Design Science

Differentiating online products using customer perceptions of sustainability

Nasreddine El Dehaibi and Erin F. MacDonald

Mechanical Engineering, Stanford University, Stanford, CA, USA

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 like a '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 enabling them to make informed purchase decisions. In this study, we validate how our previously developed method can be generalised by testing it with electric scooters and baby glass bottles. Features perceived as sustainable for both products are extracted and second, participants are tested on interpreting the features using a novel collage approach. Participants placed products on a set of two axes and selected features from a list. Our results confirm that the proposed method is effective for identifying features perceived as sustainable, and that it can generalise for different products with limitations. Positively biased Amazon reviews can limit the natural language processing performance. We recommend that designers use our method when designing products to capture feature perceptions and help inform customer-oriented design decisions.

Keywords: Customer perceptions, sustainable design, natural language processing, online reviews

Received 23 August 2021
Revised 31 May 2022
Accepted 06 June 2022

Corresponding author
N. El Dehaibi
ndehaibi@stanford.edu

Des. Sci., vol. 8, e19
journals.cambridge.org/dsj
DOI: 10.1017/dsj.2022.14

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 (Du & MacDonald 2016). These perceptions are derived from prior experiences and available information, acting as mental shortcuts for customers to simplify decision making (MacDonald & She 2015). For example, customers tend to judge how absorbent paper towels are based on the presence of quilted lines (MacDonald, Gonzalez & Papalambros 2009). Customers can simplify their decision making based on how product features align with their perceptions.

While relying on perceptions can help customers simplify decisions, it can also mislead customers to make uninformed decisions (MacDonald & She 2015). This

is often seen with sustainable products where features perceived as sustainable may not contribute to engineered sustainability. In this article, a feature is defined as either a visual aspect of a product such as material or as a descriptive aspect like 'a nice gift'. Moreover, engineered sustainability is defined as real sustainability according to well-studied methodologies such as a life cycle analysis (LCA). For example, customers may perceive a stainless-steel coffee maker as more sustainable than a plastic one, but according to an LCA, it is the energy efficiency that has the largest environmental impact. Despite this, designers tend to focus on engineered sustainability requirements while neglecting perceived sustainability (MacDonald & She 2015). This is validated by a lack of market success for sustainable products despite market research indicating customers are willing to pay more for them (The Sustainability Imperative: New Insights on Consumer Expectations 2015). Moreover, customers have grown sceptical of green marketing strategies like eco-labels (Kim & Lyon 2015).

A robust literature exists on customer perceptions of features when purchasing products (see Section 2 for an overview). MacDonald *et al.* (2009) identified a relationship between perceived and engineered requirements using a discrete choice analysis survey with paper towels. The results showed that customers constructed their perceptions on an as-needed basis and are not inherently found in people. For example, in a paper towel survey participants claimed they would not purchase nonrecycled paper towels for any price, but later reported purchasing from brands with 0% recycled paper towels the last time they went shopping. In a subsequent study, She & MacDonald (2017) demonstrated how perceived sustainable features led participants to prioritise engineered sustainability concerns in a decision scenario with toasters. For example, an embossed leaf pattern on a toaster led participants to prioritise 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. Simple design features can therefore help shape perceptions and influence decision making.

The previous literature highlights why it is important for designers to design-in perceptions, and for designers to meet customers where they are. Specifically for sustainability, it is important for designers to create a product that is both engineered to be sustainable and also perceived as sustainable by the customer. Previous literature supported designing-in features based on perceptions, but a method was lacking for identifying features as perceived by the customer. Literature has shown online reviews to be a treasure trove of customer perceptions of design features. For example, Ren, Burnap & Papalambros (2013) used Amazon Mechanical Turk (MTurk) respondents to assess the perceived safety of car designs using online reviews and machine learning. Building on this, we previously developed a method to identify features perceived as sustainable from online reviews using crowdsourced annotations and natural language processing (El Dehaibi, Goodman & MacDonald 2019) (refer to section 'Combining machine learning with collage approaches' for a deeper overview). We extracted features perceived as sustainable using French presses as a case study and demonstrated they are not fully aligned with engineered sustainability. In a subsequent study, we confirmed that participants identified the extracted French press features as sustainable using a novel collage activity (El Dehaibi, Liao & MacDonald 2021).

https://doi.org/10.1017/dsj.2022.14 Published online by Cambridge University Press

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 validated that participants identified the French press features as sustainable even if the features may not contribute directly to engineered sustainability.

In this study, the generalizability of our previous findings is tested by recreating the methods using different products and assessing the similarities in the results. The benefit of our approach is that we rely on customer reviews to determine customer perceptions. 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 article is organised as follows: Section 2 presents a background on the role of customer perceptions in decision making, the research propositions and hypotheses are in Section 3, Section 4 presents our method, the results and analysis are in Section 5, Section 6 presents our discussion, and the article is concluded in Section 7.

2. Related work

As more purchases occur online, several papers have explored the changing context in which customers form perceptions, using tools like machine learning and collage activities to extract perceptions from online reviews. Literature has revealed how product descriptions, online reviews, price, customer demographic, website features, and return policies can influence online decision making. Moreover, an active body of research has leveraged machine learning to uncover customer perceptions from online reviews. The gap in this literature lies in understanding how actual features, such as 'handle shape', would influence customer perceptions. This research aims to fill this gap by identifying generalizable methods that link specific visual or descriptive product features to customer perceptions. An overview of current literature and gaps is provided in this section. In addition, details are provided on our previous studies as we build on them in this study.

2.1. Customer perceptions in online decision making

In this section, a literature review is presented on how customer perceptions shape online decision making. As the following papers reveal, product descriptions, online reviews, price, customer demographic, website features, and return policies have all been shown to influence decision making. These factors serve as important considerations when designing a product for online sale.

Wang *et al.* (2016) investigated the impact of online reviews embedded in product descriptions on purchasing decisions. The scholars simulated a shopping experience based on Taobao, a Chinese e-commerce website that automatically bundles online review fragments into descriptions for certain products. The scholars recruited participants to explore the website while wearing an eye-tracking device that tracks eye movements in real time including when a person's eyes focus on certain objects (known as fixation time). The scholars 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

https://doi.org/10.1017/dsj.2022.14 Published online by Cambridge University Press

decisions, the scholars 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 scholars 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, Malthouse & Viswanathan (2017) studied the influence of product price and customer perceptions of reviews on online purchase decisions. The scholars 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 scholars 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 scholars used a logistic regression model with the purchase decision as a dependent variable and found that the product price plays an important role in how ratings and reviews influence the 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 the way customers perceive product reviews.

von Helversen *et al.* (2018) investigated the relationship between customer age and the influence of perceptions of product attributes and reviews on purchasing decisions. The scholars 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 scholars 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 & Pedersen (2004) explored how interactive features like content personalization and customer communities influence perceptions of customer experience on a shopping website. The scholars 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 emphasise 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.* (2019) investigated how return policies influence customer perceptions of products and decision making depending on the market stage of a business. The scholars propose a multistage 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 a varying influence on repurchase behaviour

https://doi.org/10.1017/dsj.2022.14 Published online by Cambridge University Press

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, the role of return policies on customer perceptions is crucial depending on the market stage of the seller.

