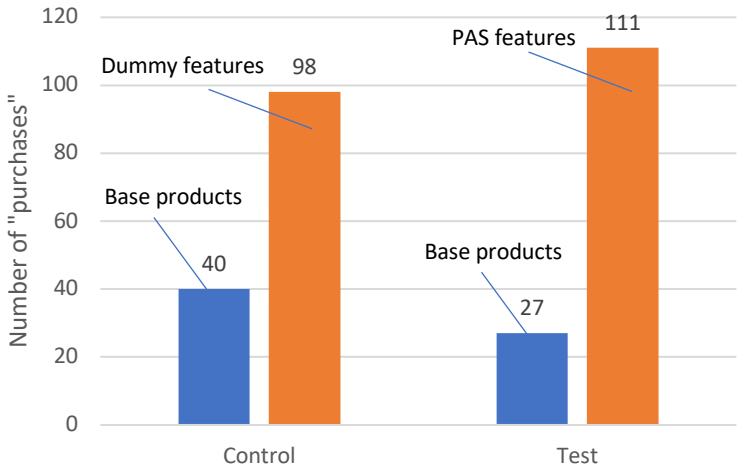

Figure 6.12: Number of purchases for base, dummy, and PAS products

To determine the influence of the experiment conditions relative to the base products, Fig. 6.13 shows the fraction of products selected for purchase with PAS features in the test condition versus with dummy features in the control condition. Approximately 80% of products selected in the test condition were products with PAS

features while 71% of products selected in the control condition where products with dummy features (as opposed to only having base features), supporting hypothesis 1.

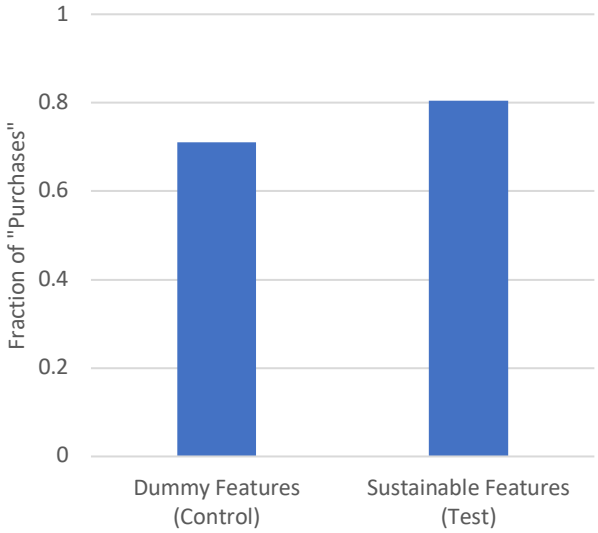

Figure 6.13: Fraction of products selected for purchase in the control condition versus the test condition

We tested if the difference in the fraction of products selected for purchase between the control and test conditions was statistically significant using a t-test, shown in Table 6.4. The difference was significant at the 0.05 level, supporting hypothesis 1. The findings therefore indicate that participants are more likely to select to purchase a product when the description is combined with PAS features rather than dummy features.

Table 6.4: Two sample t-test between control and test conditions for fraction of products selected to purchase
*: significant at p = 0.05, **: significant at p = 0.01, ***: significant at p = 0.001

	Dummy Features (Control)	**Sustainable Features (Test)**
Mean fraction of purchases	0.71	0.80
Variance	0.21	0.16
Observations	162	
P(T<=t) one-tail	0.026*	
t Critical one-tail	1.66	

6.5.2.2 Desirability Ratings

As a proxy for desirability, the survey asked participants to rate their willingness to purchase (WTP) products on a 5-point Likert scale. Figure 6.14 shows the mean willingness to pay ratings for products in the control and test conditions. The WTP for products with PAS features in the test condition is slightly higher than the WTP for products with dummy features in the control condition.

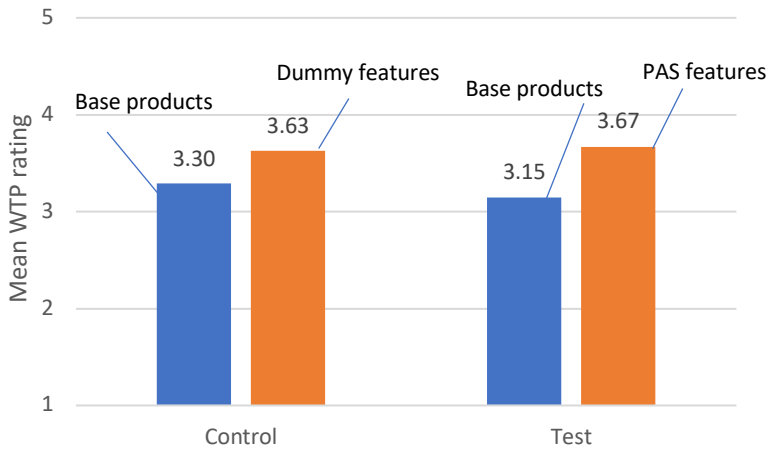

Figure 6.14: Willingness to pay rating for base, dummy, and PAS products

To determine the influence of the experiment conditions on WTP relative to the base products, Fig. 6.15 shows the mean difference in WTP in the test conditions versus the control condition. In the control condition, participants rated WTP products with dummy features 0.34 higher than the base products on a 5-point Likert scale. In the test condition, participants rated WTP products with PAS features 0.52 higher than the base products on a 5-point Likert scale. The greater difference in WTP in the test condition supports hypothesis 2a.

