Project Report: Understanding Consumer Sentiment Towards Eco-Friendly Products

---

Introduction

In an era where environmental consciousness is increasingly significant, understanding consumer attitudes towards eco-friendly products is essential for businesses aiming to align their strategies with contemporary values. This project, titled **"Understanding Consumer Sentiment Towards Eco-Friendly Products,"** explores consumer perceptions by analyzing data from Trustpilot and Reddit. The goal is to provide businesses with valuable insights to better meet the growing demand for sustainable products. This report details the project's steps, including data collection, preprocessing, analysis, and visualization, and discusses the challenges and modifications encountered throughout the process.

Description of the Task

The main objective of this project was to gather and analyze data to gauge consumer sentiment towards eco-friendly products. This involved collecting reviews and social media posts from diverse sources, processing the data for consistency and quality, and analyzing it to extract meaningful insights. The project was divided into several key phases: data collection, data preprocessing, sentiment analysis, topic modeling, trend analysis, and visualization. Each phase was crucial in providing a comprehensive understanding of consumer opinions and identifying trends in the data.

The approach involved two primary sources of data:

1. **Trustpilot:** Reviews were collected to analyze the sentiment expressed in text and compare it with the numerical ratings provided by users. This allowed for a correlation analysis between the expressed sentiment and the formal rating of eco-friendly products.

2. **Reddit:** Posts were collected to perform a trend analysis over time, focusing on how discussions and sentiments around eco-friendly topics evolved. This provided insights into the broader social discourse on sustainability and eco-friendly living.

Implementation Process for Sentiment Analysis vs. Rating Category with TrustPilot

1. Data Collection:

Initial Setup and Tool Selection: To begin data collection, a Python-based environment was established using essential libraries for web scraping and data manipulation. The requests library facilitated seamless retrieval of web pages, and BeautifulSoup enabled detailed HTML parsing to extract specific elements from the review content. Additionally, pandas was employed for data manipulation and storage, while nltk supported subsequent natural language processing tasks, including sentiment analysis.

Selecting Target Companies: A set of eco-friendly companies with a notable presence on Trustpilot was identified. The companies chosen for this study included EarthHero, The Green Company, and EcoLeaf Products. These companies were selected based on their commitment to sustainability and their significant consumer base, which provided a wealth of reviews for analysis.

Structuring the Data Collection Process: The data collection process was structured to ensure consistency and completeness. A systematic approach was used to collect reviews from multiple pages of each company's Trustpilot profile, covering a broad spectrum of

customer feedback. For each company, the process iterated through pages 1 to 6 to capture a substantial number of reviews while avoiding redundancy and diminished relevance that might occur if only the most recent reviews were considered.

Extracting Relevant Information: Upon retrieving the HTML content of each review page using the requests library, BeautifulSoup was employed to parse the HTML and navigate through the document structure. The primary focus was on extracting specific elements of each review, including the username of the reviewer, the total number of reviews they had submitted, their location, the date of the review, the content of the review itself, and the rating they assigned to the product. This required identifying and targeting the relevant HTML tags and classes encapsulating each piece of information.

Handling Data Consistency: Ensuring that all lists of extracted data (usernames, review counts, locations, dates, content, and ratings) had consistent lengths was a challenge. Discrepancies in list lengths could arise due to missing elements or variations in the HTML structure across different reviews. To address this, a method was implemented to pad shorter lists with placeholder values, such as empty strings, ensuring that all lists maintained the same length and could be accurately combined into a structured data format.

Combining and Storing Data: Once extracted and cleaned, the data was compiled into a pandas DataFrame. Each row represented a review, with columns corresponding to the extracted elements such as company name, username, total reviews, location, date, content, and rating. This structured format facilitated efficient data manipulation and analysis, enabling a comprehensive analysis of consumer sentiment towards eco-friendly products. The data was saved to a CSV file (all_reviews.csv), serving as a consolidated repository for all the reviews gathered across the selected eco-friendly companies.

