Extracting Customer Perceptions of Product Sustainability From Online Reviews

Nasreddine El Dehaibi
Department of Mechanical Engineering,
Stanford University,
Stanford, CA 94305
e-mail: ndehaibi@stanford.edu

Noah D. Goodman
Associate Professor
Department of Psychology and Department of Computer Science,
Stanford University,
Stanford, CA 94305
e-mail: ngoodman@stanford.edu

Erin F. MacDonald
Assistant Professor
Department of Mechanical Engineering,
Stanford University,
Stanford, CA 94305
e-mail: erinmacd@stanford.edu

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 versus perception of sustainability can be quite different. This paper details a design method to identify perceptions of sustainable features (PerSFs) by collecting online reviews, manually annotating them using crowdsourced 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, like sharp corners. For environmental aspects, positive PerSFs are associated with reliable materials like 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, like 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. [DOI: 10.1115/1.4044522]

Keywords: design automation, sustainable design, customer perceptions, online reviews

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 [1], 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 [2]. Customers are also skeptical of ecolabels due to misleading marketing strategies, or "greenwashing" [3]. Designers can communicate sustainability through subtle cues in 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" [2]. 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 toward online outlets with e-commerce sales making up 9.6% of total retail sales as of 2018, up from 4.2% in 2010 [4]. Roghanizad and Neufeld show that online customers tend to rely more on intuition than rational judgment when making purchasing decisions due to the higher risk of buying a product before seeing it [5]. The authors use an online book store shopping simulation with the 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 who 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 [6]. This is a large problem for sustainable product assessment. Moreover, stated preference methods are time-intensive, prone to other biases like 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. 1; each review provides different PerSFs of the product. For example, features like 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.

This study uses online reviews to identify PerSFs and to determine which of these features have positive and negative sentiments. 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: Sec. 2 presents a brief background on the use of online reviews in design, Sec. 3 presents a

This paper was also submitted to the 2019 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference (IDETC/CIE 2019).

Contributed by the Design Theory and Methodology Committee of ASME for publication in the Journal of Mechanical Design. Manuscript received March 1, 2019; final manuscript received July 13, 2019; published online September 4, 2019. Assoc. Editor: Conrad Tucker.

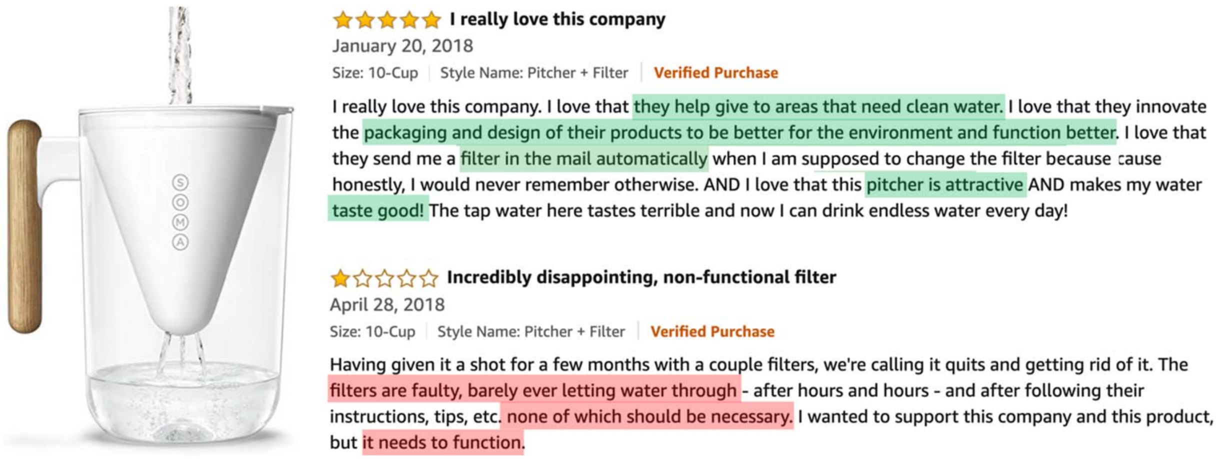

Fig. 1 Example of a product review from Amazon. Highlights in the top review indicate positive, bottom indicates negative sentiment.

literature review on NLP research, Sec. 4 describes the method used to build a machine learning model, Secs. 5 and 6 show the results and analysis, findings are discussed in Sec. 7, and conclusions are made in Sec. 8.

2 Background

In this section, we present a brief 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 Sec. 3.