2.2. Customer perceptions in online decision making

The literature discussed thus far looks at factors like age, reviews, price and website features to influence customer perceptions and decision making. A gap in previous literature is understanding how actual features, such as 'handle shape', would influence customer perceptions. This presents an opportunity for designers to determine how specific visual and descriptive product features align with customer perceptions to drive purchasing decisions. The development of e-commerce and social media provides a wealth of information that designers can tap into online. In this section, a literature review is presented on methods to extract customer perceptions from online content including machine learning and collage approaches.

Machine learning approaches

Zhang, Yan & Zhang (2020) use Airbnb, an online marketplace for accommodations, to study the influence of accommodation host self-descriptions on customer trust and booking behaviours. The scholars annotated 4179 host descriptions from Airbnb listings based on the perceived trustworthiness of the hosts. The scholars 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 scholars 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 hosts led to more booking decisions. These results point to the importance of language when describing products to drive purchasing decisions.

Liu, Jiang & Zhao (2019) use natural language processing to identify product competitive advantages from social media content. The scholars collected reviews of a Volkswagen Passat, a German car model, 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 scholars 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 scholars demonstrated how customer perceptions can drive competitor analyses to inform design decisions for next iteration products.

Zhou *et al.* (2020) developed a method to extract relevant product features from online reviews. The scholars collected 91,738 review sentences across several products created by Amazon (Fire tablet, Echo, etc.). The scholars manually

https://doi.org/10.1017/dsj.2022.14 Published online by Cambridge University Press

labelled a sample of 10,000 reviews as either relevant or not to the product and trained a fastText algorithm to filter out irrelevant reviews; about 20% of the reviews were filtered out to remove noise. The scholars then used a topic modelling approach called Latent Dirichlet Allocation (LDA) to identify 'topics' that in this case are product features from the remaining review sentences. Finally, the scholars used a sentiment analysis library called Vader to measure the sentiment of review sentences across the identified features.

Park & Kim (2020) proposed a method to improve the accuracy and diversity of extracted features from online reviews using keyword embedding and two clustering methods. The scholars tested their method with 27,201 laptop reviews, 20,823 wearable device reviews and 19,159 smartphone reviews, and also collected 27,201 reviews for 61 laptops. The reviews were preprocessed and embedded into vectors, and then clustered using X-clustering on noun phrases for noise reduction. Special clustering was then applied to extract features. The scholars extracted 10 features for laptops, nine for wearable devices and eight for smartphones.

Kim, Park & Kim (2021) build on the above method by applying it to investigate the role of COVID-19 pandemic in changing customer preferences. The scholars collected 8548 reviews for smartphones dating before the pandemic and 7263 reviews for smartphones dating after the pandemic began. Smartphones included both new and refurbished products. The reviews were preprocessed, embedded into vectors, and clustered to identify key product features: screen, memory, camera, battery, security and price. The scholars then used a sentiment library to calculate the sentiment for each review sentence according to the clustered product features. The results showed that for new phones, sentiments decreased across many of the product feature clusters, while for refurbished phones that difference was less significant. While the scholars identified sentiment changes of product features, the method does not capture how customer perception influences this sentiment.

Similarly, Bag, Tiwari & Chan (2019) develop a method to predict a customer's purchase intention based on review polarity and sentiment scores. The scholars used natural language processing to extract features from 29,069 online reviews. Then linear and nonlinear regression analysis was performed including both review polarity and brand social perception scores to extract salient product features. The method identifies salient product features but is not able to identify how customer perceptions play a role in the identified features.

Combining machine learning with collage approaches

Previous literature identified methods to extract customer perceptions from online content but has not yet identified methods that can link specific visual or descriptive product features to customer perceptions. Moreover, previous literature has not developed a method to validate perceptions of product features by testing those perceptions on users 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 (El Dehaibi *et al.* 2019), and second developed a novel collage approach to test those features in terms of how users like and evaluate products

https://doi.org/10.1017/dsj.2022.14 Published online by Cambridge University Press

(El Dehaibi *et al.* 2021). 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. Details are provided in our previous work below since we build heavily on them for this study.

In the first study, features perceived as sustainable are extracted using crowdsourced annotations of online reviews and a natural language processing algorithm (El Dehaibi *et al.* 2019). The approach combined research from identifying sustainability perceptions, rating design ideas, and natural language processing (Figure 1) and is outlined in four steps (Figure 2). Product reviews are collected for a target product type from the online marketplace Amazon, annotated the reviews using a crowdsourcing platform based on criteria related to the perceptions, modelled the reviews and annotations using natural language processing, and extracted features perceived as sustainable from the model.

The method was tested 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. A pilot study previously concluded that participants had more clarity when focusing on the pillar, so we studied each one individually. For this study, respondents are assigned

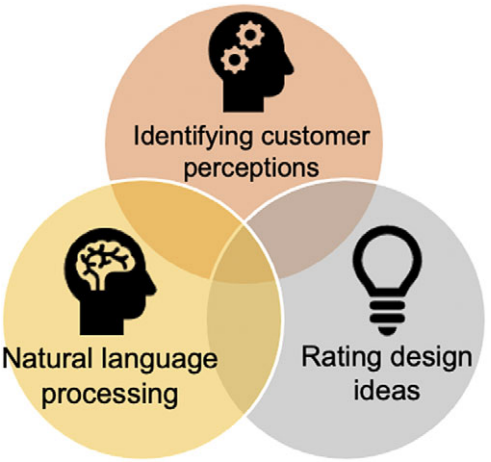

Figure 1. Interdisciplinary approach.

Collect	Collect product reviews from Amazon
Annotate	Annotate reviews via crowdsourcing
Model	Model reviews and annotations using NLP
Identify	Identify perceived sustainable product features

Figure 2. Extracting customer perceptions method flow.

https://doi.org/10.1017/dsj.2022.14 Published online by Cambridge University Press

Table 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

to one of three versions of the survey to focus on one sustainability pillar and trained on their assigned pillar using basic guidelines. Respondents then highlighted parts of reviews relevant to their pillar and rated the emotions in their highlights.

The annotations are modelled 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 1 (see section 'Machine learning model' 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.

Salient positive features are identified 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. Engineered sustainability requirements of a French press are then identified using an LCA and found that crucial engineered sustainability requirements like 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. It is crucial for designers to account for both engineered and perceived sustainability to differentiate their products and stand out to customers. For example, a customer may pass on a real sustainable product because they do not perceive it as sustainable. A product should be both sustainable but also align with customers on what they perceive as sustainable.

