ECO-Search
Project Final Report

Matteo Di Mario, Meera Kumar, Debbie Shih, Sai Sure, Faye Xiao

1 Abstract

This project introduces an AI-powered search engine designed to help consumers find sustainable clothing options by prioritizing verified eco-friendly attributes like ethical labor, recyclability, and carbon footprint. Traditional search engines often fail to highlight genuinely sustainable products due to greenwashing and an overemphasis on popularity and relevance.

By fine-tuning BERT with domain-specific sustainability keywords and integrating semantic analysis, the engine filters misleading claims and ranks results based on credible sustainability certifications and transparency. The solution addresses a growing consumer demand for ethical shopping by aggregating data from trusted sources and employing advanced information retrieval and ML techniques. Initial testing shows promising results. Challenges include dataset scalability and manual annotation, but the framework lays a foundation for future improvements. We hope this tool can empower users to make informed, sustainable choices more effortlessly, offering a niche alternative to conventional search engines for those interested in sustainability.

2 Project Description

2.1 Social Problem

The most popular search engines today (including Google, Bing, etc.) tend to prioritize information to satisfy popularity and relevance to the query in order make users' search as smooth as possible. While this makes sense from a business perspective for the companies that operate the search engines, this leaves a gap for people who would like to have a similar experience in the search process but also have other parameters be considered in the search. This specifically applies to those interested in sustainable products.

In fact, studies show that users have increasingly begun to prioritize factors related to their environmental footprint. For example, 50% of consumers identified sustainability as "one of their top criteria" when purchasing goods, according to a survey (Albella et al., 2022) of 23,000 consumers conducted by Bain & Company. Sustainability is often quoted as one of the priorities for consumers during purchases and shopping, but this cannot always turn into action if they are bombarded with unsustainable options during their searches.

For these consumers who want to make environmentally and ethically conscious choices, traditional engines are not always able to meet the demand. Sustainability-related attributes such as carbon emissions, ethical labor practices, recyclability, and supply chain transparency are rarely integrated into the core ranking systems of mainstream search tools. As a result, products that align closely with users' values may be hidden beneath more popular or heavily marketed alternatives, especially given the ad-driven business model of modern search engines. This makes discovering truly sustainable options time-consuming and frustrating, especially for those without the expertise or time to investigate each product. To bridge this gap, there is a growing need for alternative search experiences that elevate transparency and sustainability alongside traditional relevance and convenience.

Often, product listings on e-commerce platforms are filled with "greenwashing" claims: statements that suggest a product is eco-friendly when in reality, its production, sourcing, or distribution may not meet any meaningful environmental or ethical standards. Consumers are overwhelmed with conflicting information and unreliable sustainability metrics, making it difficult to make informed choices. This disconnect between consumer demand for transparency and the inability to reliably assess sustainability attributes is where the problem lies. According to an HBR article (White

et al., 2019) "65% [of people] said they want to buy purpose-driven brands that advocate sustainability, yet only about 26% actually do so", which is a problem that needs to be addressed.

With these practices making it difficult for consumers to find what they are looking for online, a need for an alternative exists. Our project does not intend to compete with currently-popular methods for general search, but more so present itself as an add-on for users to adopt if they want their search to be void of greenwashing and more specifically oriented towards sustainable shopping.

Although we started with a broad scope, we observed that one of the areas where consumers tend to struggle most is in clothing. As such, we tailored our focus to a search engine that targets clothing brands and that can handle searches related to clothing products. In this way, we return results that align user interest with sustainability practices. According to a survey conducted in Australia, "60% [of people] say companies should be transparent about their practices, while almost half said they are willing to pay more for clothes that are sourced ethically".

2.2 Technical Solution

To address this problem, we propose the development of an AI-powered search engine specifically designed to rank and recommend products based on verified sustainability attributes. This search engine is built around a comprehensive and data-driven approach that evaluates and incorporates a wide range of eco-friendly factors, including fair labor; materials sources; packaging and shipping practices; and general company sustainability measure metrics. This solution tries to tackle this greenwashing search problem for specifically for problem, hoping it can be a useful proof of concept for other sustainability areas as well.