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. showed that crowdsourcing can be used to capture perceptions of design features [7]. They recruited respondents from Amazon Mechanical Turk (MTurk) to assess the 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 a stated preference because they differentiate online reviewers from users who bought a product [8]. Netzer et al. considered online reviews as revealed preference that can be used as an auxiliary input to stated preference data [9]. In reality, 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 Sec. 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 like surveys. Decker and Trusov demonstrate this using reviews for mobile phones [10]. Online reviews are also a source of product innovation for designers. Qiao et al. examined the 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 [11]. Reviews, therefore, provide more than just a word-of-mouth effect and provide valuable information for designers.

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 Sec. 4.1 for more information).

Furthermore, customers perceive the 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 [12]. The authors find a weak correlation between the two with a 35.3% mean average error and 29.5% root mean square error. This suggests that there is a gap in perceptions for the 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.

3 Developments in Natural Language Processing 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.

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 [13]. He extracted key product features from reviews for a camcorder from epinions.com using a term-document matrix and part-of-speech (POS) tagger. Stop-words were removed from the reviews and words were stemmed. A weighted metric took into account the rate of occurrences of product features in the reviews to measure the importance of a feature. When compared with 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 [14]. The authors used a three-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 (BOW) 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 [15]. "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 "lightweight" and "WiFi," while the "deal breakers" included "battery," "screen," and "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 [16]. 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 [17–21]. 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 [22]. 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.

3.2 Extracting Implicit Customer Perceptions From Online Reviews. Implicit perceptions are phrases like, "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 [23]. To develop the co-occurrence network the authors first extracted explicit product features using a POS tagger. Sentiment extraction was performed using SentiStrength [24]. A co-word network was then generated where the nodes are ranked in order to translate the implicit message into an explicit form. The authors used Twitter data, comprising 390,000 Twitter messages of 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 [25]. Kansei engineering is a product development process that quantifies relationships between affective responses and design features [26]. 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 like 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.

4 Method

The method described in this study combines research from identifying customer perceptions, rating design ideas, and natural language processing (Fig. 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. To the best of our knowledge, this is the first rating method introduced in the design research space for evaluating customer perceptions in product reviews. Specifically, we look toward evaluating customer perceptions of sustainability, a multifaceted and abstract concept, from reviews and using natural language processing to extract sustainable value for designers at a large scale.

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 4.1 and 4.2 cover the method associated with identifying customer perceptions and rating design ideas from Fig. 2 while Secs. 4.3 and 4.4 explain the method associated with natural language processing from Fig. 2. A simplified chronological representation of the steps we took is shown in Fig. 3.

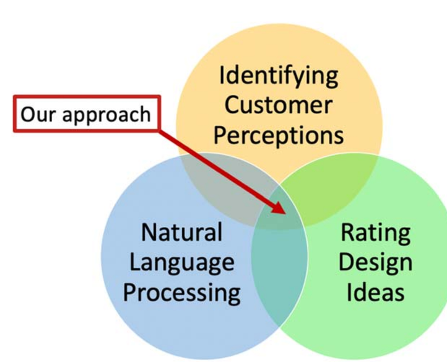

Fig. 2 High-level overview of method topics

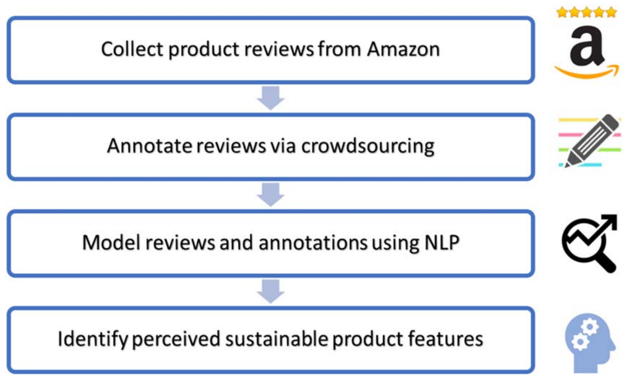

Fig. 3 Chronological method flow

We used supervised learning techniques based on the logistic classification to model the reviews. Each of the steps in Fig. 3 are explained below.

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 the 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 contains 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 Sec. 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.

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 Sec. 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 Sec. 4.3.2).

4.2.1 Survey Design. The survey consists of three versions in order to be customized for each sustainability aspect (social, environmental, economic). We distributed a total of 900 annotators evenly across each version, see Fig. 4.

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 1) along with examples of annotated reviews.[1] In the testing portion, annotators choose phrases that are relevant to a sustainability aspect from example reviews. Annotators have to 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.

After passing the test, annotators are presented with 15 reviews and are asked to complete the steps shown in Fig. 5.