In the second study, a novel collage approach is developed to test the extracted features perceived as sustainable by users in terms of how they like products and evaluate sustainability (El Dehaibi *et al.* 2021). Using the collage, a relationship is identified between features perceived as sustainable and user emotions in an engaging way without drawing attention to the features. A webapp collage activity was created with two axes: sustainability on the vertical axis (customised to one of the three pillars depending on the version of the collage) and likeability on the horizontal axis. An example of an environmental sustainability collage activity is shown in Figure 3. 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 were extracted previously as well as features 'not perceived as sustainable' that were identified for the collage study.

https://doi.org/10.1017/dsj.2022.14 Published online by Cambridge University Press

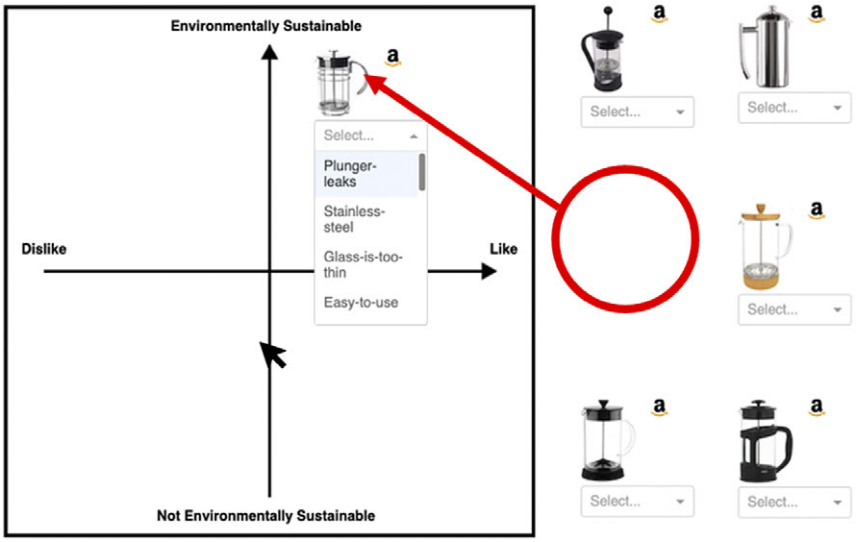

Figure 3. Dragging and dropping products on collage and selecting at least one phrase to describe each product.

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 'Collage activity') 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. A low correlation was found 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.

While our previous work validates a method to extract perceived sustainability features from online reviews, 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.

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, features perceived as sustainable are extracted 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. Participants are asked to place products along the two axes of the collage,

https://doi.org/10.1017/dsj.2022.14 Published online by Cambridge University Press

Table 2. Propositions and hypotheses from our previous studies

Propositions and hypotheses
P1: Phrases in product reviews perceived as sustainable contain semantic and syntactic characteristics that can be modelled (El Dehaibi *et al.* 2019)
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 extracted from a natural language processing algorithm as being sustainable (El Dehaibi *et al.* 2021). 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 (El Dehaibi *et al.* 2021)
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. (El Dehaibi *et al.* 2021). 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' (El Dehaibi *et al.* 2021)

sustainability and likeability, and to label the products using a list of the extracted features from the logistic classifier. In our previous studies, the following propositions and hypotheses are tested using French press products as a case study (Table 2).

Our goal for this study is to test if the same propositions and hypotheses hold when tested with multiple product types.

4. Methods

The method in this article is based on our work from two previous papers where a French press is used as the focal product (El Dehaibi *et al.* 2019, 2021). In this study, we validate how the method generalises when applied to different product types. Figure 4 provides an overview of the method. In the first part of the method, features are extracted from online reviews in three versions to account for each sustainability pillar: social, environmental and economic. In the second part of the method, each of the three sets of extracted features is tested with users on a collage. Specific steps are explained below.

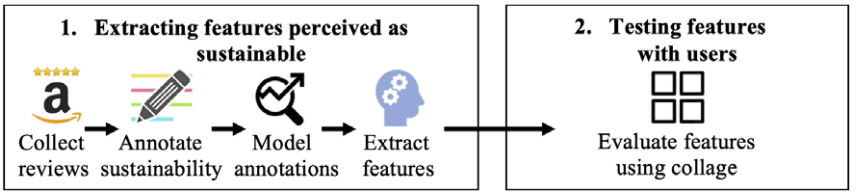

Figure 4. Method overview.

https://doi.org/10.1017/dsj.2022.14 Published online by Cambridge University Press

4.1. Extracting features perceived as sustainable from online reviews

The methods outlined in this section aim to test proposition 1. Features perceived as sustainable are extracted for electric scooters and baby glass bottles using steps in Figure 2.

Collecting reviews

Electric scooters and baby glass bottles are selected as the focal products for this study because they (a) are different in design and function from the original French press product, (b) have varying aesthetic design features available, (c) regularly receive several hundred reviews on Amazon, and (d) likely have sustainability-related concerns for customers. Products different from a French press are selected to effectively evaluate how the method can be generalised. Products like 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, e.g., 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.

A total of 1500 Amazon reviews are scraped from four electric scooters and 1444 Amazon reviews from eight baby glass bottles. Four products and eight products are selected, respectively, from Amazon so that they (a) have varying aesthetic features from one another, (b) are in a similar price range, (c) have less than 500 reviews per product, and (d) have at least 80% estimated authentic reviews according to a data analytics tool (fakespot.com) for each product. Fakespot uses a natural language processing algorithm to analyse spelling, grammar, author history and mismatch of dates between reviews and purchase dates. The algorithm then outputs an estimated review authenticity grade for the product. Estimates can include inaccuracies and products are rarely rated as having over 90% authentic reviews. While some of the collected reviews may have been fake, the amount is likely insignificant with little to no relevant information for annotators to extract from.

Products that have varying features are selected to better test different features with users. Moreover, products in a similar price range are selected so that their quality and capabilities are like each other. Each product had less than 500 reviews to have a variety of products instead of one product dominating the reviews. The motivation was to have a variety of features to test. Finally, products that are estimated to have a high number of authentic reviews are selected so that real customer opinions and perceptions are collected. All the reviews scraped came from the United States to limit the number of reviews written in a foreign language. Moreover, reviewers that were less than 10 words were filtered as they tended to be generic, for example, 'this is a great product, I highly recommend it'.

Annotating reviews

Respondents from MTurk were recruited (referred to as 'annotators', see section 'Annotators') to annotate the scraped reviews based on sustainability

https://doi.org/10.1017/dsj.2022.14 Published online by Cambridge University Press

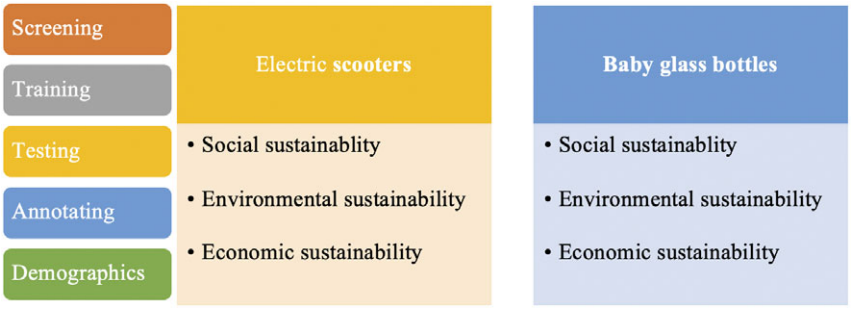

Figure 5. Three annotation survey versions per product.

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

Survey design. To guide the annotators, a survey was created 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. In total, there were 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), as shown in Figure 5.

In the training portion of the survey, sustainability criteria were displayed to the annotators 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 'Data collection') according to one of the sustainability aspects criteria.

Data collection. To annotate the reviews scraped in section 'Collecting reviews', the reviews are stored on a server so that they can be pulled live during the survey. A biased-random algorithm was used for selecting the reviews from the server to ensure that each review was presented to three different annotators. Each participant saw 15 reviews in total, one at a time. For each review, annotators were asked to highlight up to five parts of the review that they found relevant to their assigned sustainability criteria. If they highlighted parts of the review as relevant, annotators were asked 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.

