SUSTAINABLESIGNALS: An AI Approach for Inferring Consumer Product Sustainability

Tong Lin[1] , **Tianliang Xu**[1] , **Amit Zac**[2,3] and **Sabina Tomkins**[1] *
[1]School of Information, University of Michigan
[2]Center for Law and Economics, ETH Zurich
[3]Law Faculty, University of Amsterdam
{tonglin, tianlix}@umich.edu, amit.zac@gess.ethz.ch, stomkins@umich.edu

Abstract

The everyday consumption of household goods is a significant source of environmental pollution. The increase of online shopping affords an opportunity to provide consumers with actionable feedback on the social and environmental impact of potential purchases, at the exact moment when it is relevant. Unfortunately, consumers are inundated with ambiguous sustainability information. For example, *greenwashing* can make it difficult to identify environmentally friendly products. The highest-quality options, such as Life Cycle Assessment (LCA) scores or tailored impact certificates (e.g., environmentally friendly tags), designed for assessing the environmental impact of consumption, are ineffective in the setting of online shopping. They are simply too costly to provide a feasible solution when scaled up, and often rely on data from self-interested market players. We contribute an analysis of this online environment, exploring how the dynamic between sellers and consumers surfaces claims and concerns regarding sustainable consumption. In order to better provide information to consumers, we propose a machine learning method that can discover signals of sustainability from these interactions. Our method, SUSTAINABLESIGNALS, is a first step in scaling up the provision of sustainability cues to online consumers.

1 Introduction

Household consumption behavior triggers a multitude of economic activities along the supply chain of each product and service, which subsequently involves the use of resources and the release of emissions. A 2015 study found that the production and use of household goods and services consumption were responsible for 60 percent of global greenhouse gas emissions and 50% to 80% of total land, material, and water use [Ivanova *et al.*, 2016]. Increasingly, retail activities including search, comparisons, and purchases are occurring online [Jaller and Pahwa, 2020; Smith and Anderson, 2016], where the convenience of online features, such as same-day shipping, comes at a heavy cost to environmental sustainability [Jaller and Pahwa, 2020]. As more people around the world enter the middle class and move their shopping habits online, the problem is worsening. Consequently, reducing the environmental impact of household consumption is crucial to reach the 2030 Agenda for Sustainable Development (Goal n.12; n.13). The problem has gained attention from policy makers as well. The European Commission conducted a study in preparation for new legislation, assessing 150 claims about products' environmental characteristics and concluded that 53% of them contained vague, misleading, or unfounded information, and 40% of them were unsubstantiated, defining these claims as a form of greenwashing [Abnett, 2023].

Determining the environmental impact of a product is a time-intensive and expensive endeavor. Typically, this is done with a life-cycle assessment (LCA), where a product's impact is evaluated from cradle-to-grave (amongst other variants). Unfortunately, LCA procedures are complex, time-consuming, and costly. To affect household consumption at scale, one needs to generate relevant and reliable information about a large and growing set of products, across product categories. Previous attempts, like third-party certification systems, are often abused and are likewise dependent on domain experts [Brad *et al.*, 2018].

In the absence of an LCA, we ask, given a set of scores of products' sustainability, can an algorithm learn to predict these scores from noisy online signals, like reviews and product descriptions? As these are both textual signals, we wonder how well large language models (LLMs) can perform in this setting. To the extent to which online signals can be used to label products in terms of climate and social responsibility, they can assist more in-depth LCA systems by focusing attention on products that are difficult to label, providing more knowledge to consumers in years to come.

Consumers infer product quality, when such quality is not easily observable, by processing multiple-quality cues in the environment [Akdeniz *et al.*, 2013]. Yet, inferring the sustainability of a product in an online environment is even more challenging. First, the online environment includes adversarial and conflicting sources of information. Sellers can post claims about a product, including photos and descriptions. On the other hand, previous buyers can counter these claims through ratings and reviews. It is not immediately clear which components, if any, of this exchange will be useful for the

*Corresponding author.

task of inferring sustainability signals. Second, the online environment is controlled by a third-party platform, which might mask or boost signals of quality according to its own profit maximization strategy [Farronato *et al.*, 2023]. Third, sustainability quality specifically, is difficult to infer as it reflects an externality problem. That is costs are not directly internalized in the price of the product and by the manufacturer, but are instead sustained by society at large over multiple generations [Hardin, 1968].

Our Contributions. To study how different signals in the online environment can be used to infer sustainability information we investigate a sample of products from Amazon.com. We separately explored sellers' claims of sustainability in product descriptions and consumers' reviews, to disentangle potentially contradicting signals. For detecting environmental claims we used a pre-trained model and annotated text of environmental claims made by publicly listed firms [Stammbach *et al.*, 2022], while for the consumers' reviews, we developed a new annotation task to explore what aspects of sustainability are important to consumers. In the last stage, we train a LLM to predict the sustainability score of products from the Amazon environment using labeled data from public resources. We make our tools accessible to researchers and activists interested in the field [1].