Reviews are pulled from a server using weighted random sampling (see Sec. 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 6 shows a highlighting example for an environmental aspect.

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. 7).

If a phrase did not mention a specific product feature, annotators are asked to type "general." The emotional strengths are rated on a five-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. 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. Sixteen 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 versus the first). The overall emotional strength in a review phrase is then calculated as shown in Eq. (1)

$$\text{Emotional strength} = \text{Positive strength} - \text{Negative strength} \quad (1)$$

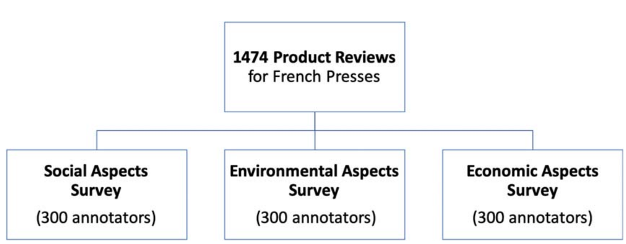

Fig. 4 Three survey versions

Table 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

Fig. 5 General annotation process

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

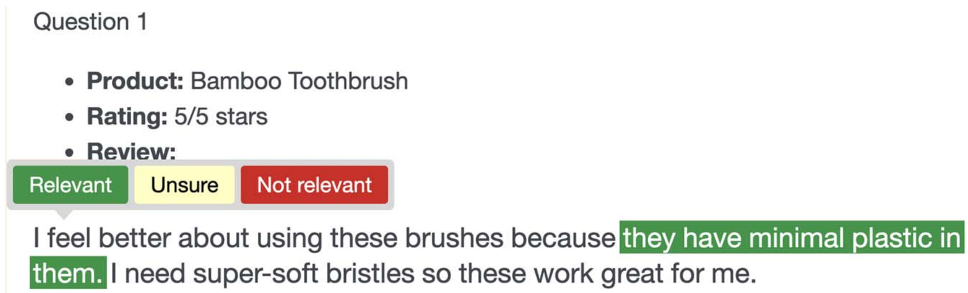

Fig. 6 Example of highlighting a phrase

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. 6). If either of these options is selected, annotators skip the questions in Fig. 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.

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 takes into account how many times a review has been previously selected and prioritizes reviews that have fewer annotations. Equation (2) provides a mathematical representation of this

$$S(r) = \left(1 - \frac{\text{counter}(r)}{3}\right) * \text{random}() \tag{2}$$

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.

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 a large number of 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 [27]. 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 of 97% approval rate. High approval rates are correlated most strongly with data quality [27]. Respondents based in the US also provide the highest response quality on average [28]. Moreover, after pilot testing, we found that the survey formatting on

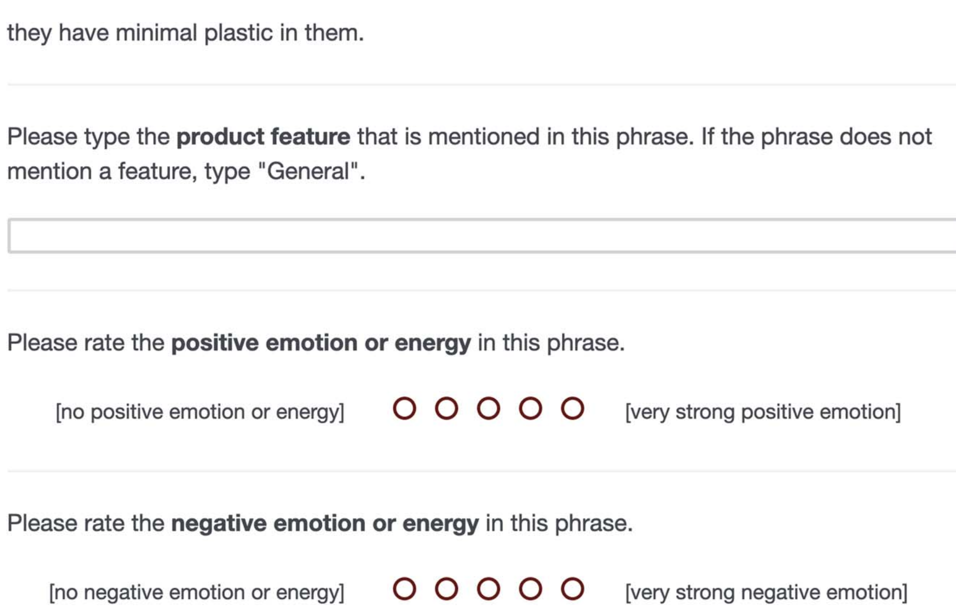

Fig. 7 Example of questions about a highlighted phrase

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 [28]. The surveys were launched using human intelligence tasks on the MTurk platform.