Annotators. A total of 1800 annotators were recruited from Amazon Mechanical Turk to complete one of the six surveys. A total of 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 were compensated $4 each. Amazon Mechanical Turk respondents were recruited 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. While the demographics of MTurk participants are not representative of the USA, they align closely with the online population (Goodman &

https://doi.org/10.1017/dsj.2022.14 Published online by Cambridge University Press

Paolacci 2017), and therefore better fit a target Amazon customer. This is desirable for the study since we focus on Amazon reviews to extract feature perceptions.

To ensure high-quality responses, respondents were required to have at least a 97% approval rating and to be based in the USA. These requirements were set in the MTurk platform and confirmed them using screening questions in the survey. Moreover, a simple checkpoint question was included 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 like what was used in our previous work (El Dehaibi *et al.* 2019). Based on these criteria 1702 out of the 1800 responses were approved.

An average task time of 20–30 minutes was targeted to avoid response quality drops per participant with longer tasks. The number of annotators and reviews specifically determines the length of the task. Knowing that sample size is considered in statistics, we determined our sample size to present a strong proof of concept and recommend working with larger sample sizes in practice.

Machine learning model

A binary logistic classification model was used to extract features perceived as sustainable from the annotated reviews. 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 (James *et al.* 2013). This enabled us to extract salient product features directly from the classifier, unlike deep learning approaches.

The logistic classification model is represented in Eq. (1). Our input 'X' consists of the (a) phrases highlighted as 'relevant' to sustainability and (b) the product features typed in by the annotators, while the output 'Y' is binary representing the emotion in each phrase. The output Y was binarized such that 0 represented negative or neutral emotion while 1 represented positive energy. We opted for a binary output instead of a multiclass model due to the limited explanatory power from our dataset for a multiclass model

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

The beta fitting parameters are optimised with a maximum likelihood shown in Eq. (2):

$$L(\beta_0,\beta_1)=\prod_{i=1}^{n} p(x_i)^{y_i}(1-p(x_i))^{1-y_i}. \tag{2}$$

We used the Scikit-Learn and Natural Language Tool Kit (NLTK) libraries in Python to build a natural language processing machine learning model. The inputs were preprocessed 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. Then the inputs were quantified in a matrix and fed into a classifier. For the highlighted phrases a bag of words, bigrams, and trigrams were used. For the typed-in features, we summarised them into a set number of 'topics' using LDA and then hot-encoded them for each highlighted phrase. LDA is a topic modelling approach that identifies a set number of 'topics' from text based on the model shown in Eq. (3):

https://doi.org/10.1017/dsj.2022.14 Published online by Cambridge University Press

$$P(t_i|d) = \sum_{i=1}^{|Z|} P(t_i|z_i = j)P(z_i = j|d), \quad (3)$$

where z_i represents a product feature, d represents a review from a collection reviews D, and $|Z|$ is a preset total number of product features.

The data was split into a 70% training and 30% test set and implemented the logistic classifier model in Python using the Scikit package. Five-fold cross validation was used on the training set and penalty terms to shrink parameters based on Ridge regularisation to address potential overfitting from high dimensionality. As an external validity check on the models, precision, recall, and F1 were used as these are often more robust measures than accuracy (James *et al.* 2013).

Extracting perceptions

Salient product features perceived as sustainable were extracted from the machine learning model that drove positive and negative sentiments. The magnitude of the beta parameters attached to a given feature indicates the influence of that feature on the model. For example, if 'stainless steel handle' was associated with a relatively large positive beta parameter, this indicates that this feature is a salient customer perception of sustainability. Alternatively, if a feature had a negative beta parameter, this would indicate a salient feature not perceived as sustainable. These features come from reviews of multiple products to capture a variety of different features. For example, multiple electric scooters were selected to capture perceptions of varying product shapes and designs. After extracting the product features perceived as sustainable, a collage experiment was conducted to validate that participants identified these features as sustainable.

4.2. Testing perceived features extracted from online reviews with participants

The method outlined in this section aims to test hypotheses 1 and 2. Totally, 300 additional respondents (referred to as 'participants' for this portion of the method) were recruited from MTurk to evaluate the products and features using the collage activity explained in section 'Combining machine learning with collage approaches'. 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, three versions of a survey were designed (accounting for each of the sustainability pillars) and assigned participants to one of the versions (see Figure 6). Like section 'Survey design', participants were asked to evaluate products for only one of the sustainability pillars based on pilot studies that demonstrated this led to more usable responses.

Presurvey

In the presurvey, participants were familiarised with their assigned sustainability criteria as well as the products that they will evaluate (Table 3). We selected the products according to the criteria explained in section 'Collecting reviews'.

During the presurvey participants were trained on their assigned sustainability pillar and led them to Amazon pages of the products in Table 3 to familiarise themselves before evaluating. They had to open each of the Amazon pages and

https://doi.org/10.1017/dsj.2022.14 Published online by Cambridge University Press

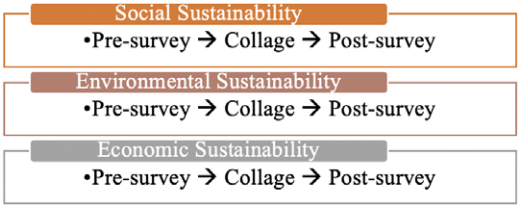

Figure 6. Three collage activity versions.

Table 3. Products in collage activity

Product name	Gotrax	Razor E300S	Mongoose	Razor EcoSmart	Segway	SKRT

spend a certain amount of time on them to proceed with the activity. This was required to ensure 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.

Collage activity

After completing the presurvey participants accessed a link to a collage webapp using the same interface shown in Figure 3. 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 presurvey. 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 Figure 3. We recorded the location of the center of the product image as a float. While this adds uncertainty, the impact is negligible since our focus is on the relative placement of products and features. Hypothesis 1 is tested based on the placement of the features on the collage, and hypothesis 2 is tested 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 'Extracting perceptions'. 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 randomised between participants. These features are presented in Section 5 as part of the results.

To further test hypothesis 1, a fourth collage activity was conducted 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, phrases were derived from the

https://doi.org/10.1017/dsj.2022.14 Published online by Cambridge University Press

unhighlighted parts of the annotated reviews collected using the method in section'Data collection'. 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. For this collage activity, an additional 100 participants were recruited using the procedures outlined in section 'Participants'.

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. This was included 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 postsurvey.

Postsurvey

In the postsurvey, participants were asked 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, participants were asked basic demographic questions.

Participants

A total of 300 participants were recruited from MTurk to complete the collage activity using the same recruiting criteria as in section 'Annotators', 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 were compensated $5 each. We did not analyse responses if they fell under one of the following: (a) participants completed the survey faster than the average time by at least one standard deviation or (b) they incorrectly answered a simple checkpoint question designed to gauge attention. Based on these criteria, we analysed 224 responses out of 300.

5. Results

First, the results that test the generalizability of proposition 1 related to the features extracted from online reviews are presented. Second, the results that test the generalizability of hypotheses 1 and 2 based on the placement of products and extracted features on the collage are presented.

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 modelled.

https://doi.org/10.1017/dsj.2022.14 Published online by Cambridge University Press

Table 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

Electric scooters model evaluation

The precision, recall and F1 scores for each of the sustainability pillars for electric scooters are shown in Table 4.