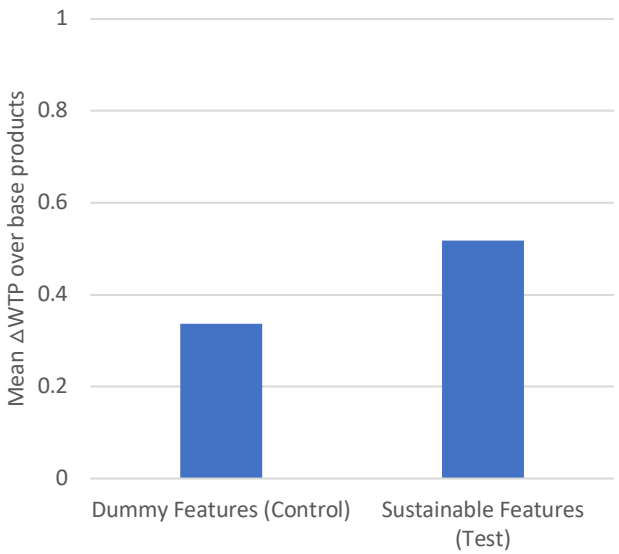

Figure 6.15: Mean ΔWTP in the control condition versus the test condition

The t-test results for mean difference in WTP are included in Table 6.5, showing that the difference between conditions is statistically significant at the 0.05 level. The findings therefore indicate that participants rate products as more desirable when the description is combined with PAS features versus dummy features.

Table 6.5: Two sample t-test between control and test conditions for mean ΔWTP
*: significant at p = 0.05, **: significant at p = 0.01, ***: significant at p = 0.001

	Dummy Features (Control)	**Sustainable Features (Test)**
Mean ΔWTP	0.34	0.52
Variance	1.95	2.23
Observations	324	
P(T<=t) one-tail	0.039*	
t Critical one-tail	1.65	

6.5.2.3 Sustainability Ratings

The survey asked participants to rate products on their sustainability using a 5-point Likert scale. Figure 6.16 shows the mean sustainability ratings for products in the control and test conditions. On average, the sustainability rating for products with PAS

features in the test condition is higher than the sustainability rating for products with dummy features in the control condition.

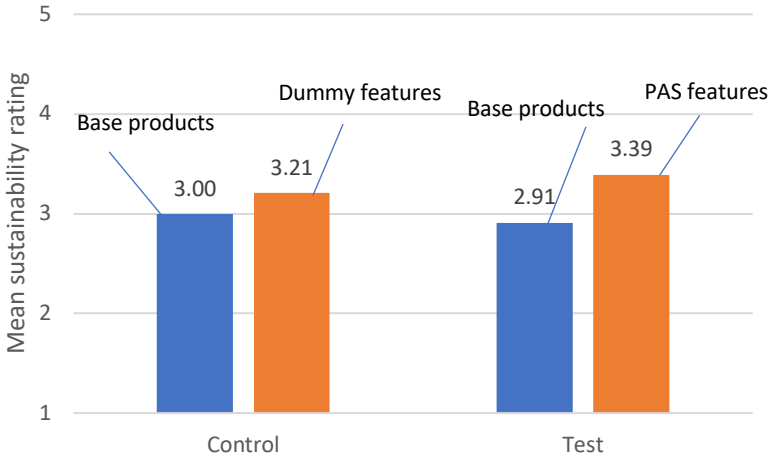

Figure 6.16: Sustainability rating for base, dummy, and PAS products

To determine the influence of the experiment conditions on sustainability ratings relative to the base products, Fig. 6.17 shows the mean difference in sustainability rating of the base products under the control conditions versus the test conditions. In the control conditions, participants rated products with dummy features 0.21 higher than base products on a 5-point Likert scale. In the test conditions, participants rated products with PAS features 0.48 higher than the base products on a 5-point Likert scale. The greater difference in mean sustainability rating in the test condition supports hypothesis 2b.

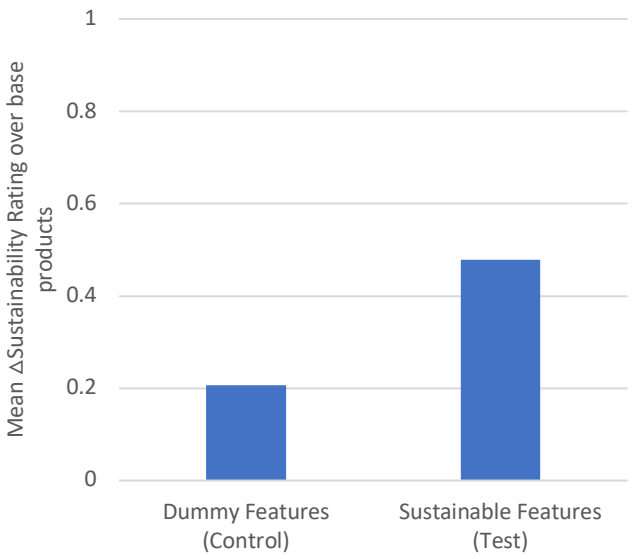

Figure 6.17: Mean ΔSustainability Rating in the control condition versus the test condition

The t-test results are included in Table 6.6 and show that the difference between the two conditions is statistically significant at the 0.001 level. The findings therefore strongly indicate that participants will rate products as more sustainable when the description is combined with PAS features versus dummy features. This also validates our previous work with assessing PAS features using a collage approach [81].