The solution proposed here is an AI-powered search engine that indexes and ranks clothing products and brands based on their verified sustainability data. This search engine aggregates data from reputable sources that include sustainability information, and integrates it into our product ranking algorithm. The engine uses relevant keywords to train our model, fine-tuning it to provide tailored search results that prioritize sustainability practices. Providing users with clear and actionable insights about the environmental and social impact of each option, our solution facilitates to a more tailored shopping experience.

3 Related Work

3.1 Sustainability Challenges in Search

According to a McKinsey & Company report (Yang et al., 2021), greenwashing practices have become increasingly common. As the report shows, these practices are particularly evident in the context of fashion. Even though consumers generally want to "go out of their way to to buy secondhand items and to look for clothing made with environmentally friendly material", a key problem in the fashion sector is that consumers often find it challenging to understand what "sustainability" truly means. This confusion is exacerbated by an overwhelming amount of unclear information provided by brands on their websites and product tags, ultimately resulting in consumer difficulty in distinguishing between genuine sustainability efforts and greenwashing tactics. An unnecessarily complex regulatory environment also contributes to a sense of confusion even for the motivated consumer (Santos-Roldán et al., 2020). Ultimately, greenwashing undermines consumer trust in genuinely sustainable brands and can reinforce mistrust towards companies that are making real environmental efforts. By simplifying and targeting the search experience for users, our solution can help combat this.

In order for our model to recognize greenwashing and highlight it in our search engine, we sought to determine the mechanisms used by companies to propagate greenwashing claims. According to a study (Mazur-Wierzbicka, 2023) companies tend to utilize the following mechanisms:

- *Vague or ambiguous terms.* Companies may use ill-defined terms like "natural," "ecological," "clean," or "green" without providing specific details or evidence to support these claims. This lack of clarity makes it difficult for consumers to understand the actual environmental impact for a specific item.

- *Lack of evidence.* Claims about a product's eco-friendly attributes may be made without providing any verifiable proof or certification. Consumers often lack the resources to independently assess the validity of such claims.

- *Hidden trade-offs.* Companies may highlight a single environmental benefit of a product while ignoring or downplaying other significant negative environmental impacts. This

selective disclosure creates a skewed perception of the product's overall sustainability.

- *Irrelevant claims.* Some companies promote environmentally friendly aspects of their business that are actually required by law or common practice within the industry. This serves to divert attention from more substantial environmental issues. These "insignificant terms" can mislead consumers into believing a company is going above and beyond.
- *False labels and certifications.* Companies may use fake or misleading eco-labels to give the impression of environmental legitimacy. Consumers may not always be able to distinguish between authentic and counterfeit certifications.
- *Misleading visuals and imagery.* The use of green imagery or natural-looking packaging can create a false sense of environmental friendliness, even if the product or the company's practices are not genuinely sustainable.
- *Promoting the "lesser of two evils".* This involves marketing a product as environmentally friendly simply because it is slightly better than a more harmful alternative, without addressing the product's own negative impacts.

Similarly, we explored research on eco-labels to identify which keywords to utilize in creating the model, in order to understand how best to weigh each label. Authors of a recent paper (Ziyeh and Cinelli, 2023) propose a framework for analyzing eco-labels in the clothing industry. They identify two main types of label assignments and weight these types to develop a stronger means of analyzing sustainability performance of a specific product. The two label assignment types are described below:

- *Binary label assignment.* The eco-label is either assigned or not assigned to a product. This includes a list of mandatory criteria that all need to be fulfilled. Examples include EU Ecolabel, Blue Angel, Nordic Swan, and Green Button.
- *Label assignment on different levels.* Products are sorted into preference-ordered levels based on their overall performance on multiple criteria. Examples of such criteria include GOTS (Global Organic Textile Standard) and Bluesign.

3.2 The Technical Approach to Resolve Search Challenges

Our model incorporates 1) domain-knowledge fine-tuning for clothing sustainable labels and 2) avoidance or rejection of empty, non-backed sustainability claims in order to effectively tackle greenwashing. Similar algorithms described in other papers are detailed below.