1.1 Online Shopping Dataset

Throughout this paper, we utilize a dataset of online consumer goods. We collect information from the product page, including full-text descriptions, all reviews appearing on the first page, prices, and ratings. Our dataset contains this product information and a sustainability score for 9512 products listed on Amazon.com. In contrast to prior work which has only concentrated on a single product type, the products in this dataset are distributed across four categories: BABY PRODUCTS, BEAUTY & PERSONAL CARE, HEALTH & HOUSEHOLD, and HOME & KITCHEN. The sustainability scores capture a range of concerns: Health Impacts, Product Durability, Product Materials, and Product Packaging. We believe this dataset is large enough to inform foundational work into the feasibility inferring product sustainability online at scale. In future work, we are exploring broadening this dataset to cover new categories, and on developing an open-source tool where contributors can update sustainability criteria.

2 Environmental Claims

In response to consumers' environmental concerns, as well as pressure from other sources, such as the media, academia, stockholders, and government [Shevchenko *et al.*, 2016], companies are increasingly incorporating environmental claims into their marketing materials [Clementino and Perkins, 2021; de Freitas Netto *et al.*, 2020b]. When factual and relevant, these claims can be useful indications of environmental sustainability. However, many of these claims may be examples of greenwashing [de Freitas Netto *et al.*, 2020a] which use information asymmetry to the advantage of the firm, providing signals which might not be particularly relevant to the consumer and our analysis. Here, our goal is to use an open-source claims classifier to explore the prevalence of claims in our dataset.

The European Commission (EC) defines claims as follows: *Environmental claims refer to the practice of suggesting or otherwise creating the impression (in the context of a commercial communication, marketing or advertising) that a product or a service is environmentally friendly (i.e., it has a positive impact on the environment) or is less damaging to the environment than competing goods or services.*[2] In the context of consumers' increasing willingness to pay for environmentally friendly products[Nielsen Media Research, 2015], sellers' boasting about environmental credentials can also be regarded as a selling point[de Freitas Netto *et al.*, 2020a].

To identity claims we utilize an opensource dataset and classifer [Stammbach *et al.*, 2022]. This claims classifier was trained on 3K claims made by European firms, particularly in finance. Each claim was annotated by 16 human experts as being an environmental claim or not. For each product we pass each sentence in the description through the classifier. We then perform max pooling to combine the resulting sentence level embeddings. This results in a single embedding for each product description which is passed through the final layer of the claims classifier, delivering a probability of a claim. Any description with a final probability greater than .5 is labeled as containing a claim.

In contrast to the original setting of finance, we are interested in the claims made in product marketing. This setting will naturally vary from the setting of financial firm claims. One key difference is that the product claims tend to be shorter, and may be sort phrases rather than the full sentences explored in the claims dataset. Thus, while we utilize this high quality and publically available data to explore claims in the setting of product marketing, we were cautious that some claims were bound to be lost in translation between the two settings. However, we see in Figure 1 that the claims discovered by this classifier seem reasonable.

- Negative Examples (No Environmental Claims Found)
  - MONEY BACK GUARANTEE - We think you're going to love this White Classic Hospital bed Sheets as much as we do! But in case you don't, you are covered by our 30 day, no questions adked, money back guarantee. We want our customers to be 100% happy.
  - FAMILI ESSENTIAL - Of course adults can use as well. A variety of functions to make life more convenient.
- Environmental Claim Examples
  - BRAND PRODUCTS - brand products are made from quality raw materials with minimal wastage at every step of production. With the goal of achieving a neutral carbon footprint, please recylce and help leave Mother Earth better off for future generations.
  - Guilt-free delivery: These environmentally-friendly bamboo paper towels arrive in a 100% post-consumer recycled box. Every part of the packaging is plastic-free and compostable (even the box tape!). And we 2x offset all carbon emissions from transportation.

Figure 1: The presence and absence of environmental claims.

[1] Code available at:
https://github.com/Sabina321/sustainable_signals.

[2] From the Commission Staff Working Document, Guidance on the implementation/application of Directive 2005/29/EC on Unfair Commercial practices, Brussels, 3 December 2009 SEC(2009) 1666. See section 2.5 on misleading environmental claims.

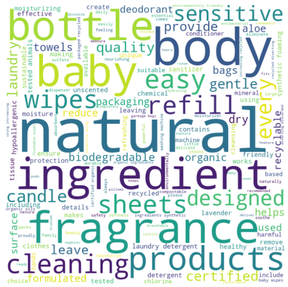

Figure 2: Keywords in environmental claims.

Using the claims classifier we found that of a total of 9512 products in our dataset, 1055 are considered to have environmental claims in their descriptions. We disclose common keywords from these claims in Figure 2. We caution that the high prevelance of environmental claims in our dataset is not representative of the online shopping marketplace. However, it does support our hypothesis that as customers parse product claims when making purchasing decisions, they must make sense of environmental claims and weigh these claims against other signals of sustainability.

3 Environmental Concerns

Do consumers think about the environment when they make online purchases? While concern over sustainability has risen over time [Szabo and Webster, 2021; Román-Augusto *et al.*, 2022], this may not translate to a quotidian concern over the health of the environment when reviewing a product purchase. Online reviews are a novel source of large data to mine for insights into consumer behavior [Archak *et al.*, 2011; Jain *et al.*, 2021]. We offer a unique study into concerns over sustainable consumption, adding to a limited body of knowledge [Brazytė *et al.*, 2017; El Dehaibi *et al.*, 2019]. Here, we explore the following questions a) do consumers bring up the environment in their reviews either directly or through relevant related topics? b) do they tend to bring up the environment when they are satisfied or dissatisfied with a product? c) in what granular context do they bring up the environment?