Table 6.6: Two sample t-test between control and test conditions for mean ΔSustainability Rating
*: significant at p = 0.05, **: significant at p = 0.01, ***: significant at p = 0.001

	Dummy Features (Control)	**Sustainable Features (Test)**
Mean ΔSustainability Rating	0.21	0.48
Variance	0.87	1.58
Observations	324	
P(T<=t) one-tail	0.001***	
t Critical one-tail	1.65	

6.6 Discussion

The results provide actionable insights for designers on how to make sustainable products more successful online. The experiment approximated purchase decisions using a simulated shopping experience with incentive alignment and measured both customer preferences and purchase decisions. In this section we discuss the value of using PAS features in design as well as using simulated shopping experiences for customer preference modeling.

First, the results showed that participants selected to purchase products with PAS features more than they did with dummy features (Fig. 6.15 and Table 6.4). It is important to note that PAS features may or may not contribute to engineered sustainability [45]. We demonstrated that, despite the importance of LCAs to inform engineered sustainability features, our proposed method to extract PAS features can drive purchasing decisions for sustainable products.

Second, the results showed that participants are willing to pay more for products with PAS product features compared with dummy features based on 5-point Likert scale ratings (Fig. 6.15 and Table 6.5). The preferences that participants stated in the Likert ratings matched with their purchase decisions in the simulated shopping experience, indicating the value of using simulated shopping experiences in design to model customer preferences. Moreover, the finding supports previous literature that participants are willing to pay more for products they perceive as sustainable [7].

Third, the results showed that participants rated products with PAS features as more sustainable compared with dummy features based on 5-point Likert scale ratings

(Fig. 6.17 and Table 6.6). It is important to note that none of the PAS, dummy, or base features contribute to engineered sustainability. The finding supports our previous work that PAS features resonate with participants as more sustainable [81], emphasizing. the value of using PAS features to communicate product sustainability to customers.

The results in this paper demonstrate that designers should use PAS design features in addition to engineering sustainable features to align sustainable products with customer perceptions. In doing so, designers can create products that are both engineered to be sustainable as well as successful in the marketplace. For example, an LCA might indicate that choosing plastic is a more sustainable manufacturing option [45] but adding some stainless-steel elements to a product might be worth the trade-off to drive online sales. An LCA could determine if dropping plastic entirely in favor of metal can actually be beneficial to the environment, due to the promotion of other "invisible" design features, such as energy-savings or shipping.

With the knowledge that not all PAS features align with engineered sustainability features, it is important to consider the implications for online shopping platforms such as Amazon. The responsibility of making informed purchase decisions ultimately lies on the customer, but Amazon could use the findings in this work to facilitate and guide informed purchase decisions. For example, Amazon could monitor PAS features mentioned in online reviews using natural language processing techniques proposed previously [45]. Moreover, Amazon could allow users to flag reviews that might be spreading misinformation. Ideally, the findings of this work can enable both designers

and e-commerce platforms to build an informed customer base that can bridge the gap between engineered and PAS features, and drive purchases for sustainable products.

There are important limitations to keep in mind with the findings in this paper. First, while we carefully designed the simulated shopping experience to be as realistic as possible, the activity did not involve real purchasing decisions. We included incentive alignments to approximate real purchase decisions, but the results may differ in a real shopping environment with real products. Second, the shopping simulation was a controlled environment with variables kept constant except for the product features. In reality, customers are exposed to varying types of images, descriptions, prices, and reviews when shopping online. The interactions between these variables and how they might influence purchase decisions were not studied in this work. Third, our experiment used French press products as a case-study, building off our previous papers, but does not study purchase decisions for different types of products. We recommend conducting an additional study to investigate the generalizability of our findings, ideally using real products and purchase decisions.

6.7 Conclusions

This paper shows that PAS features can help designers drive purchase decisions for sustainable products. We created a simulated Amazon shopping experience to control what is shown to participants and investigate purchase decisions. We studied how PAS features can influence online purchase decisions compared to dummy features in a within-subject fractional factorial experiment. We built on findings from our previous work where we extracted salient PAS features from online product reviews of

French presses [45] and demonstrated that these features resonate with participants as sustainable despite not contributing to engineered sustainability [81].

During each of the control and test conditions, participants selected a product to purchase from four options: in the control condition we included two base products and two products with dummy features, and in the test condition we included two base products and two products with PAS features. We also asked participants to rate products in terms of willingness to pay and sustainability. The results showed that more participants selected to purchase products with PAS features in the test condition than with dummy features in the control condition. Moreover, participants indicated that they are willing to pay more for products with PAS features and rated them as more sustainable too.

The findings indicate that designers should include both engineered sustainable features (from tools such as an LCA) and PAS features (from our proposed method) to drive purchasing decisions for sustainable products. Moreover, the findings demonstrate the value of conducting online shopping simulations in design research. Next steps for this work include testing the findings in a real purchasing environment as well as testing how the findings generalize with different products.

7. CHAPTER 7
CONCLUSION AND FUTURE PLAN

This dissertation contributes to product design research by providing useful insights on the value of customer perceptions in sustainable design. It presented ways to extract, validate, and use features perceived as sustainable to drive online purchasing decisions. It also presented an approach that future design researchers can use to thoroughly test similar research methods: (1) try with a test case; (2) adapt statistical methods as needed; (3) validate findings with another design method; (4) try method with further test cases; (5) test results in a mock-marketplace. Chapter 2 proposed a method to extract features perceived as sustainable from online reviews, Chapters 3 to 5 investigated metrics and approaches to validate the proposed method, and Chapter 6 tested how features perceived as sustainable can influence online purchase decisions.

