Vineetha Maddikunta
Social Media Mining, Fall 2024

Title

Analyzing Consumer Discourse on Reddit's Sustainable Fashion Community. (1480 words).

1 Introduction

The growth of social media has changed how people perceive sustainable fashion. Online communities have become important spaces where consumers share knowledge, discuss environmental concerns, and influence each other's fashion choices. Recent studies show that social media platforms significantly shape how sustainable fashion messages spread and impact consumer behavior (Orminski et al., 2021). Research has found that online communities are particularly effective at sharing knowledge about sustainable fashion practices and influencing supply chain decisions (Cervellon & Wernerfelt, 2012). While previous research has focused mainly on platforms like Twitter and Instagram, less attention has been paid to Reddit communities, where longer, more detailed discussions take place. This study examines how the r/SustainableFashion subreddit community discusses and influences sustainable fashion practices, building on earlier work that identified social media's role in shaping sustainable fashion discourse (Mukendi et al., 2020).

2 Research question

How do consumers in sustainable fashion communities navigate and discuss their consumption choices, particularly focusing on material preferences, ethical considerations, and purchasing behaviors?

3 Method

3.1 Data

For this study, the dataset is collected from the r/SustainableFashion subreddit, chosen for its focused community discussions on sustainable fashion and its position as one of Reddit's primary platforms for eco-conscious fashion discourse. The data collection was implemented using the PRAW (Python Reddit API Wrapper) library in Python, which provides authorized access to Reddit's API.

The data collection process involved searching for posts using three key search terms: "sustainable fashion", "eco-friendly fashion", and "green fashion". These keywords were selected as they represent the core terminology commonly used in sustainable fashion discussions and were implemented as case-insensitive searches to maximize coverage. To ensure a balanced representation of community perspectives, posts were retrieved using both "top" and "controversial" sorting methods, with time filters set to "all". This approach captured both highly upvoted content and contentious discussions, helping to avoid potential popularity bias. The script collected 1300 first-level comments from the retrieved posts, implementing several technical considerations: Comments were processed recursively to count all nested replies using a custom count_replies function. Rate limiting was managed through strategic sleep intervals (2 seconds between requests). The script replaced "MoreComments" objects to ensure complete thread collection.

Each comment's metadata includes the post title, comment body, score, creation timestamp, author, and reply count, offering insights into user engagement and discussion dynamics. To maintain data quality, the collection process excluded deleted comments and automated responses. The data was stored in a structured CSV format for subsequent analysis.

3.2 Analysis

The analysis of sustainable fashion discourse on Reddit employed a comprehensive methodological approach combining multiple advanced computational techniques. This section details the systematic process used to extract meaningful insights from the dataset.

1. Data Preprocessing and Cleaning: The initial phase focused on ensuring data quality and preparing the text for analysis. The dataset, comprising Reddit comments about sustainable fashion, underwent several preprocessing steps.To maintain data quality, filtering criteria was implemented to remove comments with fewer than 5 words, as these typically did not contain meaningful content for analysis. At this stage, the comments which are [delelted] or [removed] will also be filtered out. Duplicate

entries were eliminated to prevent potential skewing of results. These steps resulted in a total of 1146 comments.
Text preprocessing involved multiple stages to standardize the content. Comments were converted to lowercase, and special characters and numbers were removed using regular expressions. The text was then tokenized into individual words, and common English stop words were removed to focus on meaningful content. Finally, lemmatization was applied to reduce words to their base form, ensuring consistent treatment of variations of the same term.

2. Topic Modeling: Implemented Non-Negative Matrix Factorization (NMF) as our primary topic modeling technique. The choice of NMF was driven by its ability to identify interpretable topics and its superior performance compared to alternative methods like LDA and LSA, as demonstrated by coherence scores. First, a document-term matrix using TF-IDF vectorization is created, with parameters set to exclude terms appearing in more than 95% of documents (max_df=0.95) and terms appearing in fewer than two documents (min_df=2).
To determine the optimal number of topics, an extensive evaluation is conducted across different configurations, testing models with 3, 5, 6, 7, and 10 topics. Each model's performance was assessed using coherence scores, with the NMF 6-topic model achieving the highest score of 0.6437, indicating strong topic coherence and interpretability. The results from this model are added as a new column in the dataframe which will later be used for supervised machine learning method.