To address these questions we designed an annotation task and recruited knowledgeable students to annotate consumer product reviews, in a process described in Section 3.1. Before embarking on this study we first used off-the-shelf topic modeling tools to explore whether environmental topics appeared to be present in the reviews. This initial exploration yielded the presence of the following topics: materials, perceived sustainability, quality, and health. To verify these topics and discover topics that might only appear to someone with domain knowledge, we turned to human annotators.

3.1 Study Design

We created an online annotation task where annotators complete the following subtasks for each review:

$ST1$ Determine if the review contains terms relevant to sustainability.

$ST2$ Determine whether the review appears to be authentic.

$ST3$ Determine if the reviewer's expectations of the product were met.

$ST4$ Determine if the reviewer mentions fine-grained environment concerns of: *product materials*,*health impacts*, *product durability*, *product packaging*.

$ST5$ For each fine-grained concern determine which words and phrases are used, and if the reviewer is expressing a positive or negative sentiment with these words and phrases.

Screening Task. Before sending annotators the primary annotation task we ask that they complete a training tutorial. At the end of the tutorial they complete a short quiz. All annotators who do not pass the quiz are not asked to complete the annotations. The tutorial describes each task in detail and steps annotators through several examples for each subtask. The short quiz contains 12 questions and covers all types of subtasks described above. A passing score is assigned if 9 questions are answered correctly.

Recruitment. All annotators are graduate students from relevant backgrounds (environmental science and/or a program in sustainability) and have taken at least one core course in their program. Annotators were reimbursed at an estimated rate of $18 / hour. This study was deemed unregulated by an IRB, as annotators are not being studied, but only completing an annotation task [3].

Sampling Reviews. Our total dataset contains 68,390 reviews. To select reviews for annotation we set up the following sampling scheme. We prepared a keyword list which contains $\sim$170 keywords which are relevant to sustainability, and then counted the keywords in each review. If a review contains more than 3 keywords, we then labeled this review as potentially relevant to sustainability. Approximately 8,300 reviews fit this description. We then randomly sampled 700 reviews from the 8,300 potentially relevant reviews and sampled 300 reviews from all the non-relevant reviews respectively, and combined them together to form the review pool for the annotation task.

Implementation Details. Both the screening task and the primary annotation task are deployed on Qualtrics. We distributed the training tutorial to 33 annotators and 24 of them passed the screening quiz. Each annotator was asked to annotate 6-7 sections of reviews. For each section we replaced the last review with the first review to check annotators' attention. If an annotator annotated the same review inconsistently, their responses for that block were discarded.

Annotated Dataset. To guarantee the reliability of the obtained annotations, we check the inter-rater agreement within each section. Out of 50 sections, there are 7 sections without any annotation due to the attention check; besides, 15 sections have annotations from one annotator, 11 sections have annotations from two annotators with an average agreement score of 0.432, 11 sections have annotations from three annotators with an average agreement score of 0.183, and 4 sections have annotations from four annotators with an average agreement

[3] Exempt determination made by University of Michigan IRB.

of 0.453. We used majority vote if the annotators have a disagreement on the same question. This results in annotations for 794 reviews.

3.2 Study Results: Environmental Concerns

For our purposes it was useful to bias our sample towards reviews likely to be relevant to sustainability. We found that approximately 40% of all annotated reviews were deemed relevant to sustainability. As a result of our sampling, this number is likely to be unrepresentative of all reviews online. However, it suggests that consumers do voice concerns over the environmental impact of their purchases online.

To obtain a more realistic estimate of the overall presence of these concerns, we trained a predictive model to classify reviews as potentially relevant to sustainability or not. The input to the model was a bag-of-words representation of the reviews, and the output was a binary label of relevant to sustainability or not. We chose the optimal hyper-parameters using the model's precision in order to arrive at the most conservative estimate of whether a review is relevant to sustainability. Then, using this model to infer relevance to sustainability for the remaining reviews, we found that roughly 15% were predicted to be potentially relevant to sustainability.

Next, we calculated the positive-negative ratio in each topic (Figure 3) by counting the number of positive and negative highlighted words in each topic respectively. From Figure 3 we can infer that consumers tend to have positive sentiments while commenting on topics like product materials, health impacts and product durability; in contrast, people are more likely to give negative comments while discussing product packaging.

We then plotted the distributions for highlighted words of four topics in different categories in Figure 4. We found that a majority of consumers' concerns towards sustainability focus on the product materials across all categories. Besides, consumers tend to pay more attention to health impacts while purchasing products within the BEAUTY & PERSONAL CARE category. Another interesting finding is that when purchasing products in the BABY PRODUCTS category, concerns over packaging are relatively diminished. Instead, shoppers are more interested in product materials (e.g. whether the materials are harmful) and health impacts (e.g. whether children may be sensitive to the products).

Table 1 shows relevant keywords within each product category and environmental concern. The product categories of materials, durability and packaging, all offer straightforward interpretations. Annotators marked words to do with poor materials, durability and packaging as revealing negative sentiment, and words to do with specific environmental keywords such as organic, durable, and minimal packaging as expressing positive sentiment. Health concerns can activate consumers concern for sustainability, as there is a belief that environmentally friendly products are safer [Kim and Seock, 2009]. These highlighted words illuminate this phenomenon, showing phrases such as chemical free were frequently used positively, whereas negative phrases illuminate consumers' health concerns such as hormonal disorders.