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 analysing the most salient negative features. These metrics are like findings from our previous study with French presses (see Table 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.

Electric scooters model output

Figures 7–9 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 like '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.

Baby glass bottles model evaluation

The precision, recall and F1 scores for each of the sustainability pillars for baby glass bottles are shown in Table 5.

Like our previous findings (Table 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 emphasises the importance of data balance for using this approach to extract features. Therefore, while proposition 1 may generalise for different products there are limitations in terms of selecting products with balanced reviews.

https://doi.org/10.1017/dsj.2022.14 Published online by Cambridge University Press

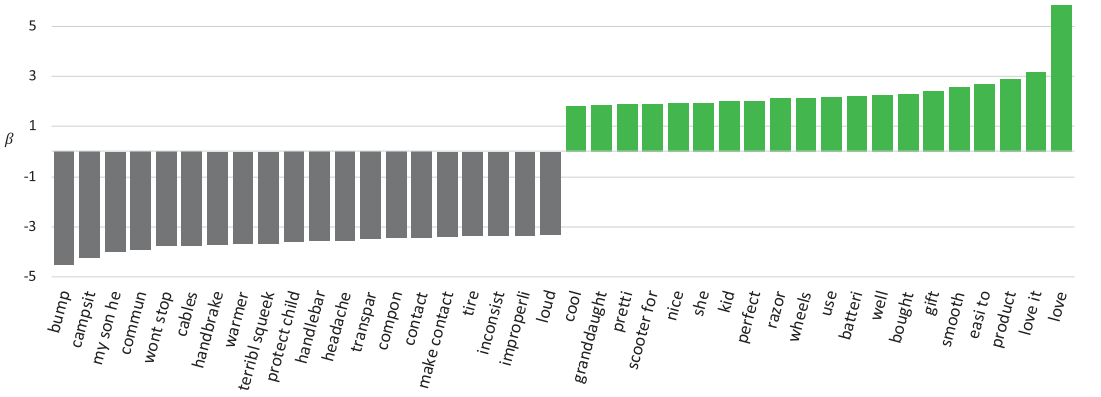

Figure 7. Most salient 20 positive and negative features of electric scooters perceived as socially sustainable.

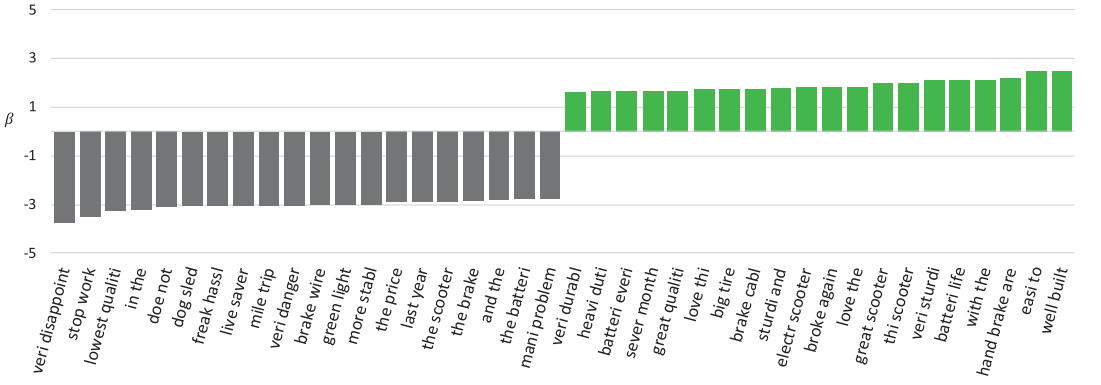

Figure 8. Most salient 20 positive and negative features of electric scooters perceived as environmentally sustainable.

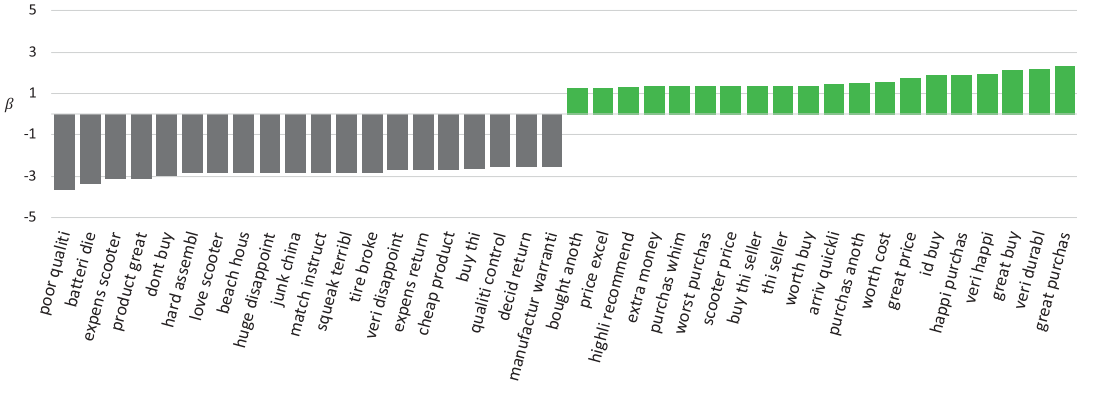

Figure 9. Most salient 20 positive and negative features of electric scooters perceived as sustainable for economic sustainability.

https://doi.org/10.1017/dsj.2022.14 Published online by Cambridge University Press

Table 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

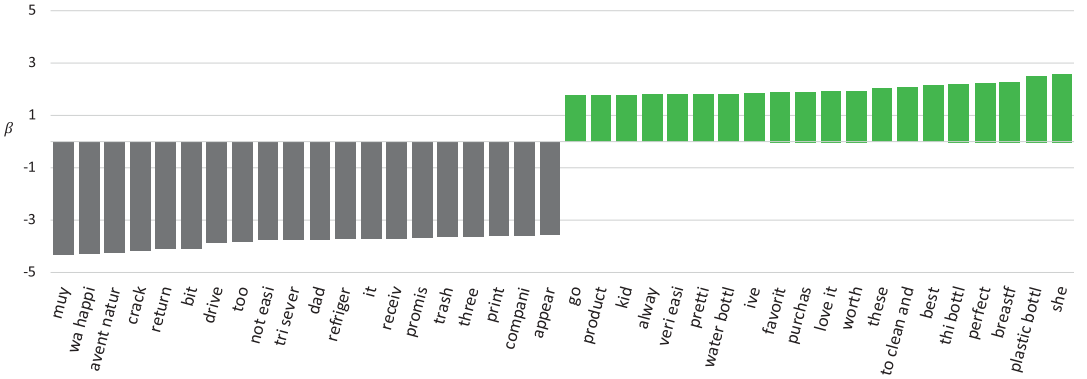

Figure 10. Most salient 20 positive and negative features of baby glass bottles perceived as socially sustainable.