In Chapter 2, the study presented a data-driven method for designers to extract features perceived as sustainable from online reviews, demonstrating that a gap exists between perceived and engineered sustainability features. The study shows that features perceived as sustainable can be extracted from online reviews using crowdsourced annotations and natural language processing. With French presses as a case study, different features were extracted for each sustainability aspect. For social aspects, the results showed that positive features were mostly intangible, such as “perfect gift” while negative features were tangible and related to safety, such as “glass cracking”. For environmental aspects, “stainless steel” was identified as a salient positive feature and “plastic “as a negative feature. For economic aspects, positive features related to high product quality and value while negative features related to

poor value. We demonstrated how a crucial engineered sustainability criterion, “energy and water consumption” was not a salient feature perceived as sustainable. The results demonstrated the value of perceived sustainability in design, and the ability of a data-driven approach to extract features perceived as sustainable from online reviews.

In Chapter 3, the study presented inter-rater reliability (IRR) analyses in the context of qualitative text annotations for machine learning datasets. Different implementations of IRR showed minimal improvements in internal validity metrics. The study showed that, for the case of qualitative text annotations, external validity metrics such as precision, recall, and F1 were more robust than internal validity metrics such as IRR. The results highlighted best practices for designers working with highly qualitative annotations such as perceptions in online reviews.

In Chapter 4, the study presented a novel collage approach to validate the extracted features perceived as sustainable from online reviews. Participants placed products on a set of two axes, sustainability and likeability, and selected product features from a dropdown menu. We built off our findings from Chapter 2, using French presses as a case study and leveraging the features we extracted. The results showed that participants selected features perceived as sustainable for products they placed higher on the sustainability axis of the collage. Moreover, the results indicated potential demographic interactions that could influence customer perceptions. Lastly the results showed a significant yet low correlation between perceived sustainability and likeability, indicating that the collage tool can measure both dimensions effectively. The study showed that while features perceived as sustainable may not contribute to engineered

sustainability, participants identified them as sustainable and highlighting their value in sustainable design. The study also demonstrated the value of the collage tool for testing customer perceptions in sustainable design.

In Chapter 5, the study validated the generalizability of the proposed methods in Chapters 2 and 4 by recreating them with different products. The study showed that the findings from our previous study do generalize with limitations. The methods generalize best when the focal products have a balanced set of positive and negative reviews. This is demonstrated by the results from the electric scooter which emulated those of the French press results. We extracted electric scooter features perceived as sustainable from online reviews and tested that participants identified these features as sustainable. Moreover, we found that the gap between perceived and engineered sustainability was closer with electric scooters than with French presses. For baby glass bottles, the findings were not able to generalize due to an imbalance of positive reviews that limited the natural language processing algorithm performance. The study provided an additional validation to our proposed methods for identifying and testing features perceived as sustainable. The insights can help shape design decisions to create sustainable products that align with customer needs.

In Chapter 6, the study presented a shopping simulation approach to investigate how features perceived as sustainable can influence and drive purchasing decisions. Using a within-subject fractional factorial experiment design, the study showed that participants more often selected to purchase products with features perceived as sustainable compared to products with dummy features. The study also validated that

products with perceived sustainable features are rated as more sustainable than products with dummy features, despite none of the features contributing to engineered sustainability concerns. Finally, the study demonstrated that participants are willing to pay more for products with features perceive as sustainable than dummy features. The study provided an ultimate validation approach for identifying practical design benefits of features perceived as sustainable to drive purchasing decisions.

The studies in this dissertation show the value of features perceived as sustainable in sustainable design. To create sustainable products that align with customer needs, they must consider both engineered and perceived sustainability requirements. This dissertation outlines methods to extract, test, and validate features perceived as sustainable for driving online purchase decisions.

It's important to address ethical concerns and implications of this work. Although this research has focused on the significance of both engineered sustainability and perceived sustainability, the results demonstrate that the two might not always be aligned in practice. If used with malintent, the findings of this work could be used to create products that customers perceive are sustainable but are not in reality. The intent of this research, however, is to shed light on the difference between perceived sustainability and engineered sustainability. It is up to the designers and sellers to encourage ethical practices when designing their products. Similarly, consumers should be aware that there can be a disconnect between perceived sustainability and engineered sustainability. This research benefits consumers by helping them make more

informed decisions about their purchases, since it is not always the case that a product they perceive as sustainable is sustainable.

For future work, I plan to first enhance the method proposed in Chapter 2 for extracting features perceived as sustainable from online reviews using crowdsourced annotations and a natural language processing algorithm. It is important this method functions more robustly against an imbalance of positive and negative reviews. In doing so, the method can better generalize effectively with products such as a baby glass bottle that tend to have overly positive reviews on Amazon, as identified in Chapter 5. Second, I plan to investigate the demographic interaction results in Chapter 4. Understanding how demographic plays a role in customer perceptions is crucial to create sustainable products that are representative and align with customer needs across different demographics. Third, I plan to build on the purchasing study in Chapter 6 by moving it beyond a simulation and into investigating real world purchase decisions. This will provide a stronger validation and motivation to include features perceived as sustainable in product design. Lastly, I aim to investigate how engineered and perceived features interact to create sustainable products. Ultimately, it is important that a sustainable product includes both. The findings of this work will likely boost sustainable products by enabling designers to create sustainable products that are truly sustainable bust also align with customer needs.

REFERENCES

[1] Visentin, C., Trentin, A. W. da S., Braun, A. B., and Thomé, A., 2020, "Life Cycle Sustainability Assessment: A Systematic Literature Review through the Application Perspective, Indicators, and Methodologies," Journal of Cleaner Production, **270**, p. 122509.