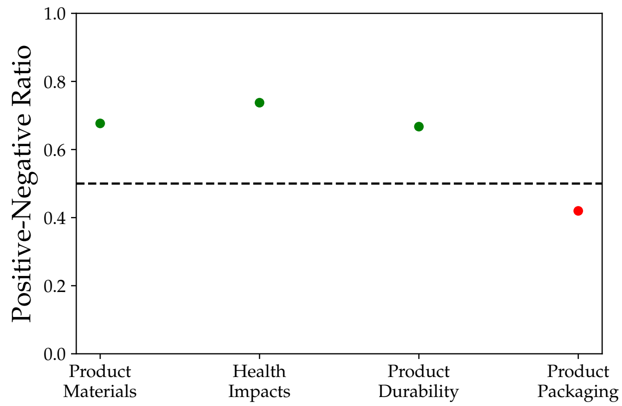

Figure 3: Across three of the four topics positive sentiment is more common than negative. The exception is product packaging, which consumers tend to bring up in a more negative context.

4 Conflicting Signals: Claims vs Concerns

An open question is whether or not the reviews can serve as a counterweight to environmental marketing claims. To test this we inspect the correlation between the sustainability score and the sentiment of the reviews for each product for which a claim is present in the description. We find a weak correlation between the sentiment (where 1 is positive and 0 is negative) and the sustainability score when a claim is present [4], of 0.043. As these reviews may be unrelated to sustainability, we also constructed a rough sentiment score using the dataset described in Section 3.1. We first construct two sets, one of all of the words the annotators tagged as being negatively related to sustainability, and one of all the words described as being positively related to sustainability. We then compute a ratio of the number of negative words across all reviews to the number of positive words across all reviews for each product.

In Figure 5, we see that as the ratio increases the sustainability score decreases. This is true regardless of whether a claim is present, i.e. claims do not mask the signal from reviews. This suggests that we can mine the information in customer reviews to serve as a counterpoint to the claims that companies make.

5 Inferring Sustainability Scores

Our goal is to solve the regression task of inferring a product's sustainability score given easily obtainable product features and category information. Given a dataset of $\mathbb{D} = \{(x_1, y_1, c_1), \ldots, (x_N, y_N, c_N)\}$ where y_i is a sustainability score $\in [0, 10]$, x_i belongs to a multimodal feature space $\mathcal{X}$ which depends on the available data, and c_i represents a product category from $\mathcal{C} = \{c^1, \ldots c^k\}$, our goal is to learn a function $f : (\mathcal{X}, \mathcal{C}) \rightarrow \mathbb{R}^d$. Alternatively, we consider the goal of inferring sustainability labels without product categories $f : \mathcal{X} \rightarrow \mathbb{R}^d$. In this setting, each x_i contains subcomponents (x_i^r, x_i^d). Each x_i^r is a product review and each x_i^d is a

[4]To compute sentiment we used the package Flair [Akbik *et al.*, 2019]

Sentiment	Topic	Frequent Highlighted Words
Positive	Product Materials	natural, ingredients, organic, quality
	Health Impacts	chemical free, sensitive, safe, gentle
	Product Durability	long lasting, durable, strong, time
	Product Packaging	packaged well, less waste, minimal packaging
Negative	Product Materials	chemicals, rough feeling, poor quality, cheap
	Health Impacts	burning, harsh, hormonal disorders, neurotoxin
	Product Durability	broken, brittle, plastic, rubbish
	Product Packaging	plastic, broken, leaked, glass

Table 1: Discovered topics pertaining to sustainability in product reviews. Each topic's assigned category is in the far right column.

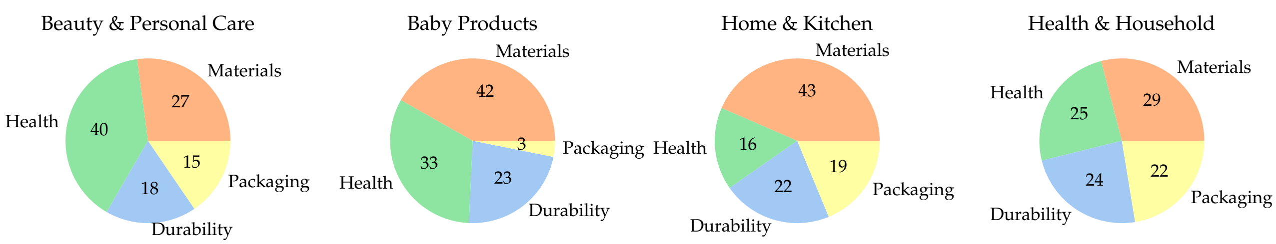

Figure 4: Distributions of four topics within each category.