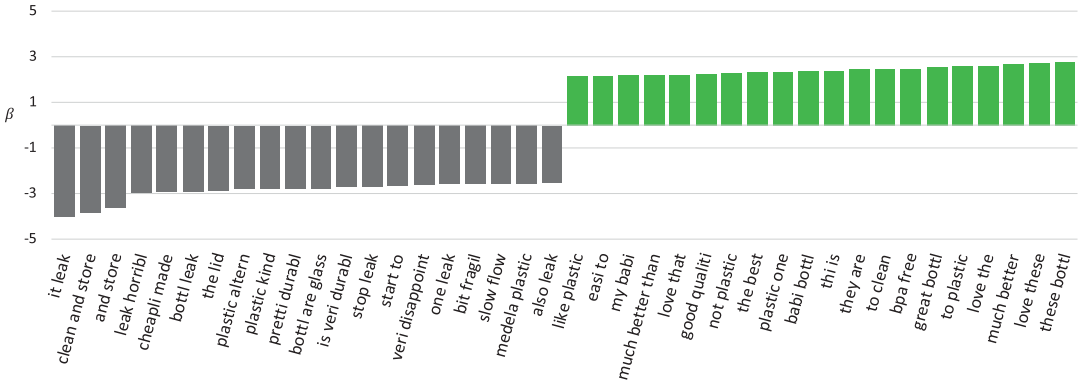

Figure 11. Most salient 20 positive and negative features of baby glass bottles perceived as environmentally sustainable.

Baby glass bottles model output

Figures 10–12 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

https://doi.org/10.1017/dsj.2022.14 Published online by Cambridge University Press

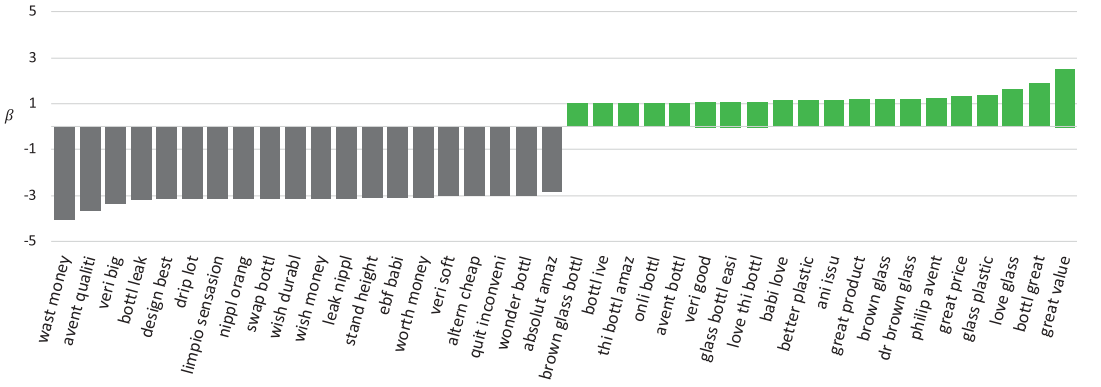

Figure 12. Most salient 20 positive and negative features of baby glass bottles perceived as sustainable for economic sustainability.

contain little meaning or are unintuitive, such as 'bit' for social sustainability or 'pretty durable' for environmental sustainability. These are likely due to the low metrics identified in Table 5. Therefore, while proposition 1 generalised with electric scooters, it could not generalise with baby glass bottles due to the severe imbalance in product review sentiments.

Based on the findings from section 'Baby glass bottles model evaluation' 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.

5.2. Collage results

This section is split into two parts: first, the location of electric scooter features on the collage is analysed, which tests hypothesis 1 and second, the placement of the products is analysed, which tests hypothesis 2. About 294 data points were excluded for products that were not moved from their starting location (starting locations are outside the collage boundaries, see Figure 3) from a total of 1834 recorded data points.

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.

Positive and negative features perceived as sustainable

Based on Figures 7–9, 10 positive and 10 negative features were identified to provide to participants during the collage activity for each of the sustainability pillars. These features are shown in Tables 6 and 7.

Table 8 shows a summary of the features selected during the collage activity.

The most selected positive feature across the three sustainability criteria was 'electric powered' while the most common negative feature was 'poor ride quality'. Figures 13–16 show the average placement of features on the collage.

https://doi.org/10.1017/dsj.2022.14 Published online by Cambridge University Press

Table 6. Positive perceptions of electric scooter sustainability

Social sustainability	Environmental sustainability	Economic sustainability
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 tyres	Buy from this seller
Stable ride	Well built	Great purchase

Table 7. Negative perceptions of electric scooter sustainability

Social sustainability	Environmental sustainability	Economic sustainability
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	Will not 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
Tyre broke	Low quality	Long wait time
Pure headache	Battery disposal	Piece of junk

The figures show distinct clusters between the positive and negative features, which supports the generalizability of hypothesis 1. A *t*-test was performed 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 9).

Like 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, a multivariate analysis (MANOVA) was performed using the '*x*' and '*y*' coordinates from the collage as dependent variables, and the rest of available information as independent

https://doi.org/10.1017/dsj.2022.14 Published online by Cambridge University Press

https://doi.org/10.1017/dsj.2022.14 Published online by Cambridge University Press

Table 8. Summary of features selected in collage

	Social		Environmental		Economic		Combined	
	Positive features	Negative features	Positive features	Negative features	Positive features	Negative features	Positive features	Negative features
# Participants	96		93		97		286	
Observations	355	222	384	154	371	242	1110	618
Average features per participant	3.7	2.31	4.13	1.66	3.82	2.49	3.88	2.16
Average features per product	59.17	37	64	25.67	61.83	40.33	185	103
Most common feature selected	Great gift	Poor ride quality	Electric powered	Low quality	Excellent price	Too expensive	Electric powered	Poor ride quality

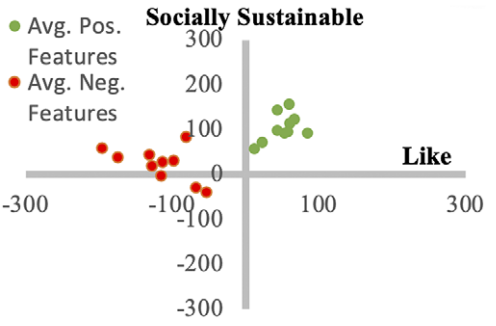

Figure 13. Average placement of positive and negative electric scooter features perceived as socially sustainable.

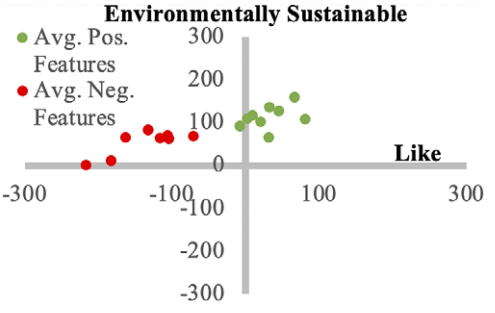

Figure 14. Average placement of positive and negative electric scooter features perceived as environmentally sustainable.