A 2021 paper (Confetto and Covucci, 2021) establishes an algorithm that collects specific semantic signals from a website's HTML code in order to improve their search. It works off of a database of sustainability terms and phrases, organized according to a taxonomy of sustainability themes (i.e. Planet, People, Profit, and Governance). This controlled vocabulary was developed from existing dictionaries and encyclopedias on sustainability and corporate social responsibility. The algorithm analyzes the presence and context of these "theme-words" in various parts of a web page, including visible content and metadata. It additionally uses Latent Semantic Indexing (LSI) techniques by identifying semantically related terms to strengthen the content's semantic context.

To implement the AI portion of the model we instead decided to utilize the latest BERT model (Devlin, 2018). As a result of the relative success of this algorithm we tried to take some of the successful developments introduced in the paper, integrating it with our AI-based approach, which we think can significantly improve user experience.

For the data-crawling portion of this project, we utilized the Aho-Corasick algorithm to improve the initial website dataset (Aho and Ullman, 1975).

We additionally sought to explore more consumer-oriented research. For example, Ecosia is a fully functional search engine that focuses on sustainability. However, this engine has a broader scope and is not exclusively focused on fashion. Also, it emphasizes the environmental impact of the AI used in its search engine more so than the sustainability of its results. This product in particular helped us identify future steps to be taken to improve the model when developing the front-end.

4 Data Collection

The data collection was mainly subdivided in three parts:

- Collecting the relevant websites that contain clothing brands and products.
- Collecting the labels and keywords to train the model on.
- Preparing some data annotations and a test dataset used to measure the model performance.

4.1 Data Collection for Relevant Websites of Clothing Brands and Products

We initially identified websites that collect brands and products for clothing brands. This portion of data collection was mostly done manually, given the domain expertise required for this phase. Through this, we identified a starting list of aggregating websites that provide relevant product clothing information. We did not focus on sustainability measures at this stage but rather the inclusion of as many clothing brands and products as possible, since the goal of this stage was only to acquire as much information as we could to feed into the model.

After that, we identified websites with brands and products known for their sustainability practices so that we would have a benchmark to compare against when looking for keywords for the model. The list for the benchmark included the following websites:

- *SPOT.* This is a comprehensive product database that lists sustainable items across industries, based on verified environmental certifications. Useful for indexing sustainability attributes.
- *Good On You.* Though primarily a rating system, this provides sustainability scores for fashion brands that could be integrated into an extension.
- *Project Cece.* This aggregates ethical fashion brands and allows users to filter based on sustainability criteria.
- *ENERGY STAR.* This is a symbol representing the satisfaction of strict standards set by the EPA.

We also manually curated a list of 20 seed websites that would be used for the web crawler. Using Good on You to verify their eco-friendliness, within the 20, we decided on 10 eco-friendly e-commerce sites and 10 non eco-friendly, fast fashion e-commerce sites. Some of these include:

- *Fair indigo.* This is a sustainable fashion brand committed to ethical production and fair wages for workers. The company emphasizes transparency, using organic cotton, recycled materials, and responsible manufacturing processes.
- *Triarchy.* This is a denim brand known for its water-saving production techniques and sustainable materials. The company focuses on reducing the environmental impact of denim manufacturing.
- *Zaful.* This is a fast fashion retailer that specializes in trendy, low cost apparel for women. The company lacks transparency in its manufacturing practices and fails to provide tangible evidence of any sustainable practices.

By utilizing websites that comprise a wide range of sustainability practices, we created a robust dataset to be crawled.

4.2 Data Collection for Labels and Keywords

Once the list of seed URLs were established, we moved to curating the keyword lists. We needed two: one list for positive keywords that indicated sustainability and another that indicated unsustainable practices. For the positive keywords, it was mixture of real sustainability certifications and terms relevant to sustainability. Some examples of positive sustainability-related keywords include certifications like CPSIA (US Consumer Product Safety Improvement Act), REACh (a European chemical regulation), Oeko-tex, and GOTS (Global, Organic Textile Standard). Other positive keywords include "biodegradable", "locally-sourced", "regenerative farming". Negative keywords, such as "flash sale" and "going fast", highlight practices relating to fast fashion and overall a lack of sustainability. For more examples of keywords collected, refer to the appendix.

