# Sentiment Analysis of Nike Shoes Using Machine Learning: BERTopic


## Abstract

This project analyzes the consumer perception and market performance of Nike's sustainable versus non-sustainable[1] footwear. Utilizing web scraping techniques, a detailed dataset was created capturing product IDs, prices, discounts, number of reviews, and ratings for both sustainable and non-sustainable shoes. The analysis included conducting statistical tests such as a t-test and chi-square test to compare the distributions of product ratings and reviews, providing insight into customer satisfaction and engagement across the two categories of analysis. The relationship between price, ratings, and review counts was explored for both sustainable and non-sustainable shoes through scatter plots. The core methodology included clustering product reviews and descriptions using BERTopic, allowing us to assess key sentiments associated with each product type. Our findings indicate that customer sentiments focus on comfort and preference, regardless of the shoe's sustainability. Additionally, the alignment between consumer sentiment and product description was strong, suggesting that Nike's product marketing strategies effectively resonate with consumers. The results suggest that Nike has the opportunity to enhance its investment in sustainable products, aligning more closely with environmental goals and consumer demand.


## Background

The fashion industry is a major contributor to global greenhouse gas emissions, responsible for an estimated 10% of global carbon emissions, according to a 2024 European Parliament report on textile production and waste [1]. As social media influences consumer trends and drives demand for the latest fashionable items, there is a growing need for more sustainable fashion production and consumption practices. Our project focuses on the largest shoe company in the world, Nike, which reported Q1 earnings of 12.4 billion dollars in 2024, according to their SEC report [2]. The project aims to understand customer sentiment regarding sustainability, offering insights into material choice and customer demand. In addition, the results provide insight to companies regarding the materials they use and their impact on consumer demand. As a global company with significant influence in the fashion industry, Nike's sustainability practices have a global impact. However, a significant gap in the sustainability field is the issue of determining whether an item is sustainable or not. The appropriate data is often unavailable to determine if a product is sustainable. For example, the lack of specific data from the Environmental Protection Agency's Supply Chain Greenhouse Gas Emission Factors, according to the NAICS-6, does not have different categories for different footwear types and categorizes all shoes under the

[1] Throughout the project, we opted for the term "non-sustainable" instead of "unsustainable." This choice was made to create a clearer contrast with "sustainable," thereby emphasizing the distinct categories within our analysis.

umbrella term “316210 Footwear Manufacturing”[3]. Thus, it is hard to determine the environmental impact of producing shoes.

Nike has made considerable efforts to reduce carbon waste through its definition and use of ‘sustainable materials’ across its product lines [4]:

- Nike Air soles consist of at least 25% recycled manufacturing waste, and all Air MI facilities operate on 100% renewable wind energy. More than 90% of the waste from materials used for Air soles is repurposed into new, innovative cushioning systems.
- Synthetic leather is produced by binding at least 50% recycled leather fibers with synthetic fibers using a water-powered process.
- Polyester fabrics are made from recycled plastic bottles.
- Cotton in Nike products is either certified organic, recycled, or sourced through the Better Cotton Initiative (BCI).
- Nylon is transformed from various materials, such as used carpets and fishing nets. The newly recycled nylon yarn cuts carbon emissions by up to 50% compared to virgin nylon.

Since we decided to go with Nike as a global company influencing the fashion industry worldwide, we needed to check if they were upholding their sustainability standard by double-checking whether the materials used were sustainable. Using the textile exchange's website, we could check if the materials used were certified, “filling the gaps in existing guidelines” on sustainability [5]. The only factory certified in sustainable materials is Nike's factory in Italy, which is responsible for 2% of all finished goods produced by Nike [6]. Nike's factories in Italy are certified by the GRS (Global Recycled Standard), RAS (Global Alpaca Standard), RMS (Responsible Mohair Standard), and RWS (Responsible Wool Standard) standards [5]. The certification of recycled and knit materials speaks to Nike's trademark Flyknit shoes, one of the company's most prevalent sustainable material styles. However, there is still much more growth that needs to occur in the certification of Nike’s factories, since 98% of products are produced without recycled standard and animal welfare certifications. Additionally, our analysis considers other global sustainability measures, such as those from The Sustainable Apparel Coalition. This coalition utilizes the HIGG index, a comprehensive tool for measuring the environmental and social impacts of products throughout their lifecycle, which was initiated by Nike in 2012 [7]. While the HIGG index is instrumental in highlighting Nike’s strides towards reducing its environmental footprint and promoting sustainability, it also raises concerns about potential biases, given that it is a standard developed by the company itself.