3. Unsupervised Learning - Sentiment Analysis: The sentiment analysis component utilized the Empath lexicon-based tool. This method allowed us to examine the emotional content of discussions across multiple dimensions simultaneously.The categories selected for sentiment analysis—positive emotion, negative emotion, money, shopping, and business—were carefully chosen for their relevance to the discourse surrounding sustainable fashion. Here's why each category was included:

Category	**Reason for Selection**
Positive Emotion	Captures general optimism and positive sentiment towards sustainable fashion.
Negative Emotion	Reflects frustration or concerns about sustainability in fashion.
Money	Addresses financial concerns, as consumers may perceive sustainable fashion as expensive.
Shopping	Focuses on consumer behavior and attitudes towards sustainable shopping habits.
Business	Reflects sentiment about businesses' role in promoting sustainability in fashion.

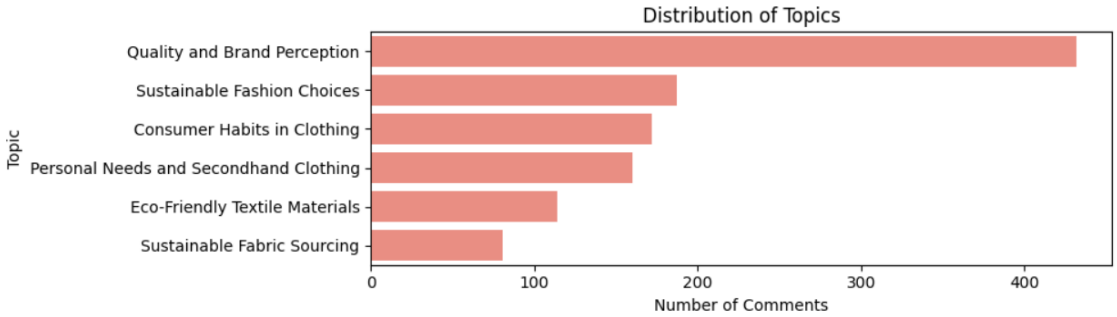
This multidimensional approach provided nuanced insights into how consumers express their attitudes toward sustainable fashion.

4. Supervised Learning - Classification Model : The supervised learning component involved developing and comparing multiple classification models to predict topic assignments based on the output from NMF topic modeling. The dataset is split into training (80%) and testing (20%) sets and evaluated several algorithms: Random Forest, Support Vector Machine (SVM), XGBoost, and Logistic Regression. Each model was trained on the document-term matrix derived from the preprocessed text data, with the target variable being the assigned topics from the NMF analysis. Standard hyperparameters were used for each model to establish a baseline performance.
To evaluate the models, standard classification metrics including accuracy, precision, recall, and F1-score are calculated.

4 Results

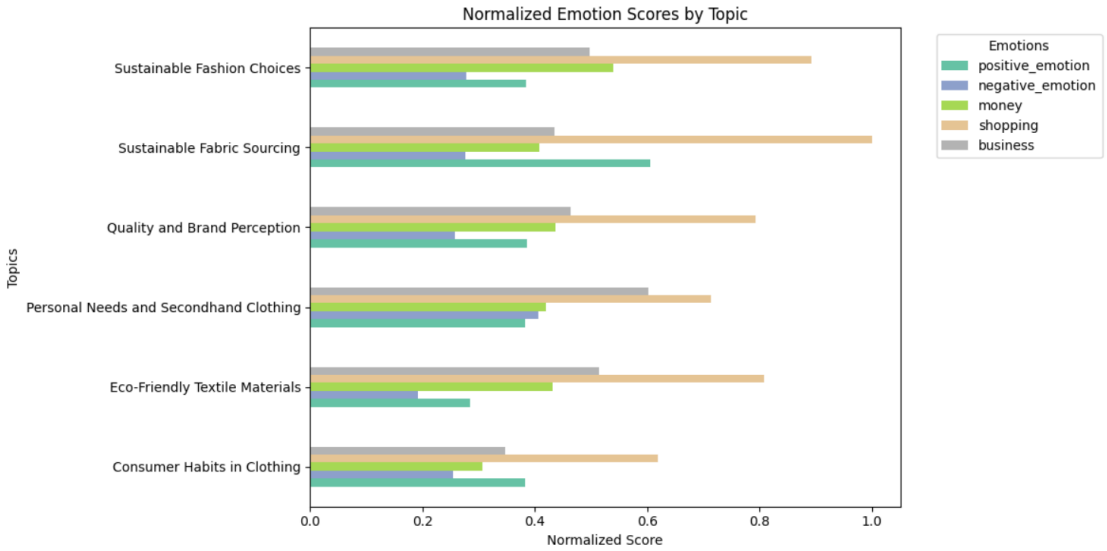
Topic Modeling **Overall Sentiment Distribution** TThe analysis of sustainable fashion discourse revealed six distinct topics through Non-Negative Matrix Factorization (NMF), achieving a coherence score of 0.6437. Here are the identified topics and their key terms:

Topic	Label	Key Terms(Top 10)
0	Sustainable Fashion Choices	sustainable, fashion, brand, fast, choice, better, need, solution, company, people
1	Consumer Habits in Clothing	clothes, new, buy, buying, clothing, wear, year, used, shop, dont
2	Personal Needs and Secondhand	people, dont, secondhand, thing, need, lot, clothing, want, size, im
3	Eco-Friendly Materials	cotton, fabric, organic, sustainable, recycled, polyester, water, plastic, fiber, good
4	Quality and Brand Perception	like, im, quality, make, brand, good, think, really, know, look
5	Sustainable Fabric Sourcing	fabric, item, buy, small, synthetic, like, question, process, place, time



Quality and Brand Perception had the highest representation.
Sustainable Fabric Sourcing had the lowest representation suggesting that discussions tend to focus more on brand perception and general sustainability practices rather than technical aspects of fabric sourcing.

Sentiment Analysis The Empath analysis revealed varying emotional patterns across topics. Sustainable Fashion Choices showed the highest positive emotion scores (0.82), while Consumer Habits displayed balanced emotional content. Eco-Friendly Materials demonstrated increased business-related terminology, and Quality and Brand Perception exhibited strong shopping-related sentiment.

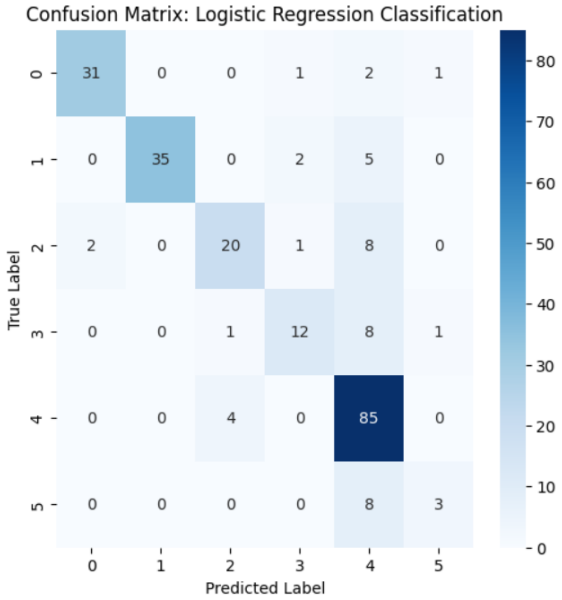


Business and shopping sentiments dominate most topics. Positive emotions generally outweigh negative emotions across all topics. Money-related sentiment shows moderate presence across topics, particularly in consumer-focused discussions. Personal Needs and Secondhand Clothing shows the most balanced distribution of all emotional categories. Sustainable Fabric Sourcing has the highest proportion of shopping-related sentiment. These patterns suggest that while the discourse is generally positive, it's heavily focused on business and shopping aspects of sustainable fashion, with varying degrees of emotional engagement across different topics.