Next, we built a web-crawler to scrape the dataset of websites identified. The crawler would find relevant keywords in the websites, determining relevancy from our keywords dataset. Starting from a set of seed URLs (ie. the websites dataset), it follows each page's robots.txt file to recursively explore links while performing keyword matching. The crawler extracts text content from each web page using BeautifulSoup and checks for the presence of keywords relevant to our context, as defined in our keywords dataset. The results are then

stored in output files. We decided to implement both exact and fuzzy keyword matching. This is because, given the nature of our keyword lists, some of the keywords need to be matched up exactly (e.g the sustainability certifications) while some can be more flexible. To account for this, we used the aho-corasick algorithm to identify words that were an exact match, and fuzzy-wuzzy to identify words that are semantically similar to the keywords.

The final output of the crawler is two files, one with positive output and one negative. Each file contains every URL crawled or parsed with a corresponding list of all keywords found on that site. The keyword list for each website is of fixed size to account for all relevant keywords in our original keywords dataset, with empty strings corresponding to keywords in our dataset that were not found on the specific website. One output file holds a list of the positive keywords found through the crawl for all websites crawled, and the other contains the negative keywords. Positive here refers to eco-friendly, while negative refers to keywords that indicate a lack of sustainable practices for a given product or company.

We do acknowledge, however, that there are certain limitations to our approach, detailed below.

First, relying on keywords could potentially under-represent sustainability-focused companies that may not be using as many keywords as other companies. That being said, we think that such companies represent a significant minority, since companies with sustainable practices often have the most up-to-date labels and standards, highlighting them on their websites.

Second, relevant keywords are likely to change over time due to trends, for example, and the performance of the model might deteriorate if the list is not updated. In the future, we plan to explore different ways to update our keyword list dynamically such that our solution stays relevant.

4.3 Preparing the Data for Performance Measurement

Some final processing was done to make sure that the dataset could be easily fed into our MiniLM model. In order to test the model, we compiled a dataset of queries and relevant websites. These queries and relevant sites were compiled through study of most popular searches made by users looking for clothing shopping options.

4.4 Challenges with Data Collection

In the data collection phase, we faced two main challenges: the size of the training dataset, and the variability of the websites we crawled.

First, given the large number of websites, brands, and products in the fashion space, we had to make decisions as to which websites to include in our scope. In our first round of data collection, we focused on compiling a list of websites that comprised a variety of sustainability practices, as described in Section 4.1. In the second round of data collection, we increased the dataset significantly and took a more systematic approach about how to include and exclude websites in the training dataset as we kept improving our model.

Second, not every website is the same, in terms of size and layout. This initially created issues in our web crawling, as our algorithm would end up spending unnecessarily long amounts of time on certain websites whose layouts were not as intuitive, inadvertently spending time on irrelevant parts of the websites. Part of this problem was resolved by utilizing a breadth-first search approach.

5 Method description

The approach of our model consisted of the following phases: 1. Preprocess the data collected to make it compatible with the MiniLM model (Devlin, 2018) 2. Utilize the very targeted and specific datasets obtained through data collection to feed them to the infrastructure of the MiniLM model and fine-tune the model 3. Adjust the embeddings 4. Extract the embeddings 5. Rank through Cosine Similarity integration. 6. Build a front-end and connect to AI-engine results.

5.1 Preprocessing

The preprocessing stage involved creating a dataset where the query would be compared against to get the websites that closely match the query. The initial data files included the websites that linked to clothing items along with positive and negative keywords associated with them. We started to create a Pandas dataframe with eco-friendly websites with their corresponding positive keywords and non eco-friendly websites with their corresponding negative keywords. Then, we utilized BeautifulSoup to extract paragraph text from each website and insert them into a new column of the dataframe. This dataset was used and encoded by the fine-tuned model for evaluation and usage.

5.2 Fine-Tuning all-MiniLM model

We utilized all-MiniLM-L6-v2 model from Sentence-Transformers because it is a compact and efficient variant of the BERT model. It is optimized to create high-quality sentence embeddings because it is trained to capture deep semantic similarities. The model maps sentences to a 384-dimensional dense vector space, making it ideal for semantic search. It also goes beyond keyword matching to retrieve results. Additionally since it is compact and lightweight, it allows for fast and scalable inference for similarity-based retrieval.