[2] She, J., and MacDonald, E. F., 2017, "Exploring the Effects of a Product's Sustainability Triggers on Pro-Environmental Decision-Making," Journal of Mechanical Design, **140**(1), p. 011102.

[3] Slimak, M. W., and Dietz, T., 2006, "Personal Values, Beliefs, and Ecological Risk Perception," Risk Analysis, **26**(6), pp. 1689–1705.

[4] Lewin, T., and Borroff, R., 2010, *How to Design Cars like a Pro*, Motorbooks, Minneapolis, MN.

[5] Gartman, D., 1994, *Auto Opium: A Social History of American Automobile Design*, Routledge, London ; New York.

[6] Reid, T. N., Frischknecht, B. D., and Papalambros, P. Y., 2012, "Perceptual Attributes in Product Design: Fuel Economy and Silhouette-Based Perceived Environmental Friendliness Tradeoffs in Automotive Vehicle Design," Journal of Mechanical Design, **134**(4), p. 041006.

[7] McCaskill, A., 2015, "Consumer-Goods' Brands That Demonstrate Commitment To Sustainability Outperform Those That Don't," Nielsen [Online]. Available: https://www.nielsen.com/us/en/press-room/2015/consumer-goods-brands-that-demonstrate-commitment-to-sustainability-outperform.html.

[8] She, J., and MacDonald, E. F., 2017, "Exploring the Effects of a Product's Sustainability Triggers on Pro-Environmental Decision-Making," Journal of Mechanical Design, **140**(1), p. 011102.

[9] Kim, E.-H., and Lyon, T. P., 2015, "Greenwash vs. Brownwash: Exaggeration and Undue Modesty in Corporate Sustainability Disclosure," Ssrn, (December 2018).

[10] 2018, "Quarterly Share of E-Commerce Sales of Total U.S. Retail Sales from 1st Quarter 2010 to 3rd Quarter 2018," Statista [Online]. Available: https://www.statista.com/statistics/187439/share-of-e-commerce-sales-in-total-us-retail-sales-in-2010/.

[11] Roghanizad, M. M., and Neufeld, D. J., 2015, "Intuition, Risk, and the Formation of Online Trust," Computers in Human Behavior, **50**, pp. 489–498.

[12] MacDonald, E. F., Gonzalez, R., and Papalambros, P. Y., 2009, "Preference Inconsistency in Multidisciplinary Design Decision Making," Journal of Mechanical Design, **131**(3), p. 031009.

[13] Ren, Y., Burnap, A., and Papalambros, P., 2013, "Quantification of Perceptual Design Attributes Using a Crowd," *International Conference on Engineering Design*, pp. 1–9.

[14] Engström, P., and Forsell, E., 2018, "Demand Effects of Consumers' Stated and Revealed Preferences," Journal of Economic Behavior & Organization, **150**, pp. 43–61.

[15] Netzer, O., Toubia, O., Bradlow, E. T., Dahan, E., Evgeniou, T., Feinberg, F. M., Feit, E. M., Hui, S. K., Johnson, J., Liechty, J. C., Orlin, J. B., and Rao, V. R., 2008,

"Beyond Conjoint Analysis: Advances in Preference Measurement," Marketing Letters, **19**(3/4,), pp. 337–354.
[16] Decker, R., and Trusov, M., 2010, "Estimating Aggregate Consumer Preferences from Online Product Reviews," International Journal of Research in Marketing, **27**(4), pp. 293–307.
[17] Qiao, Z., Wang, G. A., Zhou, M., and Fan, W., 2017, "The Impact of Customer Reviews on Product Innovation: Empirical Evidence in Mobile Apps," *Analytics and Data Science*, Springer, Cham, pp. 95–110.
[18] Liu, Y., Jin, J., Ji, P., Harding, J. A., and Fung, R. Y. K., 2013, "Identifying Helpful Online Reviews: A Product Designer's Perspective," CAD Computer Aided Design, **45**(2), pp. 180–194.
[19] Rai, R., 2012, "Identifying Key Product Attributes and Their Importance Levels From Online Customer Reviews," Proceedings of the ASME 2012 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference, **70493**(August), pp. 1–8.
[20] Stone, T., and Choi, S.-K., 2013, "Extracting Customer Preference from User-Generated Content Sources Using Classification," *Proceedings of the ASME 2013 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference*, ASME, Portland, pp. 1–9.
[21] Singh, A. S., and Tucker, C. S., 2015, "Investigating the Heterogeneity of Product Feature Preferences Mined Using Online Product Data Streams," *Volume 2B: 41st Design Automation Conference*, ASME, Boston, Massachusetts, USA, p. V02BT03A020.
[22] Singh, A., and Tucker, C. S., 2017, "A Machine Learning Approach to Product Review Disambiguation Based on Function, Form and Behavior Classification," Decision Support Systems, **97**(2016), pp. 81–91.
[23] Kataria, S., Mitra, P., and Bhatia, S., 2010, "Utilizing Context in Generative Bayesian Models for Linked Corpus," Aaai, **10**(Hofmann 1999), p. 1.
[24] Krestel, R., Fankhauser, P., and Nejdl, W., 2009, "Latent Dirichlet Allocation for Tag Recommendation," Proceedings of the third ACM conference on Recommender systems - RecSys '09, (May 2014), p. 61.
[25] Tuarob, S., Pouchard, L. C., and Giles, C. L., 2013, "Automatic Tag Recommendation for Metadata Annotation Using Probabilistic Topic Modeling," Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries - JCDL '13, p. 239.
[26] Tuarob, S., Pouchard, L. C., Noy, N., Horsburgh, J. S., and Palanisamy, G., 2012, "ONEMercury: Towards Automatic Annotation of Environmental Science Metadata," CEUR Workshop Proceedings, **951**, pp. 1–12.
[27] Zhang, X., and Mitra, P., 2010, "Learning Topical Transition Probabilities in Click through Data with Regression Models," Procceedings of the 13th International Workshop on the Web and Databases - WebDB '10, p. 1.
[28] Tuarob, S., and Tucker, C. S., 2015, "Automated Discovery of Lead Users and Latent Product Features by Mining Large Scale Social Media Networks," Journal of Mechanical Design, **137**(7), p. 071402.
[29] Tuarob, S., and Tucker, C. S., 2015, "A Product Feature Inference Model for Mining Implicit Customer Preferences Within Large Scale Social Media Networks,"