Our project aims to correlate sentiment and the environmental impact of shoes to determine whether customer favored products are better for the environment. By doing so, we anticipate providing insights that will help companies, including Nike, become more conscious of the materials they use and their adherence to social and environmental standards. This, in turn, will enhance transparency and accountability in buyer-supplier relationships within the industry.

**Previous approaches**

Previous studies have focused on the life cycle assessments of shoe brands, such as the study by Milà et al. (1998) in The International Journal of Life Cycle on the leather footwear industry [8]. However, the previous study had limitations on only assessing leather footwear produced in Spain and did not include imports or exports related to the product's manufacturing. This means that a large proportion of the emission and waste from a shoe is not getting categorized. Additionally, a significant issue in the collection of life cycle assessments is that it is hard to trace where every material was produced. This leads to a severe issue regarding the traceability of a product's life cycle. According to a Business Insider report, "approximately 99% of all shoes sold in America are made elsewhere due to the labor-intensive process that shoe-making requires and the high cost of labor in the US [9]." This is also a significant issue for companies such as Nike, which sells almost 30 pairs of shoes a second based on their quarterly revenue. According to Good On You, a fashion sustainability rating company, "the relentless churn and impossible turnaround times have made it common for garment factories to outsource work to other factories (or countless informal homeworkers) while brands turn a blind eye [10]." Constant production and adherence to trends make it hard for life cycle assessments to compile accurate details about a brand's sustainability goals.

Furthermore, customer preferences are the driving force behind brand sustainability policies. It is crucial to focus on how consumers perceive and value sustainable practices and products. If there were no customer demand for sustainability, companies would have little incentive to implement these practices. Understanding customer sentiment is important for making a meaningful impact on how companies can better tailor their products towards better the environment while also pleasing consumers. Suresh et al. Mining Effective Strategies for Climate Change Communication is an essay with the purpose of understanding which tweets; the goal was to understand which tweets most effectively engage users [11]. Using a similar methodology, we analyzed whether customer reviews aligned with product descriptions regarding sustainable and non-sustainable shoes. This approach involved a sentiment analysis of customer reviews, comparing the language used in sustainable and non-sustainable shoe reviews to identify key themes and sentiments. It allowed us to understand if product descriptions were accurate and gauge consumer perceptions and satisfaction. Such insights are crucial for understanding the impact of marketing on consumer behavior and ensuring that the sustainability claims by shoe brands are accurate and resonate with consumers.

**Methods**

The original dataset we intended to use contained details on Nike and Adidas products. Unfortunately, it was outdated and didn't match any of the product IDs of the current sustainable shoes from that file. We also decided to focus solely on Nike due to Adidas' privacy policies and web scraped the product information from the current Nike website to create our own dataset.

Using the BeautifulSoup package, we obtained the sustainable shoe product IDs, full and current prices, whether it's discounted, product URLs, descriptions, average rating scores, and total review counts and stored it into one csv file. The separate csv file contained the first 3 default reviews of all sustainable shoes. We repeated this process with all Nike shoes and removed duplicate product IDs in order to differentiate between Nike's sustainable and non-sustainable shoes. Nike had a category for shoes made with sustainable materials but not for 'non-sustainable' materials.

We preprocessed the data to ensure that the raw text from product descriptions was suitable for further analysis. This began with text normalization, where all text was converted to lowercase to help standardize the data. Next, we performed tokenization, removing common stop words such as "the", "and", and "or" using the NLTK library. This helps in focusing on more meaningful words. We then tagged each word with its part of speech using NLTK's tagging function and words were then converted to their base form. Finally, the frequency of each word was determined to help identify the most common words in the data and provide insights into common topics.

Statistical tests were performed to provide a better understanding of how sustainability impacts consumer engagement and satisfaction. We utilized a t-test to determine if there were significant differences in the average product ratings for sustainable and non-sustainable shoes.The goal of this test is to help determine if consumers rated one category of shoes higher than the other, which could indicate a preference based on sustainability. To analyze consumer engagement, we performed a chi-square test, which is utilized for categorical data, on the distribution of the number of reviews between the two categories. By applying the chi-square test to the review counts of the two categories, we could determine whether there was a significant difference in how often consumers engaged with sustainable versus non-sustainable products.