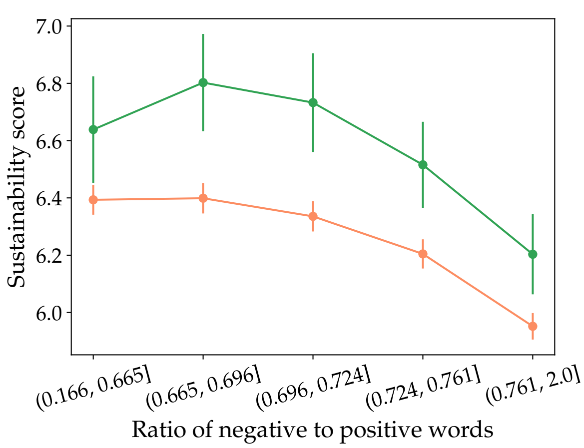

Figure 5: As the reviews increase in negativity, the sustainability score decreases, even in the presence of claims.

product description. Each product belongs to a single product category.

Large Language Models

Our approach is agnostic to the exact large language model deployed, allowing us to experiment with different choices. For example, a natural choice for this setting may be CLIMATEBERT[Webersinke *et al.*, 2022], as it was developed for somewhat relevant tasks related to climate change. Additionally, we evaluate our approach on the state-of-the-art models DISTILBERT[Sanh *et al.*, 2019], and DISTILROBERTA[Liu *et al.*, 2019; Sanh *et al.*, 2019].

Webersinke *et al.* utilized a domain-adaptive technique and proposed CLIMATEBERT, a transformer-based language model pretrained on 2 million paragraphs of climate-related texts from sources such as common news, research articles, and climate reporting. Their model has remarkable achievements on various climate-related tasks such as text classification, sentiment analysis, and fact checking

DISTILBERT exploits recent developments in deep neural network (DNN) language models. DISTILBERT builds on the recent advances of applying transfer learning and deep neural networks to Natural Language Processing (NLP) tasks. The DNNs can then be used with either no additional training, or with some customization on a specific dataset. While additional training is preferable, retraining large language models can be computationally expensive [Wu *et al.*, 2019]. DISTILBERT offers a lightweight version of BERT [Devlin *et al.*, 2018] which can be updated with new data more efficiently.

DISTILROBERTA is a process for training BERT which results in a new large language model. While much of the structure is the same as BERT, the authors of DISTILROBERTA have shown that by training for a long time and on large batches of data, their pre-trained model can outperform BERT on many NLP tasks.

We propose SUSTAINABLESIGNALS, a method which utilizes on LLMs to extract sustainable signals from product descriptions and customer reviews. These LLMs should be complex enough to learn from even conflicting signals, a hypothesis we explore in Section 6. SUSTAINABLESIGNALS is able to fine-tune predictions to specific product categories,

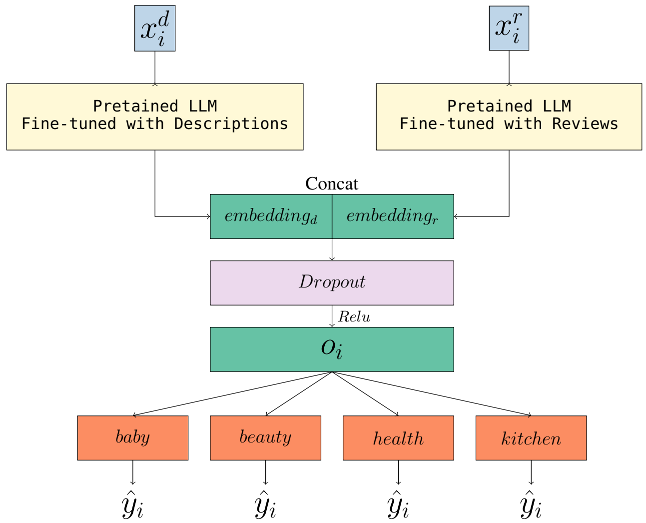

Figure 6: Architecture for SUSTAINABLESIGNALS. Note that each o_i will travel along exactly one path.

while exploiting textual similarities across them.

In SUSTAINABLESIGNALS, we pass both descriptions and reviews into their own pretrained LLM layer each of which outputs a fine-tuned embedding. We concatenate the textual representations before passing them through Dropout and Relu layers. At this point, we form an intermediate representation o_i. This o_i shares information across the four product categories in a global embedding. Before predicting the score for each product, we learn a category-aware representation. This allows us to express the variance we observe in correlations between sustainability and different features across product categories. For example, many products in both the BEAUTY & PERSONAL CARE and BABY PRODUCTS categories contain similar chemicals which can have negative health and environmental impacts. Thus, for each x_i the label y_i depends on c_i:

$$y_i = \begin{cases} \texttt{baby}(o_i), & \text{if } c_i \equiv \textsc{Baby Products} \\ \texttt{beauty}(o_i), & \text{if } c_i \equiv \textsc{Beauty \& Personal Care} \\ \texttt{health}(o_i), & \text{if } c_i \equiv \textsc{Health \& Household} \\ \texttt{kitchen}(o_i), & \text{if } c_i \equiv \textsc{Home \& Kitchen} \end{cases}$$

where each of the category blocks consist of two fully connected linear layers, and the number of inner nodes depends on the category.

6 Experiments

To test the ability of a machine learning model to infer whether a product is sustainable, we utilized an openly available product sustainability score, called a Finch score. Finch has developed a plugin for Amazon, scoring products in terms of sustainability from 0 to 10. However, we found in practice that the scores follow a Normal distribution with most (25%-50%) of the scores being in the range of 5-7. As this was the most comprehensive dataset we could find for products (rather than brands), we focus our attention on the online marketplace of Amazon.com, a reasonable choice given it's online popularity.