Volume 1B: 35th Computers and Information in Engineering Conference, p. V01BT02A002.
[30] Thelwall, M., Buckley, K., Paltoglou, G., and Cai, D., 2010, "Sentiment Strength Detection in Short Informal Text," The American Society for Informational science and technology, **61**(12), pp. 2544–2558.
[31] Wang, W. M., Li, Z., Tian, Z. G., Wang, J. W., and Cheng, M. N., 2018, "Extracting and Summarizing Affective Features and Responses from Online Product Descriptions and Reviews: A Kansei Text Mining Approach," Engineering Applications of Artificial Intelligence, **73**(February), pp. 149–162.
[32] Nagamachi, M., and Imada, A. S., 1995, "Kansei Engineering: An Ergonomic Technology for Product Development," International Journal of Industrial Ergonomics, **15**(1), p. 1.
[33] Paolacci, G., and Chandler, J., 2014, "Inside the Turk: Understanding Mechanical Turk as a Participant Pool," Current Directions in Psychological Science, **23**(3), pp. 184–188.
[34] Goodman, J. K., and Paolacci, G., 2017, "Crowdsourcing Consumer Research," Journal of Consumer Research, **44**(1), pp. 196–210.
[35] Jurafsky, D., "N-Gram Language Models," *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*.
[36] Blei, D. M., "Latent Dirichlet Allocation," p. 30.
[37] James, G., Witten, D., Haste, T., and Tibshirani, R., 2006, *An Introduction to Statistical Learning with Applications in R*.
[38] MacDonald, E. F., and She, J., 2015, "Seven Cognitive Concepts for Successful Eco-Design," Journal of Cleaner Production, **92**, pp. 23–36.
[39] Meinel, C., Leifer, L., and Plattner, H., eds., 2011, *Design Thinking*, Springer Berlin Heidelberg, Berlin, Heidelberg.
[40] Stone, T., and Choi, S.-K., 2013, "Extracting Consumer Preference From User-Generated Content Sources Using Classification," *Volume 3A: 39th Design Automation Conference*, ASME, Portland, Oregon, USA, p. V03AT03A031.
[41] Kudrowitz, B. M., and Wallace, D. R., 2010, "Assessing the Quality of Ideas From Prolific, Early-Stage Product Ideation," *Volume 5: 22nd International Conference on Design Theory and Methodology; Special Conference on Mechanical Vibration and Noise*, ASMEDC, Montreal, Quebec, Canada, pp. 381–391.
[42] Kwon, J., and Kudrowitz, B., 2019, "The Sketch Quality Bias: Evaluating Descriptions of Product Ideas With and Without Visuals," p. 9.
[43] Kilem Li Gwet, 2014, *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuing the Extent of Agreement Among Raters*, Advanced Analytics.
[44] Toh, C. A., Miller, S. R., and Okudan Kremer, G. E., 2014, "The Impact of Team-Based Product Dissection on Design Novelty," Journal of Mechanical Design, **136**(4), p. 041004.
[45] El Dehaibi, N., Goodman, N. D., and MacDonald, E. F., 2019, "Extracting Customer Perceptions of Product Sustainability From Online Reviews," Journal of Mechanical Design, **141**(12), p. 121103.
[46] Hallgren, K. A., 2012, "Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial," TQMP, **8**(1), pp. 23–34.