To fine-tune the model to favor eco-friendly websites, we had to create another dataset. The dataset was created using sample list of clothing items like sweater, jeans, and many others. It contained relevant eco-friendly websites and non eco-friendly websites along with their extracted text and keywords for each clothing item. The label was 1 if the website was eco-friendly and 0 otherwise. Additionally, we used a special token [ECO] to signal eco-friendly and enrich the website text. The model was fine-tuned on cosine similarity loss (Espejel, 2022). Since the dataset only contained 1749 rows, we ran the model for 1 epoch to avoid overfitting the dataset.

5.3 Embedding-Based Retrieval

After fine-tuning, the MiniLM is used to generate high-dimensional embeddings for the extracted text concatenated with keywords for each website. The query is also embedded using the fine-tuned model. Since we are promoting sustainability, we decided to use CPU to run every time a user provided a query since GPU creates a bigger carbon footprint comparatively. As a result, all the embeddings are moved to CPU when computing cosine similarity. The similarity between the query and each website is computed using cosine similarity using the function from the sentence transformers library. The results are ranked in descending order based on similarity scores and return the websites associated with the five highest similarity scores.

5.4 Front-end

We built a front-end for users to utilize the search engine. The URL is currently hosted locally. To build the front-end we used a combination of traditional search engine design like Google and more sustainability-focused ones like Ecosia, mentioned in the Related Work section. Once the main layout of the front-end was built we connected the front-end to the back-end. In particular, each search sends requests to the MiniLM model that, already fine-tuned sends responses in the form of a search engine result, providing relevant website to access.

6 Experiments and Results

The search engine's effectiveness was assessed by testing it with various queries to evaluate its ability to retrieve semantically relevant documents. Queries were transformed into embeddings using the fine-tuned MiniLM model, and cosine similarity was used to rank document embeddings. Ahead of time, 20 test queries were pre-prepared and relevant results were identified. This process required manually building a set of queries and relevant results and was for this reason not as extensive as hoped, and might have influenced the results one way or another. More extensive queries and associated relevant websites should be produced moving forward to to improve the retrieval quality. Originally, we had calculated certain metrics such as accuracy and precision to give an idea of the model performance, but we realized this was not very representative since it was extremely subjective and dependent on what was considered a environmentally friendly website or product. So we instead provide example results for the reader to evaluate on their own on the appendix. The results highlight a few things:

- The approach taken has yielded reasonable results on average given the constraints mentioned. We think that between expanded datasets and more extensive test query results, there is a good amount of room of improvement, starting from a good baseline.

- The user can expect on average a relevant and sustainable product/brand in the first five search results. This result highlights that our model can find itself a niche for interested consumers that do not want to spend a lot of time getting familiar with the clothing sustainability space, but are nonetheless careful about sustainability.

Finally, we wanted to note one more limitation of our approach on top of the other ones already raised throughout the paper. Given that we have used a pre-trained model it is possible that the model already had inherent biases, and it is possible that

these biases carry over to the results of our project as well.

7 Conclusions

While not as comprehensive as other search options might be, we hope that our solution can offer a small but useful alternative option to those that are encountering barriers in finding more sustainable option to their clothing shopping and searches. Given the initial ambitious goal of the project, we think we reached a satisfactory phase of utility of the the engine, while there remains of course a lot of work to be done to improve the model, but the current model as is can be preferred in the limited cases mentioned and targeted.

In particular, data collection and selection will be a significant focus of the next work, to extend the dataset and curate it more extensively, as this is necessary to allow the MiniLM to work more effectively. In addition, measuring performance has proved hard, because of the amount of manual work required. So we think better reference queries and results can be produced to check the real performance of the model.

8 Addressing Reviewer's Feedback

We carefully looked at the reviewer's feedback provided and attempted to implement correction accordingly as best as possible to the paper.

Below we report the feedback with a brief note for each on how we tackled it either in the paper or directly in the model implementation. The feedback is reported below listed by recurrence of comment's theme:

- Multiple comments were raised about the testing section, especially regarding more transparency of the evaluation metrics. We decided that calculating metrics on the performance of an engine on a subject that is not completely objective and has many nuances like sustainability. So instead we deemed more appropriate to provide direct results to queries (reported in the appendix) in order to let the user decide the performance of the model.

- There were some comments on the lack of figures/tables, so we ensured to add these to the final model. These are mostly provided in the final Appendix, since there was not enough space to include them while remaining within the 10 page limit. The figures include the final result of the front-end, examples of keywords utilized (comment raised elsewhere too), and others.