We approved 871 responses out of the 900 total annotators. We used two criteria to approve responses: (1) time to complete the survey (t) is within 1 standard deviation (σ) of mean completion time (μ) or longer (i.e., $t \geq \mu - \sigma$) and (2) passing a checkpoint question. For participants who 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.

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.

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 BOW model as well as bigrams and trigrams [29]. 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 a given phrase. Table 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.

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 predefined and tuned for optimal results. The LDA model is presented mathematically in Eq. (3)

$$P(t_i|d) = \sum_{j=1}^{|Z|} P(t_i|z_i = j) * P(z_i = j|d) \tag{3}$$

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.

Table 2 Simple BOW model example

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

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 has to 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 Sec. 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.

4.3.2 Build a Logistic Classifier. We implemented a 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. (4)

$$p(Y = 1|X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} \tag{4}$$

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) [30]. The βs are fitting parameters that are optimized using a maximum likelihood function shown in Eq. (5)

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} p(x_i)^{y_i}(1 - p(x_i))^{1-y_i} \tag{5}$$

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. (4) yields a number close to 1 for reviews that have a positive sentiment and a number close to 0 for reviews that have a 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 fivefold cross-validation to optimize penalty terms.

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 Sec. 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.

5 Preprocessing and Model Evaluation

Before featurizing the annotations, we first preprocessed the text data collected. This includes the phrases highlighted as relevant and the product features typed in by the annotators. Preprocessing text is done to minimize the amount of noise in the data by removing information that is unlikely to add value. The following preprocessing steps were taken: lowercasing, removing punctuation, removing stop-words (words like "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 fivefold 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. (6)–(8), respectively

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{6}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{7}$$

$$F1 = 2 * \frac{\text{Precision*Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

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 five reviews with positive sentiment and the model predicts that only two 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 3–5 for social, environmental, and economic aspects, respectively. These scores evaluate how well the model predicts positive and negative sentiments in phrases that contain sustainability aspects.

Table 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 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 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 and 0.90), while they are lower for predicting negative sentiment (between 0.53 and 0.69). This is likely because there were more annotated phrases related to sustainability that have positive sentiment compared with negative across the three sustainability aspects. 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. Section 6 presents the PerSFs extracted in this study.

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.

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 five-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 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 six relevant reviews per annotator followed by a spike at 15 reviews. The distributions are skewed toward 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.

Figures 9–11 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. 8–11 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.

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.

6.2.1 Topic Modeling Output. Table 6 shows the extracted product features using the topic modeling approach outlined in Sec. 4.3.1. The features are in order of highest occurrence in the