	MSE
Baseline	1.173
Lasso	0.848
Gradient Boosting	0.818
NOCATE(RB)	0.776
NOCATE(CB)	0.763
NOCATE(DB)	0.756
SUSTAINABLESIGNALS(RB)	0.762
SUSTAINABLESIGNALS(CB)	0.753
SUSTAINABLESIGNALS(DB)	**0.736**

Table 2: We see that the deep learning models achieve the highest performance, and that incorporating a category-aware representation yields the best results. Our approach performs the best with DISTILBERT, yielding a statistically significant improvement (paired t-test, $p < .05$) over Gradient Boosting and Lasso.

		Avg. True Score	Avg. Predicted Score	Avg. MSE
Has claims	High word ratio	6.32	6.49	0.63
	Low word ratio	7.11	6.95	0.74
No claims	High word ratio	6.05	6.07	0.67
	Low word ratio	6.28	6.39	0.81

Table 3: High/low word ratio refers to a higher/lower than the median ratio of negative to positive sustainability relevant key words across all reviews for this product.

Comparators. We compared our proposed approach to several machine learning models, here we show the two with the best performance: Lasso and Gradient Boosting. We also compare SUSTAINABLESIGNALS to NOCATE, a deep learning model which does not differentiate between product categories. and report the performance of predicting the average score for each product (Baseline). Our proposed method depends on large language models, and we experiment with three variations: DISTILROBERTA(RB), CLIMATEBERT(CB), and DISTILBERT(DB).

Implementation. We perform 5 fold cross-validation to search over the hyperparameters. Here, we search over the number of nodes for each category specific block. For each block we try either 512 or 768 inner nodes. We also choose the learning rate for the deep learning models during the validation stage, searching over the values of $[.00001, .0001, .001, .01, .1, 1]$. For the baselines we also searched over the learning rate, and for the GradientBoostedRegressor the number of estimators ($[1, 50, 100]$), and for the Lasso Regression the coefficient alpha ($[.001, .01, .1, 1, 10, 100, 1000]$). All results in in Table 2 are on the final held out test set which comprises 10% of the dataset with 952 labeled products. All experiments are run with 8 24 GB Nvidia GPUs. Each of the large language models was implemented in Hugging Face [Wolf *et al.*, 2020], and all of the deep learning was done in PyTorch [Paszke *et al.*, 2019]. Here, we favored lighter-weight language models, however, fine-tuning the LLM models comprised the majority of the computational burden for our approach.

Conflicting Signals and Error Heterogeneity. A concern with this problem setting is that models will over-rely on companies' claims, assigning overly-optimistic scores to products companies advertise as environmentally friendly.

Overall, we find that the model has a positive bias, even in the absence of claims. As we found in Section 4, the customer reviews can serve as a counterweight to these claims. We see in Table 3 that SUSTAINABLESIGNALS, learns to use these reviews as well, not only producing more conservative sustainability estimates, but reducing error.

7 Discussion

We begin by analyzing two potentially conflicting sources of sustainability signals, environmental claims and concerns. Environmental claims are present in roughly 10% of all product descriptions, and it is not uncommon for reviews to contain environmental concerns. Some reviews counter the product's claims, particularly in the fine-grained subjects of durability and packaging, where consumers most often voice that products are not durable and that the packaging is not sustainable. We hope that our initial study into environmental concerns can serve as a reference point for future work, and stress that such work should be conducted longitudinally, to measure changing attitudes towards the environment in general, and sustainable consumption in particular.

Additionally, we investigate the extent to which a machine learning model can learn to infer sustainability scores in this online environment. We find that the average MSE is less than 1 when using our proposed model SUSTAINABLESIGNALS. This is a promising first result, showing that in general the inferred score is close to the true score. However, there is much to be explored in future work. While a deep learning model reduces the MSE by 10% over a classic model and by 37% over the simplest baseline, future improvements can likely further boost performance gains.

8 Contributions to Related Work

Given the potential impact of AI systems towards reducing the environmental burden of consumption, there is relatively little work in this area. Existing work falls in several categories: mining sustainability cues from online data, inferring sustainability labels and integrating machine learning into the LCA process. These approaches are challenged by the availability of high quality ground truth sustainability labels. When the goal is to inform LCA or product design it is not clear if online data sources can provide the requisite fine-grained data on a product's attributes and life-cycle.

Mining Sustainability Cues From Online Data. El Dehaibi *et al.* discover sustainability perceptions from product reviews and find that consumers' perceptions of sustainability are not necessarily grounded in fact. Their work is an important contribution towards understanding consumer perceptions with reviews, however, they consider only one type of product. Our work spans product categories and uses the signal from reviews to infer overall sustainability scores. Similarly, [Saidani *et al.*, 2021] propose online reviews as a source of information for informing product design.