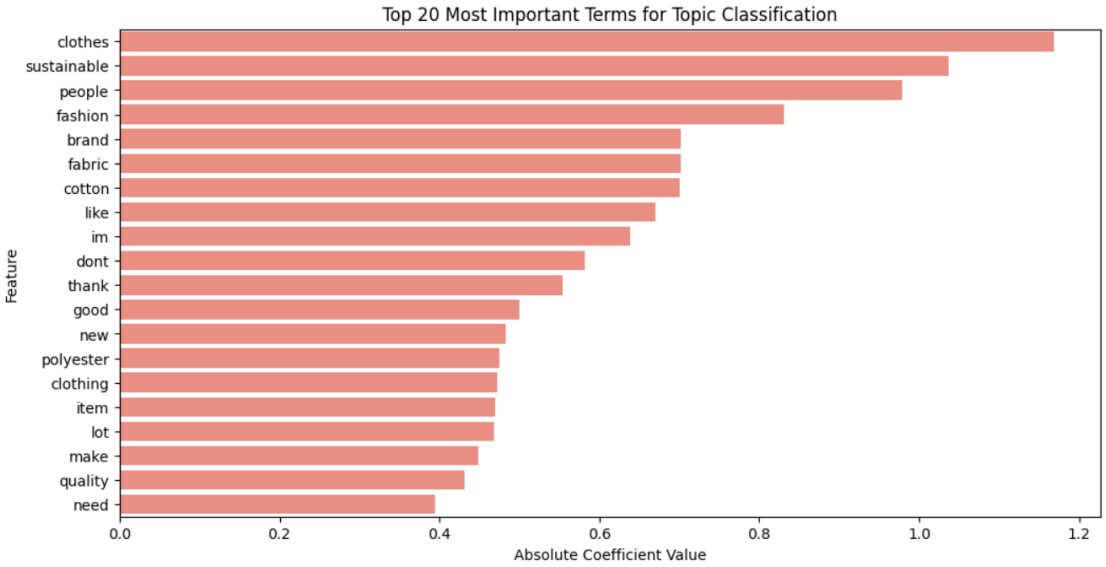
Classification Performance The Logistic Regression classifier achieved superior performance among tested models:

Model	Accuracy	Weighted Avg F1-Score
Logistic Regression	81%	0.80
XGBoost	75%	0.74
Random Forest	73%	0.72
SVM	67%	0.63

Performance across topics varied significantly, with Sustainable Fashion Choices and Consumer Habits achieving the highest F1-scores (91%), while Sustainable Fabric Sourcing proved more challenging to classify (38% F1-score).



Strong diagonal pattern indicating good classification accuracy Most misclassifications occurred between related topics Topic 5 (Sustainable Fabric Sourcing) had the most false negatives Topics 0 and 4 showed the strongest classification performance.



Material-specific words (cotton, polyester, fabric) showed high importance. Consumer behavior terms (buy, wear, need) were also significant. Brand-related terminology appeared frequently among important features

5 Conclusion and Limitations

Conclusion The analysis of sustainable fashion discourse on Reddit revealed several key insights about how consumers navigate and discuss their consumption choices. Through topic modeling and sentiment analysis, we identified six distinct conversation themes, with Quality and Brand Perception dominating the discourse, followed by Sustainable Fashion Choices and Consumer Habits. The high coherence score (0.6437) of our topic model indicates strong thematic consistency in these discussions. Consumer behavior patterns emerged clearly through the emotional analysis. Shopping-related sentiment was particularly strong in discussions about Sustainable Fabric Sourcing and Sustainable Fashion Choices, suggesting that

consumers actively seek information about sustainable purchasing options. Business-related terminology dominated conversations about Eco-Friendly Textile Materials, indicating a strong interest in industry practices and manufacturing processes. The relatively balanced distribution of positive and negative emotions across topics suggests that consumers approach sustainable fashion with both optimism and critical awareness.

Limitations The study faced several notable limitations. First, the dataset was limited to Reddit discussions, which may not represent the broader consumer population due to the platform's demographic skew. Second, the sentiment analysis through Empath revealed some challenges in capturing nuanced emotional content, particularly in technical discussions about materials and manufacturing processes. This is evident in the normalized emotion scores, where Sustainable Fabric Sourcing showed unexpectedly high shopping-related sentiment despite its technical nature. Finally, the classification model's performance varied significantly across topics, with particularly low accuracy in identifying discussions about Sustainable Fabric Sourcing (38% F1-score), suggesting that some nuanced conversations may have been misclassified or oversimplified in our analysis.

References

Orminski, J., Tandoc Jr, E. C., & Detenber, B. H. (2021). Sustainable fashion discourse on Twitter: Network and content analysis of top influencers. Environmental Communication, 15(8), 1060-1079.

Mukendi, A., Davies, I., Glozer, S., & McDonagh, P. (2020). Sustainable fashion: current and future research directions. European Journal of Marketing, 54(11), 2873-2909.

Cervellon, M. C., & Wernerfelt, A. S. (2012). Knowledge sharing among green fashion communities online: Lessons for the sustainable supply chain. Journal of Fashion Marketing and Management, 16(2), 176-192.

Abbate, S., et al. (2024). The evolution of the academic discourse on sustainability in the fashion industry: A bibliometric analysis. Contemporary Social Science.