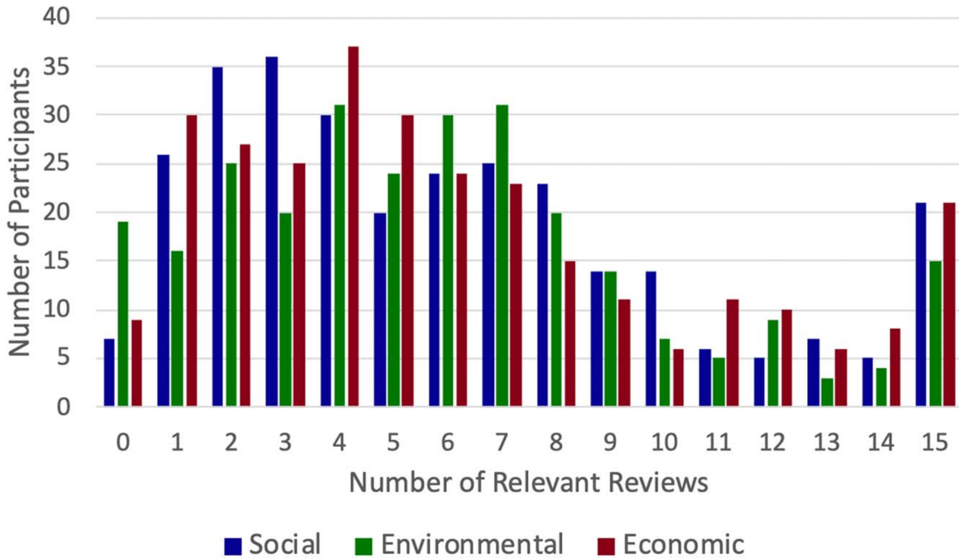

Fig. 8 Number of relevant reviews per annotator

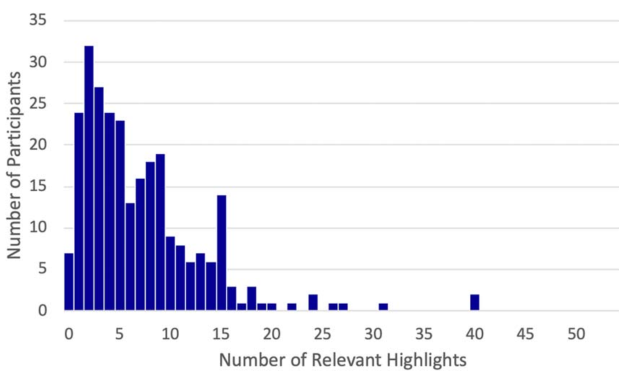

Fig. 9 Number of highlights per annotator of social aspects

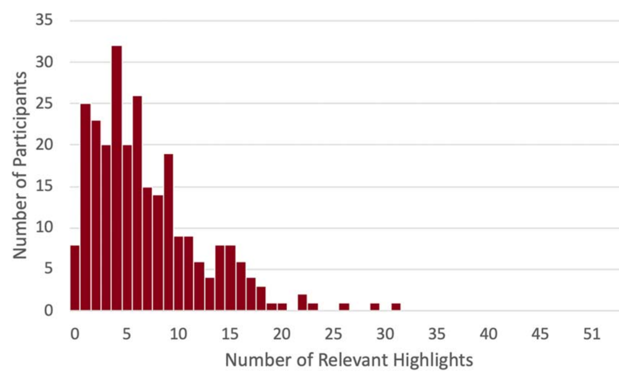

Fig. 11 Number of highlights per annotator for economic aspects

annotated phrases. Note that we manually categorized the product features shown in Table 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."

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 1) and specific product features generated by the annotators (such as "glass carafe"). Product

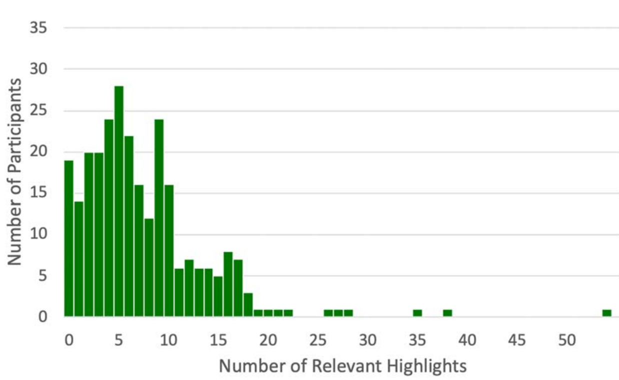

Fig. 10 Number of highlights per annotator for environmental aspects

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 6, we can see that features tend to become more product-

Table 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	–	–

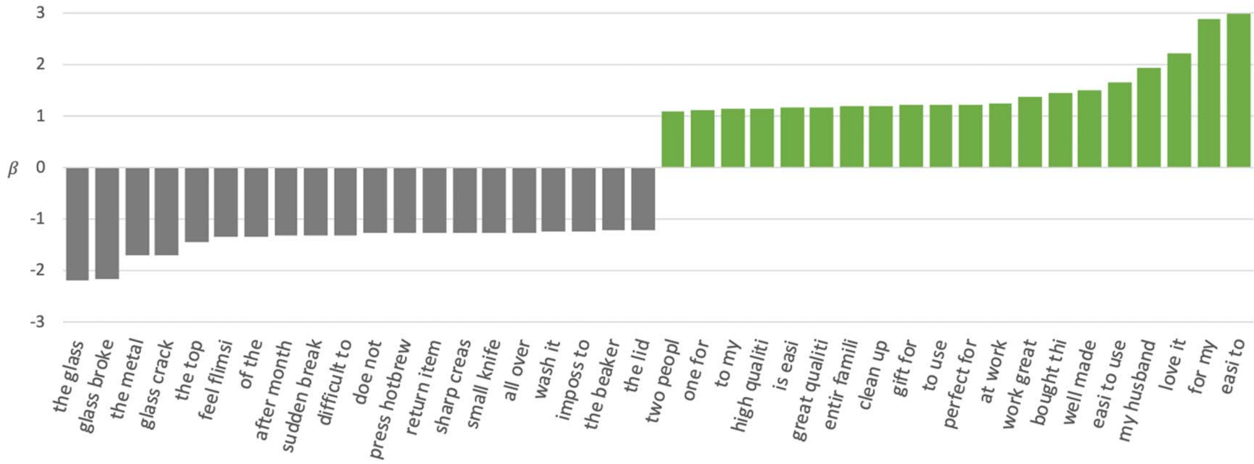

Fig. 12 Top 20 most positive (right side) and negative (left side) logistic classification parameters for social aspects

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.