Inferring Sustainability Labels in Online Settings. There is limited work inferring product sustainability in online settings. One approach uses a probabilistic model to infer latent sustainability scores for Amazon grocery products [Tomkins *et al.*, 2018]. However, they do not use labeled data and are restricted to a single product category, while also requiring proprietary shopping history. Kauffmann et al. utilize NLP techniques to classify products based on positive/neutral/negative sentiment features [Kauffmann *et al.*, 2019]. However, their work focus on inferring sentiment from reviews in the online market for cell phones only, without directly predicting sustainable features. To the best of our knowledge, there is no existing AI approach which can predict sustainability scores for multiple categories. Existing approaches rely on generally proprietary data from a single domain, for example fashion or food products. In contrast, we develop a multi-task model which can use easily obtained information from product pages to infer labels in multiple product categories. We show that this approach can learn from potentially noisy online signals to infer third-party sustainability categorizations.

Integrating Machine Learning with LCA. While work on scaling LCA to perform on new products with limited human supervision is very sparse, there is a more active area of directly integrating machine learning into the work-flow of human LCA analysts. Wistoff et al. utilized machine learning to score products by various LCA metrics given physical attributes [Wisthoff *et al.*, 2016]. This approach is promising as a tool within LCA, however it requires data not available from online shopping sites. Satinet and Fouss select meaningful life cycle characteristics, integrate them with the results of LCAs from 75 clothing products and further apply 9 supervised machine learning algorithms to classify products' sustainability [Satinet and Fouss, 2022]. However, their focus is mainly on the clothing industry; in addition, the LCA results are not available online. Bracke et al. select sustainability-related parameters of devices, applying cluster analytics to identify those devices with similar features and further utilizing data envelopment analysis (DEA) to extrapolate the energy efficiency of these devices [Bracke *et al.*, 2017]. They mainly focus on electronic devices with physical device parameters which could be hard to collect in practice.

9 Conclusion

In order to promote sustainable consumption we must be able to assess the sustainability of different products. While gold standards like LCA can do this reliably, they cannot do it at scale. We show that AI models can infer sustainability scores from conflicting signals in the online shopping environment. We propose this as a first step for developing open data sets which can be used in conjunction with machine learning models, to communicate high standards of sustainability at scale.

This work relies on a single metric of sustainability. Surprisingly, there are only a few open-source datasets containing fine-grained information on environmental metrics which can be mapped to products found on popular shopping platforms. As a next step, we are developing a method for collecting fine-grained environmental indicators in online marketplaces. These indicators can be used to infer more nuanced measures of sustainability. Finally, we stress that although we have focused on AI's ability to discover relevant sustainability information to inform responsible consumption, we also believe it can be used to avoid consumption.

Contribution Statement

Tong Lin and Tianliang Xu contributed equally to this paper.

References

[Abnett, 2023] Kate Abnett. EU Plans Law Forcing Companies to Prove Green Claims Are Real-Draft. *Reuters*, 2023.

[Akbik *et al.*, 2019] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 54–59, 2019.

[Akdeniz *et al.*, 2013] Billur Akdeniz, Roger J Calantone, and Clay M Voorhees. Effectiveness of Marketing Cues on Consumer Perceptions of Quality: The Moderating Roles of Brand Reputation and Third-Party Information. *Psychology & Marketing*, 30(1):76–89, 2013.

[Archak *et al.*, 2011] Nikolay Archak, Anindya Ghose, and Panagiotis G Ipeirotis. Deriving the pricing power of product features by mining consumer reviews. *Management science*, 57(8):1485–1509, 2011.

[Bracke *et al.*, 2017] Stefan Bracke, Berna Ulutas, and Christoph Rosebrock. Concept for Analysing Product Sustainability Regarding Eco-related Product Perception and Efficiency Within a Product Spectrum. *Procedia Manufacturing*, 8:28–35, 2017.

[Brad *et al.*, 2018] Alina Brad, Alice Delemare, Natasha Hurley, Valerie Lenikus, Rachel Mulrenan, Noemi Nemes, Urska Trunk, and Urbancic Urbancic. The false promise of certification. *Changing Markets Foundation, Utrecht*, 2018.

[Brazytė *et al.*, 2017] Karolina Brazytė, Fabian Weber, and Dorothea Schaffner. Sustainability management of hotels: how do customers respond in online reviews? *Journal of Quality Assurance in Hospitality & Tourism*, 18(3):282–307, 2017.

[Clementino and Perkins, 2021] Ester Clementino and Richard Perkins. How do companies respond to environmental, social and governance (ESG) ratings? Evidence from Italy. *Journal of Business Ethics*, 171:379–397, 2021.

[de Freitas Netto *et al.*, 2020a] Sebastião Vieira de Freitas Netto, Marcos Felipe Falcão Sobral, Ana Ribeiro, and Gleibson Robert da Luz Soares. Concepts and forms of greenwashing: a systematic review. *Environmental Sciences Europe*, 32:1–12, 2020.

[de Freitas Netto *et al.*, 2020b] Sebastião Vieira de Freitas Netto, Marcos Felipe Falcão Sobral, Ana Regina Bezerra Ribeiro, and Gleibson Robert da Luz Soares. Concepts and forms of greenwashing: A systematic review. *Environmental Sciences Europe*, 32(1):1–12, 2020.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[El Dehaibi *et al.*, 2019] Nasreddine El Dehaibi, Noah D Goodman, and Erin F MacDonald. Extracting Customer Perceptions of Product Sustainability From Online Reviews. *Journal of Mechanical Design*, 141(12), 2019.