[47] Cohen, J., 1960, “A Coefficient of Agreement for Nominal Scales,” Educational and Psychological Measurement, **20**(1), pp. 37–46.
[48] Kennedy, L. E., Hosig, K. L., Ju, Y., and Serrano, E. L., 2019, “Evaluation of a Mindfulness-Based Stress Management and Nutrition Education Program for Mothers,” Cogent Social Sciences, **5**(1), p. 1682928.
[49] Liang, Y., Kirilenko, A. P., Stepchenkova, S. O., and Ma, S. (David), 2020, “Using Social Media to Discover Unwanted Behaviours Displayed by Visitors to Nature Parks: Comparisons of Nationally and Privately Owned Parks in the Greater Kruger National Park, South Africa,” Tourism Recreation Research, **45**(2), pp. 271–276.
[50] Fleiss, J. L., 1971, “Measuring Nominal Scale Agreement among Many Raters.,” Psychological Bulletin, **76**(5), pp. 378–382.
[51] Rash, J. A., Prkachin, K. M., Solomon, P. E., and Campbell, T. S., 2019, “Assessing the Efficacy of a Manual-based Intervention for Improving the Detection of Facial Pain Expression,” Eur J Pain, **23**(5), pp. 1006–1019.
[52] Lai, V. K. W., Li, J. C.-H., and Lee, A., 2019, “Psychometric Validation of the Chinese Patient- and Family Satisfaction in the Intensive Care Unit Questionnaires,” Journal of Critical Care, **54**, pp. 58–64.
[53] krippendorff, K., 2004, “Measuring the Reliability of Qualitative Text Analysis Data,” Qual Quant, **38**(6), pp. 787–800.
[54] Stab, C., and Gurevych, I., 2014, “Identifying Argumentative Discourse Structures in Persuasive Essays,” *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Doha, Qatar, pp. 46–56.
[55] Card, D., Boydstun, A. E., Gross, J. H., Resnik, P., and Smith, N. A., 2015, “The Media Frames Corpus: Annotations of Frames Across Issues,” *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Association for Computational Linguistics, Beijing, China, pp. 438–444.
[56] de Medeiros, J. F., Ribeiro, J. L. D., and Cortimiglia, M. N., 2016, “Influence of Perceived Value on Purchasing Decisions of Green Products in Brazil,” Journal of Cleaner Production, **110**, pp. 158–169.
[57] Sheth, J. N., Sethia, N. K., and Srinivas, S., 2011, “Mindful Consumption: A Customer-Centric Approach to Sustainability,” J. of the Acad. Mark. Sci., **39**(1), pp. 21–39.
[58] Petersen, M., and Brockhaus, S., 2017, “Dancing in the Dark: Challenges for Product Developers to Improve and Communicate Product Sustainability,” Journal of Cleaner Production, **161**, pp. 345–354.
[59] Gary Levin, 1993, “Too Green for Their Own Good,” Advertising Age.
[60] O’Rourke, D., and Ringer, A., 2016, “The Impact of Sustainability Information on Consumer Decision Making: Impact of Sustainability Information on Consumers,” Journal of Industrial Ecology, **20**(4), pp. 882–892.
[61] Joung, J., and Kim, H. M., 2021, “Automated Keyword Filtering in Latent Dirichlet Allocation for Identifying Product Attributes From Online Reviews,” Journal of Mechanical Design, **143**(8), p. 084501.

[62] Hou, T., Yannou, B., Leroy, Y., and Poirson, E., 2019, "Mining Changes in User Expectation Over Time From Online Reviews," Journal of Mechanical Design, **141**(9), p. 091102.
[63] E. Bruce Goldstein and James R. Brockmole, 2017, "Introduction to Perception," *Sensation & Perception*, Cengage Learning.
[64] Papista, E., and Krystallis, A., 2013, "Investigating the Types of Value and Cost of Green Brands: Proposition of a Conceptual Framework," J Bus Ethics, **115**(1), pp. 75–92.
[65] MacDonald, E. F., Gonzalez, R., and Papalambros, P., 2009, "The Construction of Preferences for Crux and Sentinel Product Attributes," Journal of Engineering Design, **20**(6), pp. 609–626.
[66] Borin, N., Cerf, D. C., and Krishnan, R., 2011, "Consumer Effects of Environmental Impact in Product Labeling," Journal of Consumer Marketing, **28**(1), pp. 76–86.
[67] Maccioni, L., Borgianni, Y., and Basso, D., 2019, "Value Perception of Green Products: An Exploratory Study Combining Conscious Answers and Unconscious Behavioral Aspects," Sustainability, **11**(5), p. 1226.
[68] Steenis, N. D., van Herpen, E., van der Lans, I. A., Ligthart, T. N., and van Trijp, H. C. M., 2017, "Consumer Response to Packaging Design: The Role of Packaging Materials and Graphics in Sustainability Perceptions and Product Evaluations," Journal of Cleaner Production, **162**, pp. 286–298.
[69] Catlin, J. R., Luchs, M. G., and Phipps, M., 2017, "Consumer Perceptions of the Social Vs. Environmental Dimensions of Sustainability," Journal of Consumer Policy, **40**(3), pp. 245–277.
[70] Rai, R., 2012, "Identifying Key Product Attributes and Their Importance Levels From Online Customer Reviews," *Volume 3: 38th Design Automation Conference, Parts A and B*, ASME, Chicago, Illinois, USA, p. 533.
[71] Amis Guyton, A., "Developing Sustainable Product Semantics for Consumer Products: A Sustainable Designer's Guide," Georgia Institute of Technology.
[72] Liao, T., Tanner, K., and MacDonald, E., 2019, "Revealing Insights of Users' Perceptions: An Approach to Evaluate Wearable Products Based on Emotions," Proceedings of the Design Society: International Conference on Engineering Design, **1**(1), pp. 3969–3978.
[73] HD Delaney, 2003, "Higher Order Designs WIth Within-Subjects Factors: The Multivariate Approach," *Designing Experiments and Analyzing Data*.
[74] Bakdash, J. Z., and Marusich, L. R., 2017, "Repeated Measures Correlation," Front. Psychol., **8**, p. 456.
[75] Luchs, M. G., Naylor, R. W., Irwin, J. R., and Raghunathan, R., "The Sustainability Liability: Potential Negative Effects of Ethicality on Product Preference," p. 14.
[76] Gao, J., Zhang, C., Wang, K., and Ba, S., 2012, "Understanding Online Purchase Decision Making: The Effects of Unconscious Thought, Information Quality, and Information Quantity," Decision Support Systems, **53**(4), pp. 772–781.
[77] Zhang, K. Z. K., Zhao, S. J., Cheung, C. M. K., and Lee, M. K. O., 2014, "Examining the Influence of Online Reviews on Consumers' Decision-Making: A Heuristic–Systematic Model," Decision Support Systems, **67**, pp. 78–89.