6.2.2 Logistic Classification Output. The largest and smallest logistic classification parameters from each of the sustainability aspect models are shown in Figs. 12–14. The larger (positive) parameters correspond to features that the model predicts have a positive sentiment while the smaller (negative) parameters

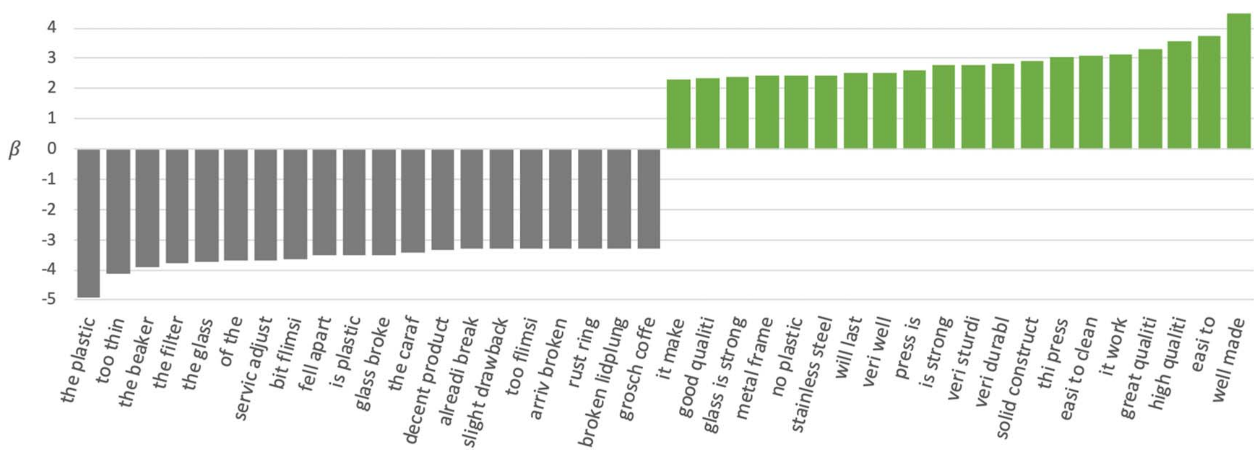

Fig. 13 Top 20 most positive (right side) and negative (left side) logistic classification parameters for environmental aspects

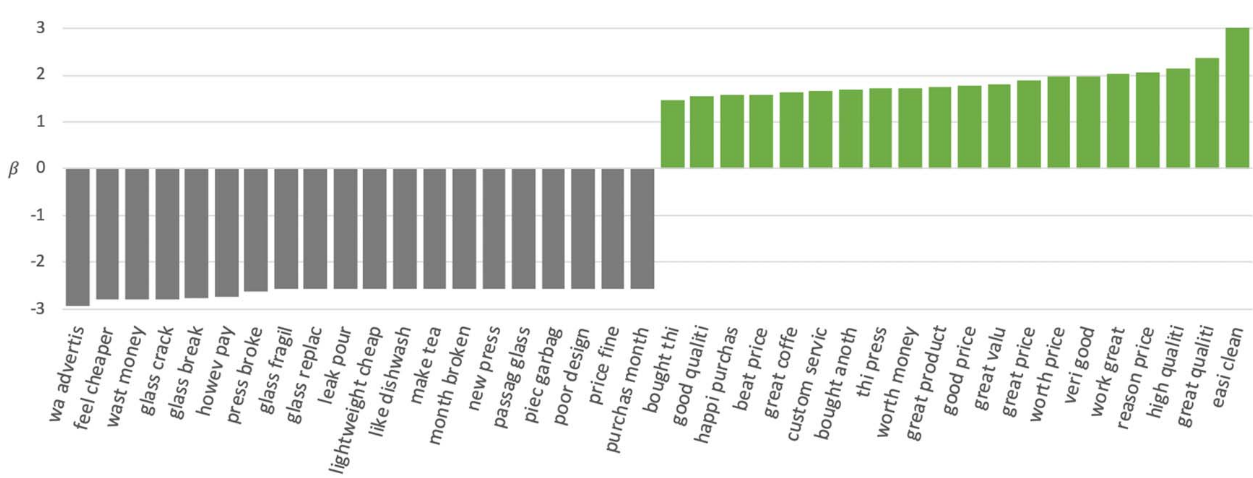

Fig. 14 Top 20 most positive (right side) and negative (left side) logistic classification parameters for economic aspects

Table 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	–

correspond to features that the model predicts have a negative sentiment. Note that the features displayed in the figures have been stemmed as part of preprocessing the highlighted phrases such as "bought thi" in Fig. 12, which may originally have been "bought this" or "bought these" (see Sec. 5 on stemming). Moreover, note that synonyms are present in the results (for example, "great valu" and "worth money" in Fig. 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).