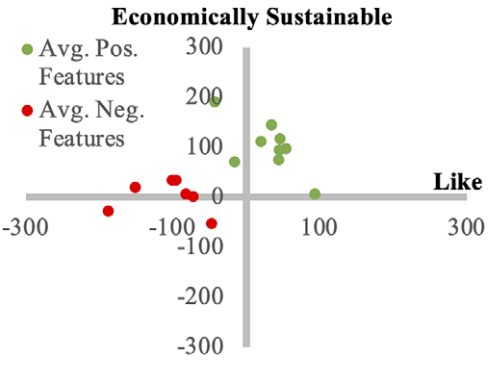

Figure 15. Average placement of positive and negative electric scooter features perceived as economically sustainable.

variables (Table 10). We chose the Pillai criterion for its robustness when linearity assumptions are not met (Delaney 2003). Across all sustainability criteria, the features were statistically significant, like our findings with features extracted for

https://doi.org/10.1017/dsj.2022.14 Published online by Cambridge University Press

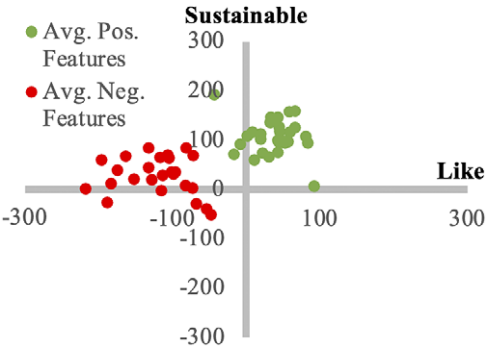

Figure 16. Average placement of positive and negative electric scooter features perceived as sustainable for all criteria.

Table 9. Two-sample *t*-test between positive and negative features perceived as sustainable

	Social		Environmental		Economic		Combined	
	Positive features	Negative features	Positive features	Negative features	Positive features	Negative features	Positive features	Negative features
Mean Y	103	21	118	56	106	7	110	23.9
Observation	246	137	278	88	209	140	733	365
$p(T \leq t)$	<0.001**		0.004*		<0.001**		<0.001**	
t critical	1.97		1.98		1.97		1.96	

*Significant at $p = 0.01$;
**Significant at $p = 0.001$.

Table 10. MANOVA output with positive and negative features perceived as sustainable

	Social			Environmental			Economic			Combined		
	Pillai	F	Pr(>F)	Pillai	F	Pr(>F)	Pillai	F	Pr(>F)	Pillai	F	Pr(>F)
Product	0.14	5.35	<0.001*	0.21	8.51	<0.001*	0.23	8.36	<0.001*	0.12	14	0.001*
Criteria	—	—	—	—	—	—	—	—	—	0.02	4.79	<0.001*
Feature type	0.29	73.2	<0.001*	0.21	48	<0.001*	0.19	37.5	<0.001*	0.24	171	0.001*

*Significant at $p = 0.001$.

French presses. Thus, our findings fail to reject hypothesis 1 for electric scooters and validate the generalizability of the hypothesis.

https://doi.org/10.1017/dsj.2022.14 Published online by Cambridge University Press

Table 11. Phrases not containing perceptions of electric scooter sustainability

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

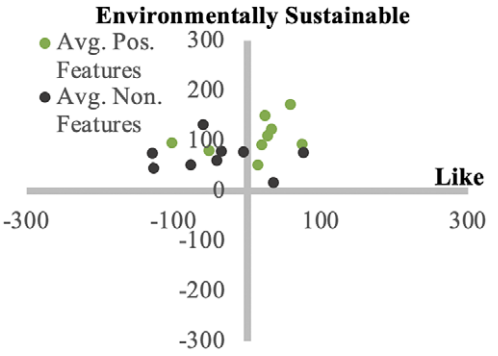

Figure 17. Average placement of positive features perceived as sustainable and features not related to sustainability.

Table 12. Two-sample *t*-test between positive features perceived as environmentally sustainable and features not related to sustainability

	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
$p(T \leq t)$ two-tail	0.026*	
t critical two-tail	1.96	

*Significant at $p = 0.05$

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 11. Figure 17 shows the placement of the new set of features.

A *t*-test using the y-coordinates of the two sets of features is shown in Table 12.

The *t*-test shows a statistically significant difference, supporting the generalizability of hypothesis 1. A MANOVA analysis with repeated measures is shown in Table 13.

https://doi.org/10.1017/dsj.2022.14 Published online by Cambridge University Press

Table 13. MANOVA output with positive features perceived as sustainable and features not related to sustainability

	Pillai	F	Num *df*	Den *df*	Pr(>F)
Product	0.192	8.13	10	768	<0.001*
Feature type	0.056	11.3	2	383	<0.001*

*Significant at $p = 0.001$.

Table 14. Repeated measures correlation between perceived sustainability of a product and liking the product

	Social	Environmental	Economic	Combined
Repeated measure correlation	0.18	0.09	0.08	0.11
p-value	0.006	0.042	0.034	0.001

The electric scooter features are statistically significant even with the more challenging list, like 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.

Product analysis

In this section, we present analyses for testing the generalizability of hypothesis 2. A repeated measured correlation was used 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 (Bakdash & Marusich 2017). The results are shown in Table 14.

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 demonstrating the usefulness of the collage tool for assessing sustainability perceptions.

6. Discussion

Our findings support the generalizability of proposition 1 that phrases perceived as sustainable in reviews contain semantic and syntactic characteristics that can be modelled. Looking at the machine learning model metrics in Tables 4 and 5 for electric scooters and baby glass bottles, respectively, we see that they are like the ones in our previous study with French presses (Table 1). The metrics for negative

https://doi.org/10.1017/dsj.2022.14 Published online by Cambridge University Press

sentiment faired poorer in this study, suggesting potential limitations on the generalizability (see below for more on limitations).

Similarities and differences were found between features perceived as sustainable for electric scooters and baby glass bottles. For social sustainability in Figure 7, many of the positive features of electric scooters were intangible, such as relating to family or gift-giving. This is like what was found in our previous test using French presses. For baby glass bottles in Figure 10, 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 of both electric scooters and baby glass bottles are mainly tangible. In the case of baby glass bottles in Figure 11, 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 the material. For electric scooters in Figure 8, the positive features included many components such as the battery life, brakes, and tyres. 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 scooters in Figure 9 related to how great of a value the product is, such as 'great purchase' or 'poor quality'. This is like what we found in our previous study with French presses. For baby glass bottles in Figure 12, positive features included different brands, as well as tangible features like 'plastic' while for negative features they included tangible features like 'bottl leak'.

The results from the collage activity demonstrate a strong support for the generalizability of hypotheses 1 and 2. Tables 10 and 13 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 like our results with French presses and supports the generalizability of hypothesis 1. It demonstrates that the method in this article can be generalised 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 14. The results indicate that the way customers perceive sustainability in products plays a role in how they like products, like 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 14 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

https://doi.org/10.1017/dsj.2022.14 Published online by Cambridge University Press

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.

Our findings reveal crucial practical implications for guiding customers to make informed purchase decisions. We proposed a novel sustainable design approach to augment existing design science approaches and recommend that designers use the method so that they may bridge the gap between perceived and engineered sustainability. For example, designers may use the proposed method to extract features perceived as sustainable for a product of interest and compare those features alongside engineered sustainability features derived from established tools like an LCA. In the case of French presses, we found a large gap between perceived and engineered sustainability while with electric scooters that gap was smaller. The proposed method provides insight to designers on what features are important to include in a sustainable product so that it is both engineered to be sustainable and perceived as sustainable by the customer, thus potentially driving purchase decisions.

The findings 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 4 and 5), resulting in noise in the extracted features. We saw this in our results, for example, 'love scooter' appears as a salient negative economic sustainability feature for electric scooters in Figure 9. There was 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 collecting reviews from a different platform that may have more balanced ratings. Moreover, certain features overlapped across sustainability pillars, such as 'love it' for social and 'love this' for environmental. We saw similar overlaps in our previous paper with French press features perceived as sustainable. Therefore, it is important to keep in mind the context for which reviews were annotated in and the possibility of machine learning noise in the extracted features. Finally, it would have been ideal to test the hypotheses on a larger set of products. While the methods are easily scalable, there were project constraints and we decided to follow a common design approach where a finite set of products are tested to investigate generalisability of the hypothesis.

