Lexical Markers of Sustainable Fashion: A Text-Based Machine Learning Study

Farha Shaikh
BS Data Science and Applications
Indian Institute of Technology Madras

Abstract

This study investigates whether textual product descriptions contain identifiable lexical markers distinguishing sustainable fashion products from mainstream fast-fashion items. Using TF-IDF vectorization and Logistic Regression, we analyze lexical patterns across balanced datasets. While a single train-test split suggested strong performance, 5-fold cross-validation revealed performance variability (Mean F1 0.78, Std 0.16), indicating dataset-source bias and lexical clustering. Feature analysis demonstrates that sustainable products emphasize natural fiber terminology, whereas fast fashion focuses on demographic and aesthetic descriptors.

1 Introduction

The rapid expansion of fast fashion has significantly increased textile waste, water consumption, and carbon emissions. Sustainable fashion brands attempt to differentiate themselves through eco-conscious materials and responsible branding.

This study explores whether textual descriptors alone can distinguish sustainable fashion products from mainstream fast-fashion items. Rather than verifying sustainability claims directly, we analyze linguistic and lexical patterns embedded within product descriptions.

2 Dataset Construction

Two datasets were used:

- Curated sustainable Indian fashion products
- Large-scale general fashion dataset (44,000+ items)

Products were labeled as:

- 1 – Sustainable
- 0 – Non-sustainable

Due to significant class imbalance, majority downsampling was applied to create a balanced dataset for model training and evaluation.

3 Methodology

Text preprocessing included:

- Lowercasing

- Removal of non-alphabet characters
- Removal of explicit brand markers
- Removal of obvious sustainability tokens

TF-IDF vectorization (maximum 300 features) was used for feature extraction. A Logistic Regression classifier was trained for classification.

4 Evaluation

Model performance was evaluated using both a single train-test split and 5-fold cross-validation.

Train-Test Split Results

Class	Precision	Recall	F1-score
Sustainable (1)	0.99	0.99	0.99
Non-Sustainable (0)	0.99	0.99	0.99

5-Fold Cross-Validation

Metric	Value
Mean F1 Score	0.78
Standard Deviation	0.16

While the single split suggested strong separation, cross-validation revealed significant variability across folds. This indicates lexical clustering and dataset-source bias rather than uniformly stable predictive behavior.

5 Lexical Feature Analysis

Feature importance analysis revealed distinct lexical patterns.

Top predictors of sustainable products included:

- cotton
- linen
- hemp
- denim
- cupro

These terms primarily represent natural or semi-natural fibers commonly associated with sustainable branding.

In contrast, fast-fashion predictors included:

- men
- women
- black

- printed
- shoes
- watch

These terms reflect demographic segmentation, color-based marketing, and accessory-oriented product categories.

6 Limitations and Future Work

Several limitations were identified:

- Dataset-source bias due to curated sustainable brands
- Lexical clustering affecting cross-validation stability
- Sustainability labels derived from dataset origin rather than independent verification

Future work may include transformer-based language models (e.g., BERT), cross-domain validation, and multimodal integration of text and image features.

7 Conclusion

Text-based analysis reveals distinct lexical branding strategies between sustainable and fast-fashion products. However, robust evaluation is essential to prevent inflated performance claims. This study highlights the importance of cross-validation and bias detection in sustainability-focused AI research.