Table 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. 12–14, 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.

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 [31]. 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 sentiments. For example, energy and water consumption or recycling did not have a significant effect on the environmental aspects model in Fig. 13. To investigate this further, we performed a life cycle analysis (LCA) using Sustainable Minds [32] 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. 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.

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 8 shows the carbon footprint for materials of two French Presses; the first is the original design from Fig. 15, the second replaces plastic parts with 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 [31,33]. 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.

Turning to the PerSFs that were found to have a significant effect, we will now offer some recommendations for designers. For social aspects, in Fig. 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 like "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. 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 like "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 a negative sentiment.

Table 8 CO_2 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

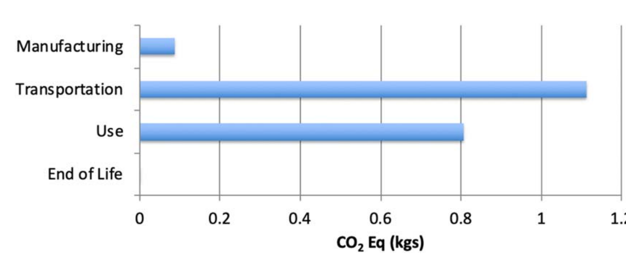

Fig. 15 Life cycle analysis of French Press

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. 14, the extracted PerSFs that have a 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 of using "reliable" materials like metal. Designers can perform further semantic testing to identify metals and finishes that read as "reliable." Glass can also be perceived as positive as long as it does not impair the durability of the product. For economic aspects, PerSFs revolve around how well the product works in general and if it is a good price, 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. The study does not investigate the generalizability of the method. 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 3–5 suggest that there is noise in the features associated with negative sentiment, which could explain why terms like "dishwasher" and "make tea" appear as negative features for economic aspects (Fig. 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. 8–11 by simplifying questions in the survey and emphasizing highlighting instructions.

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, like giving the product as a gift to a relative, while negative PerSFs include tangible features that could be unsafe to a user, like "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 underrepresented as compared with "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 five-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. 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 the demand for sustainable products.

Acknowledgment

We would like to thank Dr. W. Ross Morrow for writing the JavaScript code and implementing the server that provides weighted random sampling of product reviews in Qualtrics [34]. We would also like to thank Daniel Book for writing the JavaScript code that enables custom highlighting features in Qualtrics.

References

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

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

[3] Kim, E.-H., and Lyon, T. P., 2015, "Greenwash vs. Brownwash: Exaggeration and Undue Modesty in Corporate Sustainability Disclosure," Organ. Sci., **26**(3), pp. 705–723.

[4] 2018, "Quarterly Share of E-Commerce Sales of Total U.S. Retail Sales From 1st Quarter 2010 to 3rd Quarter 2018," Statista. https://www.statista.com/statistics/187439/share-of-e-commerce-sales-in-total-us-retail-sales-in-2010/.

[5] Roghanizad, M. M., and Neufeld, D. J., 2015, "Intuition, Risk, and the Formation of Online Trust," Comput. Hum. Behav., **50**(Sept.), pp. 489–498.

[6] MacDonald, E. F., Gonzalez, R., and Papalambros, P. Y., 2009, "Preference Inconsistency in Multidisciplinary Design Decision Making," ASME J. Mech. Des., **131**(3), p. 031009.

[7] Ren, Y., Burnap, A., and Papalambros, P., 2013, "Quantification of Perceptual Design Attributes Using a Crowd," International Conference on Engineering Design, Seoul, Korea, Aug. 19–22.

[8] Engström, P., and Forsell, E., 2018, "Demand Effects of Consumers' Stated and Revealed Preferences," J. Econ. Behav. Organ., **150**(June), pp. 43–61.

[9] 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," Mark. Lett., **19**(3/4), pp. 337–354.

[10] Decker, R., and Trusov, M., 2010, "Estimating Aggregate Consumer Preferences From Online Product Reviews," Int. J. Res. Mark., **27**(4), pp. 293–307.

[11] 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.