[78] Kordzadeh, N., 2019, "Investigating Bias in the Online Physician Reviews Published on Healthcare Organizations' Websites," Decision Support Systems, **118**, pp. 70–82.
[79] Lu, X., He, S., Lian, S., Ba, S., and Wu, J., 2020, "Is User-Generated Content Always Helpful? The Effects of Online Forum Browsing on Consumers' Travel Purchase Decisions," Decision Support Systems, **137**, p. 113368.
[80] Kim, E.-H., and Lyon, T. P., 2015, "Greenwash vs. Brownwash: Exaggeration and Undue Modesty in Corporate Sustainability Disclosure," Organization Science, **26**(3), pp. 705–723.
[81] El-Dehaibi, N., Liao, T., and MacDonald, E. F., 2021, "Validating Perceived Sustainable Design Features Using a Novel Collage Approach," Proceedings of the ASME 2021 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference, August 17-19, Online.
[82] Wang, Z., Li, H., Ye, Q., and Law, R., 2016, "Saliency Effects of Online Reviews Embedded in the Description on Sales: Moderating Role of Reputation," Decision Support Systems, **87**, pp. 50–58.
[83] Maslowska, E., Malthouse, E. C., and Viswanathan, V., 2017, "Do Customer Reviews Drive Purchase Decisions? The Moderating Roles of Review Exposure and Price," Decision Support Systems, **98**, pp. 1–9.
[84] von Helversen, B., Abramczuk, K., Kopeć, W., and Nielek, R., 2018, "Influence of Consumer Reviews on Online Purchasing Decisions in Older and Younger Adults," Decision Support Systems, **113**, pp. 1–10.
[85] Nysveen, H., and Pedersen, P. E., 2004, "An Exploratory Study of Customers' Perception of Company Web Sites Offering Various Interactive Applications: Moderating Effects of Customers' Internet Experience," Decision Support Systems, **37**(1), pp. 137–150.
[86] Li, X., Zhuang, Y., Lu, B., and Chen, G., 2019, "A Multi-Stage Hidden Markov Model of Customer Repurchase Motivation in Online Shopping," Decision Support Systems, **120**, pp. 72–80.
[87] Zhang, L., Yan, Q., and Zhang, L., 2020, "A Text Analytics Framework for Understanding the Relationships among Host Self-Description, Trust Perception and Purchase Behavior on Airbnb," Decision Support Systems, **133**, p. 113288.
[88] Liu, Y., Jiang, C., and Zhao, H., 2019, "Assessing Product Competitive Advantages from the Perspective of Customers by Mining User-Generated Content on Social Media," Decision Support Systems, **123**, p. 113079.
[89] Sun, X., Han, M., and Feng, J., 2019, "Helpfulness of Online Reviews: Examining Review Informativeness and Classification Thresholds by Search Products and Experience Products," Decision Support Systems, **124**, p. 113099.
[90] James, G., Witten, D., Hastie, T., and Tibshirani, R., 2013, *An Introduction to Statistical Learning*, Springer New York, New York, NY.
[91] Johnstone, M.-L., and Tan, L. P., 2015, "Exploring the Gap Between Consumers' Green Rhetoric and Purchasing Behaviour," J Bus Ethics, **132**(2), pp. 311–328.
[92] MacDonald, E. F., Gonzalez, R., and Papalambros, P. Y., 2009, "Preference Inconsistency in Multidisciplinary Design Decision Making," Journal of Mechanical Design, **131**(3), p. 031009.

[93] Green, P. E., and Rao, V. R., 1971, "Conjoint Measurement for Quantifying Judgmental Data," Journal of Marketing Research, **8**(3), p. 355.
[94] Suryadi, D., and Kim, H. M., 2019, "A Data-Driven Methodology to Construct Customer Choice Sets Using Online Data and Customer Reviews," Journal of Mechanical Design, **141**(11), p. 111103.
[95] Goucher-Lambert, K., Moss, J., and Cagan, J., 2017, "Inside the Mind: Using Neuroimaging to Understand Moral Product Preference Judgments Involving Sustainability," Journal of Mechanical Design, **139**(4), p. 041103.
[96] Goucher-Lambert, K., and Cagan, J., 2015, "The Impact of Sustainability on Consumer Preference Judgments of Product Attributes," Journal of Mechanical Design, **137**(8), p. 081401.
[97] Tovares, N., Boatwright, P., and Cagan, J., 2014, "Experiential Conjoint Analysis: An Experience-Based Method for Eliciting, Capturing, and Modeling Consumer Preference," Journal of Mechanical Design, **136**(10), p. 101404.
[98] Chevalier, J. A., and Mayzlin, D., 2006, "The Effect of Word of Mouth on Sales: Online Book Reviews," Journal of Marketing Research, **43**(3), pp. 345–354.
[99] CHEN, Y., WANG, Q., and XIE, J., 2011, "Online Social Interactions: A Natural Experiment on Word of Mouth Versus Observational Learning," Journal of Marketing Research, **48**(2), pp. 238–254.
[100] Liu, Y., 2006, "Word of Mouth for Movies: Its Dynamics and Impact on Box Office Revenue," Journal of Marketing, **70**(3), pp. 74–89.
[101] Dhar, V., and Chang, E. A., 2009, "Does Chatter Matter? The Impact of User-Generated Content on Music Sales," Journal of Interactive Marketing, **23**(4), pp. 300–307.