BERTopic was used to cluster text data derived from product descriptions and customer reviews. BERTopic utilizes transformers to vectorize text and employs UMAP for dimensionality reduction, which makes the text more interpretable. HDBSCAN is then implemented to cluster the vectorized words, and predominant keywords within each cluster are identified. We identified distinct clusters representing aspects of customer reviews and product descriptions, including comfort, quality, and aesthetics.

**Results**

Figures 1a and 1b. the relationship between ratings and the number of reviews for both sustainable and non-sustainable groups. The graphs show similar trends for both groups. Also, there are more ratings between 4-5 for both groups, which suggest high satisfaction rates for both groups. From this, we could infer that the material doesn't affect the satisfaction rate significantly. Figure 2a and 2b show the relationship between full price and reviews for both

groups as well. Two groups have similar trends as well, from which we could infer that price didn't affect consumer demand that much. To accurately support this, we conducted statistical tests - t test and chi square test. The resulting p-values were 0.698 for the t-test assessing average ratings and 0.608 for the chi-square test analyzing the distribution of the number of reviews. Both values were way above the threshold of 0.05, proving the insignificance in difference. These results help validate that being labeled as 'made with sustainable material' itself does not impact consumer experience, showcasing that including more sustainable products will not result in lower customer satisfaction.

Initially, we thought that the BERTopic model came pre-trained on text data, which was why we utilized it to test our data. Figures 3a and Figure 3b were originally shown for the presentation, illustrating that four topic clusters were found in the descriptions for both types of shoes.

In Figure 3a the identified topics for sustainable shoe descriptions were as follows: Topic 0 - Comfort and recycled materials, Topic 1 - Performance sport shoes, Topic 2 - Advanced innovation, and Topic 3 - Miscellaneous. In Figure 3b, topics within non-sustainable shoe descriptions were categorized as: Topic 0 - Classic, timeless styles, Topic 1 - Sport shoes, Topic 2 - Comfort, and Topic 3 - Specific sport shoes.

However, upon further investigation, we realized that the BERTopic model needed to be pre-trained on our specific dataset. Instead, these initial clusters represented patterns within the training data used to train the model.

Similarly we made the same oversight for the shoe reviews, shown in Figure 4a and Figure 4b Sustainable shoe reviews were overall positive. The topics were named as: Topic 0 - Comfort, Preference, Topic 1 - Quality. For non-sustainable shoe reviews, it showed four clusters, named as the following: Topic 0 - Comfort, Preference, Topic 1 - Quality, Aesthetics, Topic 2 - Product Concerns, Topic 3 - Negativity.

Therefore, after properly training and testing a new model, the results did not significantly deviate much from our initial findings. Figure 5a illustrates the clusters found in the sustainable shoe description test set, still identifying Topic 0 - Comfort and recycled materials, Topic 1 - Performance sport shoes, and Topic 2 - Advanced innovation. The only difference was the absence of a miscellaneous cluster. Figure 5b shows the clusters found in the non-sustainable shoes description test set with clusters labeled as Topic 0 - Sport shoes, Topic 1 - Comfort, Topic 2 - Classic, timeless shoes, and Topic 3 - Innovative.

These test results show that both shoes still had similar topics, which suggests that despite the difference in materials used to create these two types of shoes, they both prioritize comfort in sportswear.

Figure 6a represents the clusters for the sustainable shoe reviews in the test set, with topics such as Topic 0 - Comfort, Love, and Topic 1 - Positivity. However, Figure 6b shows topic clusters labeled as Topic 0 - Error, Topic 1 - Negative, and Topic 2 - Positivity. These two figures show that both types of shoes had positive reviews, but the model found negativity as a main cluster. Additionally, non-sustainable shoes' Topic 0 may indicate there were some errors when trying to cluster the test set. As a result, this reveals that customer sentiment is related to concerns over comfort and preference regardless of the sustainability of the shoe and quality.

**Conclusion**

Our findings, as depicted in the figures, underscore the fact that there's a very small difference between sustainable and non-sustainable shoes in terms of product descriptions and consumer sentiment, so the material used does not have a significant impact on consumer demand. This supports our claim that customer satisfaction for both types of shoes is primarily driven by comfort, preference, and quality, which we have driven from our machine learning. However, there is a noticeable divergence in aesthetics and sustainability preferences. Future research could involve consumer insights studies to better understand specific elements of design that consumers feel are lacking in sustainable products.