[12] Liu, Y., Jin, J., Ji, P., Harding, J. A., and Fung, R. Y. K., 2013, "Identifying Helpful Online Reviews: A Product Designer's Perspective," Comput. Aided Des., **45**(2), pp. 180–194.

[13] Rai, R., 2012, "Identifying Key Product Attributes and Their Importance Levels From Online Customer Reviews," ASME 2012 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference, Chicago, IL, Aug. 12–15.

[14] Stone, T., and Choi, S.-K., 2013, "Extracting Customer Preference From User-Generated Content Sources Using Classification," ASME 2013 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference, Portland, OR, Aug. 4–7.

[15] Singh, A. S., and Tucker, C. S., 2015, "Investigating the Heterogeneity of Product Feature Preferences Mined Using Online Product Data Streams," ASME 2015 International Design Engineering Technical Conferences & Computers and Information in Engineering Conferences, Boston, MA, Aug. 2–5.

[16] Singh, A., and Tucker, C. S., 2017, "A Machine Learning Approach to Product Review Disambiguation Based on Function, Form and Behavior Classification," Decis. Support Syst., **97**(2016), pp. 81–91.

[17] Kataria, S., Mitra, P., and Bhatia, S., 2010, "Utilizing Context in Generative Bayesian Models for Linked Corpus," 24th AAAI Conference on Artificial Intelligence, Atlanta, GA, July 11–15.
[18] Krestel, R., Fankhauser, P., and Nejdl, W., 2009, "Latent Dirichlet Allocation for Tag Recommendation," Third ACM Conference on Recommender Systems, New York, NY, Oct. 23–25.
[19] Tuarob, S., Pouchard, L. C., and Giles, C. L., 2013, "Automatic Tag Recommendation for Metadata Annotation Using Probabilistic Topic Modeling," 13th ACM/IEEE-CS Joint Conference on Digital Libraries, Indianapolis, IN, July 22–26.
[20] Tuarob, S., Pouchard, L. C., Noy, N., Horsburgh, J. S., and Palanisamy, G., 2012, "ONEMercury: Towards Automatic Annotation of Environmental Science Metadata," Second International Workshop on Linked Science, Boston, MA, Nov. 12.
[21] Zhang, X., and Mitra, P., 2010, "Learning Topical Transition Probabilities in Click Through Data With Regression Models," 13th International Workshop on the Web and Databases, Indianapolis, IN, June 6.
[22] Tuarob, S., and Tucker, C. S., 2015, "Automated Discovery of Lead Users and Latent Product Features by Mining Large Scale Social Media Networks," ASME J. Mech. Des., **137**(7), p. 071402.
[23] Tuarob, S., and Tucker, C. S., 2015, "A Product Feature Inference Model for Mining Implicit Customer Preferences Within Large Scale Social Media Networks," ASME 2015 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference, Boston, MA, Aug. 2–5.
[24] Thelwall, M., Buckley, K., Paltoglou, G., and Cai, D., 2010, "Sentiment Strength Detection in Short Informal Text," Am. Soc. Inf. Sci. Technol., **61**(12), pp. 2544–2558.
[25] 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," Eng. Appl. Artif. Intell., **73**(Aug.), pp. 149–162.
[26] Nagamachi, M., and Imada, A. S., 1995, "Kansei Engineering: An Ergonomic Technology for Product Development," Int. J. Ind. Ergon., **15**(1), pp. 3–11.
[27] Paolacci, G., and Chandler, J., 2014, "Inside the Turk: Understanding Mechanical Turk as a Participant Pool," Curr. Dir. Psychol. Sci., **23**(3), pp. 184–188.
[28] Goodman, J. K., and Paolacci, G., 2017, "Crowdsourcing Consumer Research," J. Consum. Res., **44**(1), pp. 196–210.
[29] Jurafsky, D., 2018, "N-Gram Language Models," *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*.
[30] James, G., Witten, D., Haste, T., and Tibshirani, R., 2006, An Introduction to Statistical Learning With Applications in R.
[31] MacDonald, E. F., and She, J., 2015, "Seven Cognitive Concepts for Successful Eco-Design," J. Cleaner Prod., **92**(Apr.), pp. 23–36.
[32] 2019, "Sustainable Minds," Sustainable Minds, LLC. https://www.sustainableminds.com.
[33] Slimak, M. W., and Dietz, T., 2006, "Personal Values, Beliefs, and Ecological Risk Perception," Risk Anal., **26**(6), p. 1689e1705.
[34] 2018, "reviewsampler," Github. https://github.com/wrossmorrow/reviewsampler.