[Farronato *et al.*, 2023] Chiara Farronato, Andrey Fradkin, and Alexander MacKay. Self-Preferencing at Amazon: Evidence from Search Rankings. 2023. NBER Working Paper 30894.

[Hardin, 1968] Garrett Hardin. The Tragedy of the Commons. *Science*, 162(3859):1243–1248, 1968.

[Ivanova *et al.*, 2016] Diana Ivanova, Konstantin Stadler, Kjartan Steen-Olsen, Richard Wood, Gibran Vita, Arnold Tukker, and Edgar G Hertwich. Environmental Impact Assessment of Household Consumption. *Journal of Industrial Ecology*, 20(3):526–536, 2016.

[Jain *et al.*, 2021] Praphula Kumar Jain, Rajendra Pamula, and Gautam Srivastava. A systematic literature review on machine learning applications for consumer sentiment analysis using online reviews. *Computer Science Review*, 41, 2021. Article 100413.

[Jaller and Pahwa, 2020] Miguel Jaller and Anmol Pahwa. Evaluating the environmental impacts of online shopping: A behavioral and transportation approach. *Transportation Research Part D: Transport and Environment*, 80, 2020. Article 102223.

[Kauffmann *et al.*, 2019] Erick Kauffmann, Jesús Peral, David Gil, Antonio Ferrández, Ricardo Sellers, and Higinio Mora. Managing Marketing Decision-Making with Sentiment Analysis: An Evaluation of the Main Product Features Using Text Data Mining. *Sustainability*, 11(15), 2019.

[Kim and Seock, 2009] Soyoung Kim and Yoo-Kyoung Seock. Impacts of health and environmental consciousness on young female consumers' attitude towards and purchase of natural beauty products. *International Journal of Consumer Studies*, 33(6):627–638, 2009.

[Liu *et al.*, 2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*, 2019.

[Nielsen Media Research, 2015] Nielsen Media Research. The sustainability imperative. https://nielseniq.com/global/en/insights/analysis/2015/the-sustainability-imperative-2/, 2015. Accessed: 2022-07-04.

[Paszke *et al.*, 2019] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in*

Neural Information Processing Systems (NeurIPS), pages 8024–8035. 2019.

[Román-Augusto *et al.*, 2022] Jose Antonio Román-Augusto, Camila Garrido-Lecca-Vera, Manuel Luis Lodeiros-Zubiria, and Martin Mauricio-Andia. Green Marketing: Drivers in the Process of Buying Green Products—The Role of Green Satisfaction, Green Trust, Green WOM and Green Perceived Value. *Sustainability*, 14(17):10580, 2022.

[Saidani *et al.*, 2021] Michael Saidani, Harrison Kim, and Bernard Yannou. Can Machine Learning Tools Support the Identification of Sustainable Design Leads From Product Reviews? Opportunities and Challenges. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference (IDETC-CIE)*, 2021.

[Sanh *et al.*, 2019] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[Satinet and Fouss, 2022] Chloe Satinet and François Fouss. A Supervised Machine Learning Classification Framework for Clothing Products' Sustainability. *Sustainability*, 14(3), 2022.

[Shevchenko *et al.*, 2016] Anton Shevchenko, Moren Lévesque, and Mark Pagell. Why Firms Delay Reaching True Sustainability. *Journal of Management Studies*, 53(5):911–935, 2016.

[Smith and Anderson, 2016] Aaron Smith and Monica Anderson. Online shopping and e-commerce. https://www.pewresearch.org/internet/2016/12/19/online-shopping-and-e-commerce/, 2016. Accessed: 2023-02-10.

[Stammbach *et al.*, 2022] Dominik Stammbach, Nicolas Webersinke, Julia Anna Bingler, Mathias Kraus, and Markus Leippold. A Dataset for Detecting Real-World Environmental Claims. *arXiv preprint arXiv:2209.00507*, 2022.

[Szabo and Webster, 2021] Szerena Szabo and Jane Webster. Perceived Greenwashing: The Effects of Green Marketing on Environmental and Product Perceptions. *Journal of Business Ethics*, 171:719–739, 2021.

[Tomkins *et al.*, 2018] Sabina Tomkins, Steven Isley, Ben London, and Lise Getoor. Sustainability at Scale: Towards Bridging the Intention-Behavior Gap with Sustainable Recommendations. In *Proceedings of the ACM Conference on Recommender Systems (RecSys)*, pages 214–218, 2018.

[Webersinke *et al.*, 2022] Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. Climatebert: A pretrained language model for climate-related text. *arXiv preprint arXiv:2110.12010*, 2022.

[Wisthoff *et al.*, 2016] Addison Wisthoff, Vincenzo Ferrero, Tony Huynh, and Bryony DuPont. Quantifying the impact of sustainable product design decisions in the early design phase through machine learning. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference (IDETC-CIE)*, 2016.

[Wolf *et al.*, 2020] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, 2020.

[Wu *et al.*, 2019] Honghan Wu, Karen Hodgson, Sue Dyson, Katherine I Morley, Zina M Ibrahim, Ehtesham Iqbal, Robert Stewart, Richard JB Dobson, Cathie Sudlow, et al. Efficient Reuse of Natural Language Processing Models for Phenotype-Mention Identification in Free-text Electronic Medical Records: A Phenotype Embedding Approach. *JMIR Medical Informatics*, 7(4):e14782, 2019.