- It was noted to add sources in the social problem section to support the claim of consumers caring about environmental practices when shopping for clothes. Therefore we added a section at the end citing sources, to back our claim more strongly.

- Good comments on the limitations of using keywords to train the model were provided. These were later addressed partially both in the paper and in the implementation of the model itself, and those that could not be tackled were added as limitations in the data collection portion. More specifically, one very good comment highlighted how there was no mention of what greenwashing looks in practice, especially with regards to keywords. Since this was a big part of the project, and the model relied on being able to recognize between these two categories we added examples in the data collection section that could give a sense of this difference.

- One comment mentioned the paper was missing the abstract. We had originally not added it given that it was not mentioned in the section, but we decided to add it as a result of the reviewer's comment.

- There were some comments on lack of clarity regarding what was done manually and what was not. So we tried to address this throughout the paper to make it more clear to the reader.

- There was a comment about the lack of a user interface demonstration. Given this comment we decided to add a fully built front-end that connected to the already built back-end. Both the changes make by adding the front-end and some demonstrations of it were provided in the paper.

- Various comments were made about increasing the datasets. We increased the seed links from 3 to 20 (10 eco-friendly sites and 10 non eco-friendly sites). This resulted in 2000 entries in our dataset – 100 links per domain. Given this increase in size, we hope that we can generalize the model better, improve the

performance of the model, and provide more tailored responses to user's searches.

References

Elisa Albella, Anita Balchandani, Nic Cornbleet, and Libbi Lee. 2022. In search of fashion's sustainability seekers. *McKinsey & Company*.

Maria Giovanna Confetto and Claudia Covucci. 2021. "sustainability-contents seo": a semantic algorithm to improve the quality rating of sustainability web contents. *The TQM Journal*, 33(7):295–317.

J Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding/arxiv preprint. *arXiv preprint arXiv:1810.04805*.

Omar Espejel. 2022. Train and fine-tune sentence transformers models. *Hugging Face Blog*. https://huggingface.co/blog/how-to-train-sentence-transformers.

Ewa Mazur-Wierzbicka. 2023. Greenwashing–consumer's perspective. *Scientific Papers of Silesian University of Technology Organization and Management Series*, 164:283–297.

Luna Santos-Roldán, Beatriz Palacios-Florencio, and Juan Manuel Berbel-Pineda. 2020. The textile products labelling analysis and requirements. *Fashion and Textiles*, 7:1–24.

Katherine White, David J Hardisty, Rishad Habib, et al. 2019. The elusive green consumer. *Harvard Business Review*, 11(1):124–133.

Yang Yang, Shiwei Liu, Cunde Xiao, Cuiyang Feng, and Chenyu Li. 2021. Evaluating cryospheric water withdrawal and virtual water flows in tarim river basin of china: An input–output analysis. *Sustainability*, 13(14):7589.

Paula Ziyeh and Marco Cinelli. 2023. A framework to navigate eco-labels in the textile and clothing industry. *Sustainability*, 15(19):14170.

9 Appendix

Figures and Tables:

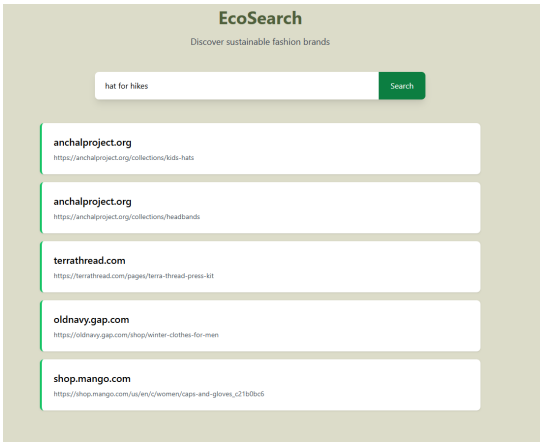

Figure 1: Search engine example search results

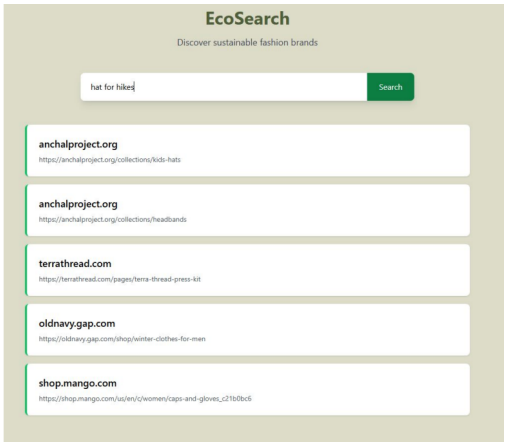

Figure 2: Search engine example test results

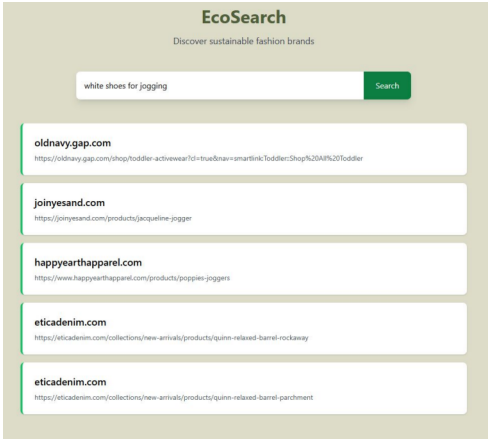

Figure 3: Search engine example test results

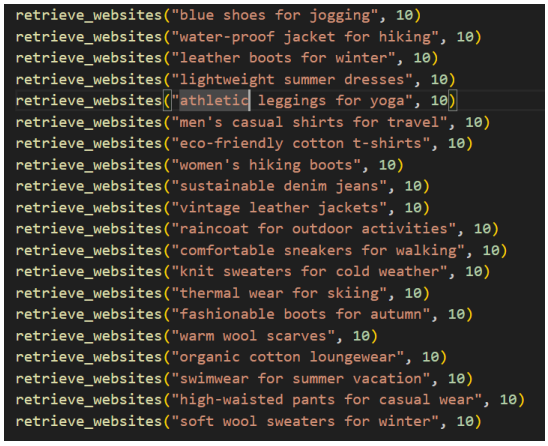

Figure 4: Test set for engine queries

Positive Keywords	Negative Keywords
WRAP certified	minutes ago
B corp	flash sale
PFC-free	ends in
bluesign	order within
UN Global Compact	last few

Figure 5: Example of positive keywords vs. negative keywords

10 Individual Contributions

Contributors:

- Matteo Di Mario: I developed the code to connect the front-end with the back-end. This includes the POST and GET requests to have the website display results and dynamically update according to the user's search. I also did the testing for the engine. Finally, I contributed to the checkpoint and the final paper, especially for the project description, related work, experiments and results.

- Meera Kumar: I developed the front-end for our solution, using HTML and React. As specified in the paper, it was designed after popular search engines like Google for increased accessibility. Additionally, I helped compile the seed URL list. I also led review efforts for the entire paper after we received peer review feedback, making sure all comments were addressed and that the paper read smoothly.

- Debbie Shih: I developed the web crawler designed to crawl ecommerce sites and identify sustainability related (and fast-fashion / non ecofriendly related) keywords on these pages. I did the research on what algorithms to use for keyword detection, and contributed to the manual curation of the seed URL list. I was also responsible for testing and making any tweaks to the crawler. I also contributed to editing and writing parts of the checkpoints.

- Sai Sure: I extracted paragraph text from all eco-friendly and non eco-friendly websites. We created a dataset with the website url, the extracted text, and the url's corresponding keywords. I fine-tuned MiniLM model using small dataset. I later encoded each website's text with the fine-tuned model and created a function to encode the query as well as return top 5 websites that are eco-friendly and relevant.

- Faye Xiao: I created the dataset used to train and evaluate the web crawler. This included identifying 20 seed ecommerce websites and manually collecting a total of 100 keywords by reviewing each site—5 relevant keywords such as "GOTS certified" and "carbon neutral," and 5 non-relevant keywords such as "trend-setting" and "going fast." I curated and formatted the seed URLs to ensure the crawler started from sustainability-related product pages. I also generated query-to-website mappings for terms like "shirt" and "jacket," labeling links as relevant or irrelevant to support a presence/absence keyword detection model.