Also, the strong alignment between product descriptions and customer reviews demonstrates that Nike's product descriptions accurately capture customer sentiment. Hence, their precision in creating shoes to meet their intended purpose suggests Nike could fulfill specific needs and functions when making new products, or renewing their popular products with sustainable materials. Furthermore, they could even increase investment in research and development to improve the quality and performance of these sustainable materials so that they match or exceed the qualities of traditional, less sustainable materials.

A limitation of our study is the available data for the project, which may not fully capture all consumer opinions, especially those who did not review the product, but still purchased. Moreover, our analysis could benefit from a deeper examination into the long-term sustainability impacts of different materials, which was beyond the scope of this project.

This project highlights the need for global changes in the traceability of shoe production. Consumers need more information beyond the vague labels of "sustainable" or "non-sustainable." Instead, consumers should have access to who made their shoes, where their shoes were manufactured, and what is in their shoes. Unfortunately, many global companies, like Nike, fail to provide answers to these questions for a significant number of their products.

**Graphs & Outputs**

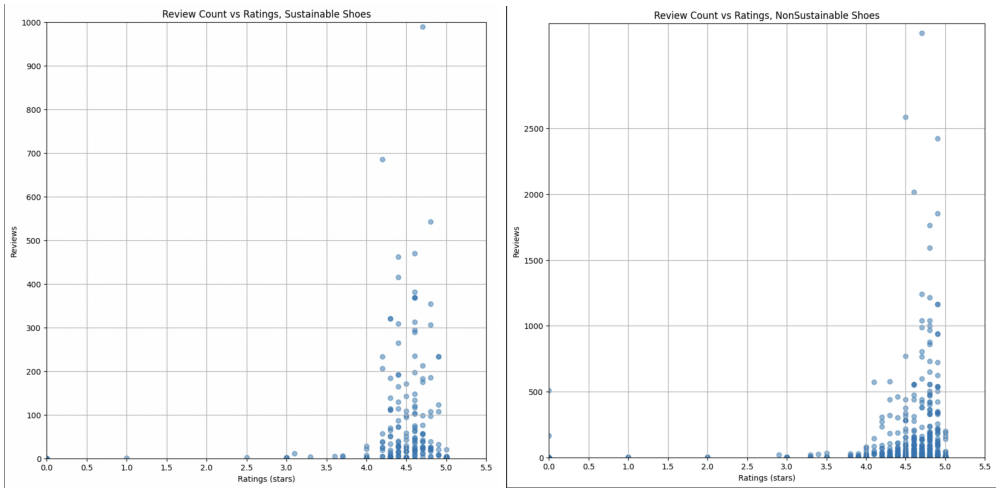


**Figure 1a, 1b**: Relationship between ratings and reviews on both types of shoes

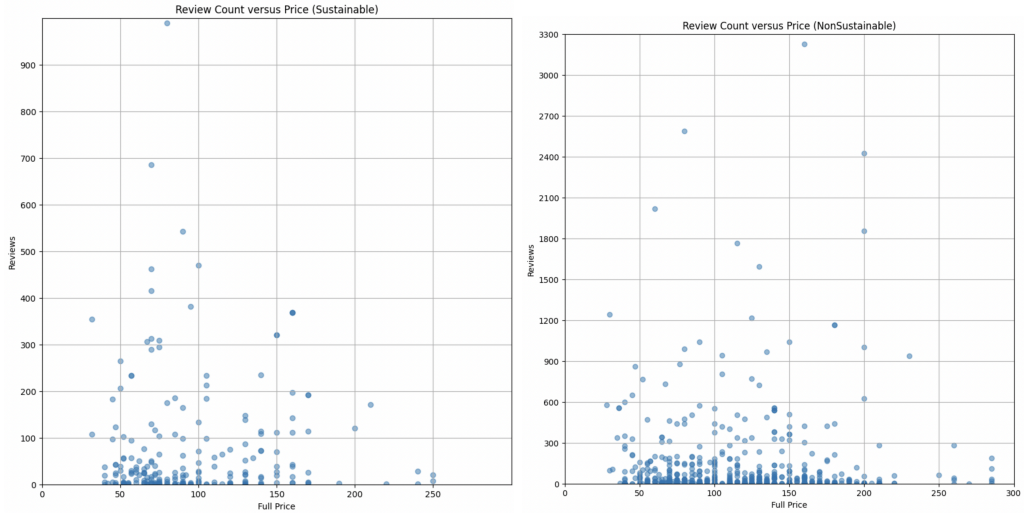


**Figure 2a, 2b:** Relationship between full price and reviews

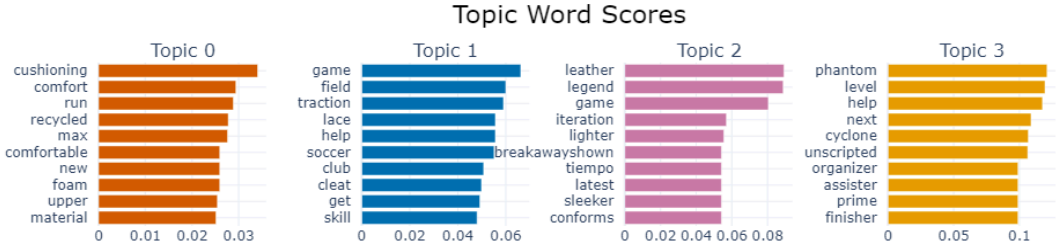


**Figure 3a:** Topic clusters on sustainable shoes descriptions (Initial)

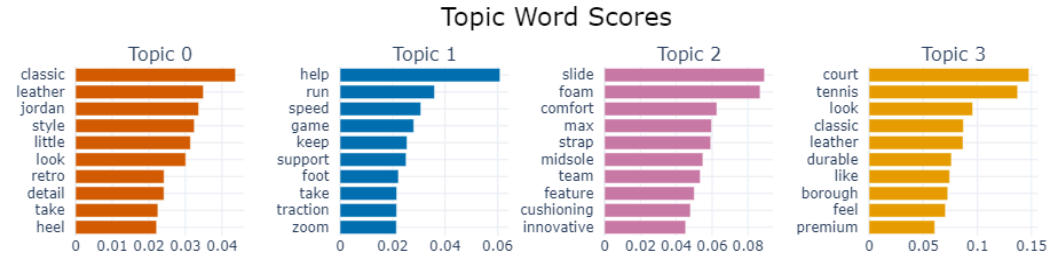


**Figure 3b:** Topic clusters on non-sustainable shoes descriptions (Initial)

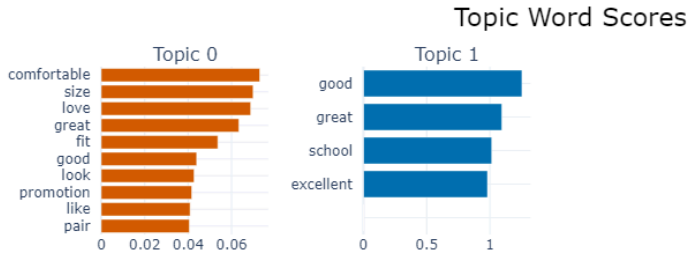


**Figure 4a:** Topic clusters on the sustainable shoes reviews (Initial)

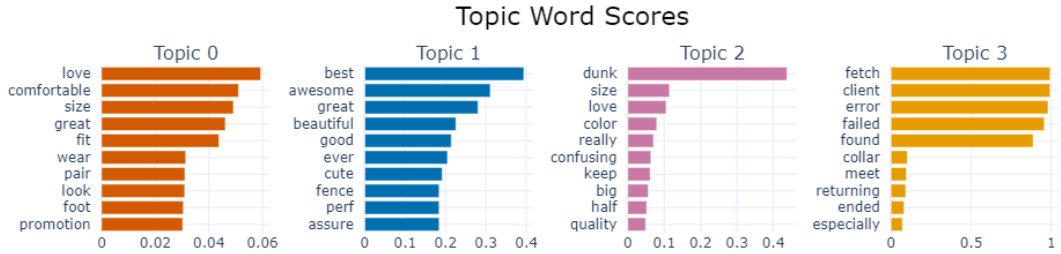


**Figure 4b:** Topic clusters on the non-sustainable shoes reviews (Initial)

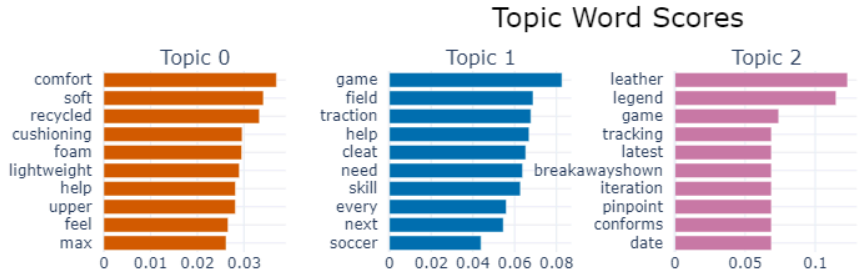


**Figure 5a:** Topic clusters on sustainable shoes descriptions test set

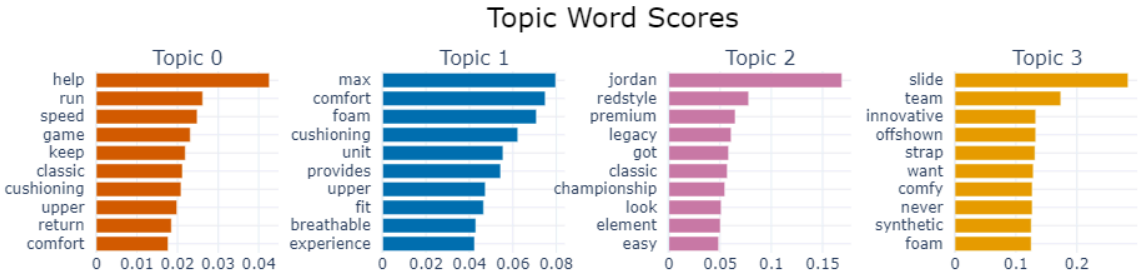


**Figure 5b:** Topic clusters on non sustainable shoes descriptions test set

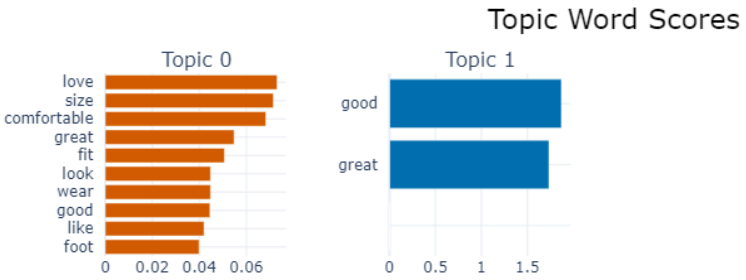


**Figure 6a:** Topic clusters on sustainable shoes reviews test set

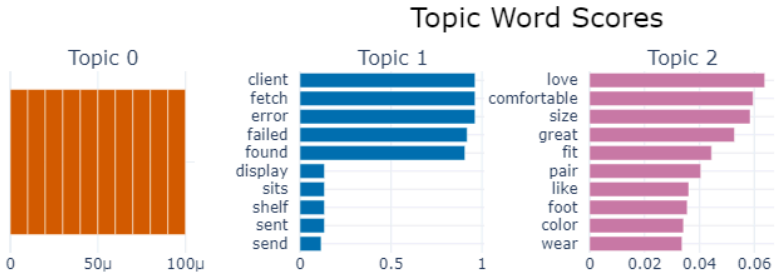