It is important to address the 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,

https://doi.org/10.1017/dsj.2022.14 Published online by Cambridge University Press

since it is not always the case that a product they perceive as sustainable is sustainable.

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, two focal products were used, electric scooters and baby glass bottles, and recreated our previous work where French presses were used (El Dehaibi *et al.* 2019, 2021). Amazon reviews were collected 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. The annotations were then modelled 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. Participants were tasked with placing products along the two axes of the collage, sustainability and like, and to select from a dropdown menu features that were extracted from the machine learning model.

Based on the results we found that our previously proposed method does generalise with limitations. Crucial insights were shared that can help designers make customer-oriented design 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. For example, designers may use the method to design a sustainable lamp but collecting crowdsourced annotations of lamp reviews and extracting features perceived as sustainable. The features can then inform design decisions so that designers create a lamp that is both engineered as sustainable and also perceived as sustainable by the customer. We 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 on specific pillars of sustainability.

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 this analysis on a wider set of products with 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 the 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.

https://doi.org/10.1017/dsj.2022.14 Published online by Cambridge University Press

Financial support

We would like to thank Qatar National Research Fund (QNRF) for supporting this work. This work was funded by QNRF under the Qatar Research Leadership Program (QRLP).

References

Bag, S., **Tiwari, M. K.** & **Chan, F. T.** 2019 Predicting the consumer's purchase intention of durable goods: an attribute-level analysis. *Journal of Business Research* **94**, 408–419; doi: 10.1016/j.jbusres.2017.11.031.

Bakdash, J. Z. & **Marusich, L. R.** 2017 Repeated measures correlation. *Frontiers in Psychology* **8**, 456; doi:10.3389/fpsyg.2017.00456.

Delaney, H. D. 2003 Higher order designs with within-subjects factors: the multivariate approach. In *Designing Experiments and Analyzing Data*. Routledge.

Du, P. & **MacDonald, E. F.** 2016 Product body shapes, not features, provide fast and frugal cues for environmental friendliness. In *Volume 7: 28th International Conference on Design Theory and Methodology. ASME 2016 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Charlotte, North Carolina, USA*, p. V007T06A046. American Society of Mechanical Engineers; doi: 10.1115/DETC2016-60283.

El Dehaibi, N., **Goodman, N. D.** & **MacDonald, E. F.** 2019 Extracting customer perceptions of product sustainability from online reviews. *Journal of Mechanical Design* **141** (12), 121103; doi:10.1115/1.4044522.

El Dehaibi, N., **Liao, T.** & **MacDonald, E. F.** 2021 Validating perceived sustainable design features using a novel collage approach. In *2021 ASME International Design Engineering Technical Conferences*, p. 14. American Society of Mechanical Engineers.

Goodman, J. K. & **Paolacci, G.** 2017 Crowdsourcing consumer research. *Journal of Consumer Research* **44** (1), 196–210; doi:10.1093/jcr/ucx047.

James, G., **Witten, D.**, **Hastie, T.** & **Tibshirani, R.** 2013 *An Introduction to Statistical Learning*. Springer New York (Springer Texts in Statistics); doi:10.1007/978-1-4614-7138-7.

Kim, E.-H. & **Lyon, T. P.** 2015 Greenwash vs. brownwash: exaggeration and undue modesty in corporate sustainability disclosure. *Organization Science* **26** (3), 705–723; doi: 10.1287/orsc.2014.0949.

Kim, J., **Park, S.** & **Kim, H.** 2021 Analysis of customer sentiment on product features after the outbreak of coronavirus disease (COVID-19) based on online reviews. In *International Conference on Engineering Design, Gothenburg, Sweden, August 16–20*. ICED.

Li, X., **Zhuang, Y.**, **Lu, B.** & **Chen, G.** 2019 A multi-stage hidden Markov model of customer repurchase motivation in online shopping. *Decision Support Systems* **120**, 72–80; doi: 10.1016/j.dss.2019.03.012.

Liu, Y., **Jiang, C.** & **Zhao, H.** 2019 Assessing product competitive advantages from the perspective of customers by mining user-generated content on social media. *Decision Support Systems* **123**, 113079; doi:10.1016/j.dss.2019.113079.

MacDonald, E. F., **Gonzalez, R.** & **Papalambros, P.** 2009 The construction of preferences for crux and sentinel product attributes. *Journal of Engineering Design* **20** (6), 609–626; doi:10.1080/09544820802132428.

https://doi.org/10.1017/dsj.2022.14 Published online by Cambridge University Press

MacDonald, E. F., **Gonzalez, R.** & **Papalambros, P. Y.** 2009 Preference inconsistency in multidisciplinary design decision making. *Journal of Mechanical Design* **131** (3), 031009; doi:10.1115/1.3066526.

MacDonald, E. F. & **She, J.** 2015 Seven cognitive concepts for successful eco-design. *Journal of Cleaner Production* **92**, 23–36; doi:10.1016/j.jclepro.2014.12.096.

Maslowska, E., **Malthouse, E. C.** & **Viswanathan, V.** 2017 Do customer reviews drive purchase decisions? The moderating roles of review exposure and price. *Decision Support Systems* **98**, 1–9; doi:10.1016/j.dss.2017.03.010.

Nysveen, H. & **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), 137–150; doi:10.1016/S0167-9236(02)00212-9.

Park, S. & **Kim, H. M.** 2020 Improving the accuracy and diversity of feature extraction from online reviews using keyword embedding and two clustering methods. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Virtual, August 17–19*. ASME.

Ren, Y., **Burnap, A.** & **Papalambros, P.** 2013 Quantification of perceptual design attributes using a crowd. In *International Conference on Engineering Design, Seoul, Korea, August 19–22*. Design Society.

She, J. & **MacDonald, E. F.** 2017 Exploring the effects of a product's sustainability triggers on pro-environmental decision-making. *Journal of Mechanical Design* **140** (1), 011102; doi:10.1115/1.4038252.

The Sustainability Imperative: New Insights on Consumer Expectations 2015 *Nielsen*, pp. 1–19.

von Helversen, B., **Abramczuk, K.**, **Kopeć, W.** & **Nielek, R.** 2018 Influence of consumer reviews on online purchasing decisions in older and younger adults. *Decision Support Systems* **113**, 1–10; doi:10.1016/j.dss.2018.05.006.

Wang, Z., **Li, H.**, **Ye, Q.** & **Law, R**. 2016 Saliency effects of online reviews embedded in the description on sales: moderating role of reputation. *Decision Support Systems* **87**, 50–58; doi:10.1016/j.dss.2016.04.008.

Zhang, L., **Yan, Q.** & **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**, 113288; doi:10.1016/j.dss.2020.113288.

Zhou, F., **Ayoub, J.**, **Xu, Q.** & **Yang, X. J.** 2020 A machine learning approach to customer needs analysis for product ecosystems. *Journal of Mechanical Design* **142** (1), 011101; doi:10.1115/1.4044435.

https://doi.org/10.1017/dsj.2022.14 Published online by Cambridge University Press