**Figure 6b:** Topic clusters on non sustainable shoes reviews test set

## References

1. "The Impact of Textile Production and Waste on the Environment." *European Parliament*, 2024, www.europarl.europa.eu/topics/en/article/20201208STO93327/the-impact-of-textile-production-and-waste-on-the-environment-infographics.
2. "Q1 Earnings Report." *Nike, Inc.*, 2024, investors.nike.com/investors/news-events-and-reports/default.aspx.
3. "Find Certified Company." *Textile Exchange*, www.textileexchange.org/find-certified-company/.

4. Nike. “Sustainability: Materials.” *Nike*, www.nike.com/sustainability/materials.
5. “Standards.” *Textile Exchange*, www.textileexchange.org/standards/.
6. "Manufacturing Map." *Nike Manufacturing Map*, www.manufacturingmap.nikeinc.com/.
7. "The HIGG Index." *Sustainable Apparel Coalition*, www.apparelcoalition.org/the-higg-index/.
8. Milà, L., et al. "Life Cycle Assessment of the Leather Footwear Industry." *The International Journal of Life Cycle Assessment*, 1998.
9. Bhattarai, Abha. "The Future of Shoe Manufacturing in America." *Business Insider*, 2017, www.businessinsider.com/the-future-of-shoe-manufacturing-in-america-2017-3.
10. “Fashion Sustainability Ratings” *Good On You*, www.goodonyou.eco/.
11. Suresh, A., Milikic, L., Murray, F., Zhu, Y., & Grossglauser, M. (2023). Mining effective strategies for climate change communication. In *Proceedings of the Tackling Climate Change with Machine Learning Workshop at ICLR 2023*. EPFL.