2. Data Cleaning and Preparation:

Clean and consistent text data is crucial for reliable analysis. A clean_text function was developed to standardize the text by converting it to lowercase and removing URLs, punctuation, special characters, and extra spaces. This function used regular expressions (re.sub) to remove unwanted elements, resulting in a clean and uniform text dataset. Missing values in the 'Content' column were replaced with empty strings, and the 'Rating' column's missing values were filled with the median rating to avoid bias in the analysis. This step ensured that the text data was ready for accurate sentiment analysis and topic modeling.

3. Sentiment Analysis:

The VADER sentiment analyzer from nltk was used to classify the sentiment of each review. VADER is effective for analyzing sentiment in short texts, such as reviews. The analyze_sentiment function applied VADER to calculate a compound score for each review, categorizing them into positive, neutral, or negative sentiments based on the score. Reviews with a compound score above 0.05 were classified as positive, those below -0.05 as negative, and those in between as neutral. The sentiment classification was added to the dataset, providing a clear measure of customer feelings towards eco-friendly products.

4. Text Preprocessing for Topic Modeling:

Text preprocessing was essential for effective topic modeling. A preprocess_text function was used to tokenize the text, remove non-alphabetic characters, and filter out common stop words. This preprocessing ensured that the text was focused on meaningful words relevant to the analysis. The function used nltk's word_tokenize to split the text into tokens, retaining only alphabetic words and removing stop words. The cleaned and tokenized text was then used to create a document-term matrix with CountVectorizer.

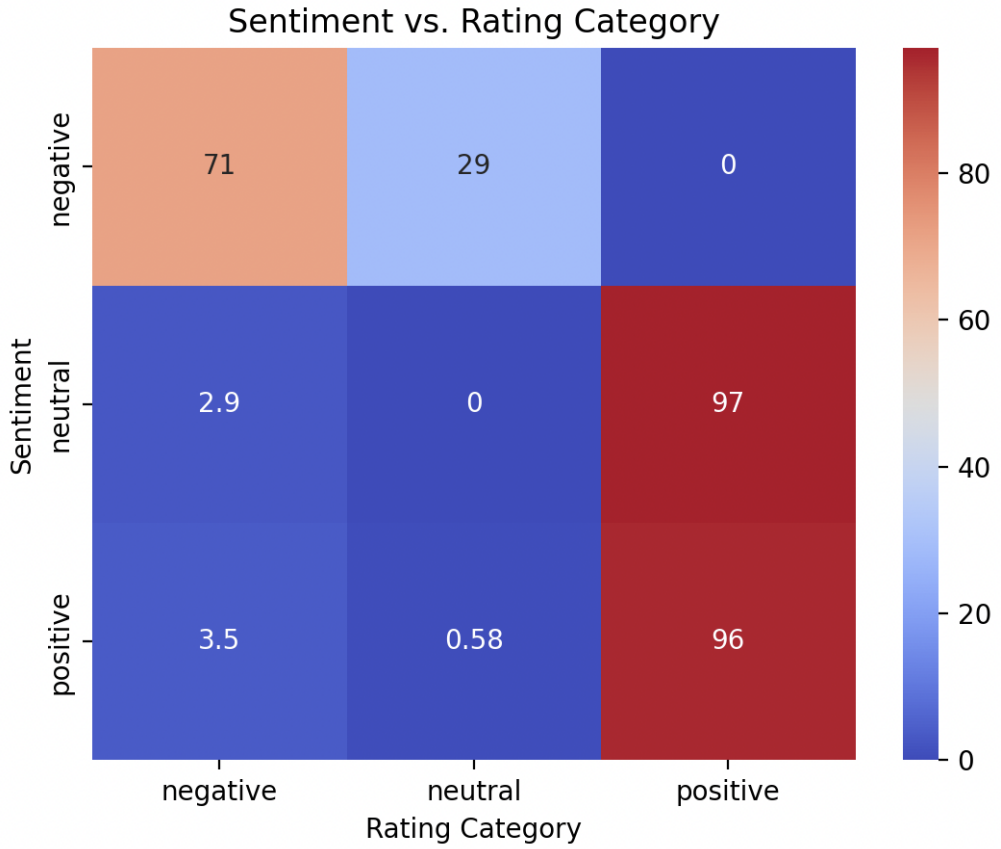
5. Topic Modeling with LDA:

The Latent Dirichlet Allocation (LDA) model was used to uncover hidden topics within the reviews. LDA is a powerful tool for identifying thematic structures in large text collections. The CountVectorizer transformed the cleaned text into a document-term matrix, capturing word frequencies across the reviews. The LDA model then used this matrix to identify five distinct topics, each represented by a set of key words. A function was defined to display the top words for each topic, providing insights into the main themes discussed by the reviewers.

Sentiment and Rating Analysis: To ensure the accuracy of the sentiment analysis, a cross-tabulation of sentiment and Trustpilot ratings was performed. Trustpilot ratings, ranging from 1 to 5, were categorized as positive (4 and 5), neutral (3), and negative (1 and 2). The cross-tabulation showed the distribution of sentiment within each rating group, providing a nuanced view of how sentiment corresponds to ratings. A heatmap visualized this relationship, helping to identify any discrepancies between sentiment scores and ratings.

The correlation analysis served as a validation check for the sentiment analysis results. By comparing sentiment scores with ratings, the analysis verified that positive sentiments generally corresponded to higher ratings and negative sentiments to lower ratings. This validation ensured that the sentiment classification accurately reflected customer opinions. Accurate sentiment analysis is critical for understanding consumer attitudes and guiding strategic decisions to enhance customer satisfaction in the eco-friendly market.

Summary of Sentiment vs. Rating Category Heatmap:



- **Negative Sentiment and Ratings:** 71% of reviews with negative sentiment were accompanied by negative ratings. This strong alignment indicates that customers who express dissatisfaction in their reviews tend to reinforce their negative feedback with lower numerical ratings, demonstrating a consistent reflection of their disappointment both textually and numerically. This underscores the reliability of using sentiment analysis as a proxy for understanding customer dissatisfaction.

- **Neutral Sentiment and Ratings:** 97% of reviews categorized as neutral in sentiment received positive ratings. This finding is intriguing as it suggests that even when customers do not express strong emotions in their text, their overall experience with the product is sufficiently positive to warrant a higher rating. It reflects a scenario

where the lack of emotional expression does not equate to dissatisfaction but rather indicates contentment or a satisfactory experience that meets their expectations.

- **Positive Sentiment and Ratings:** 96% of reviews with positive sentiment were paired with positive ratings. This high level of consistency underscores that satisfied customers not only express their approval through positive language but also reinforce it with high ratings. The correlation between positive textual feedback and high numerical ratings highlights that customers' satisfaction with eco-friendly products is clearly communicated through both forms of feedback.

- **Discrepancies and Anomalies:** 29% of reviews with negative sentiment had neutral ratings, and 3.5% of reviews with positive sentiment received negative ratings. These anomalies suggest that there are instances where the sentiment expressed in the text does not fully align with the numerical rating. This misalignment indicates that other factors may influence the ratings that are not captured by the sentiment in the text alone, highlighting the complexity of customer feedback and the need for nuanced analysis.

Overall, the heatmap demonstrates a robust alignment between customer sentiment and their ratings, particularly at the extremes of negative and positive feedback. This alignment suggests that textual feedback and numerical ratings together provide a reliable indicator of customer sentiment, even in cases where neutral sentiment is expressed, as it tends to lean towards positive ratings. The consistent correlation between sentiment and ratings offers a clear and reliable measure of customer satisfaction through their words and ratings.

Sentiment Trend Analysis on Reddit

1.Data Collection:

Initial Setup and Tool Selection: To collect data from Reddit, I established a Python environment equipped with the necessary libraries for API interaction and data management. The Python Reddit API Wrapper (PRAW) was chosen for its ease of integration with Reddit's API, allowing seamless access to Reddit data. Pandas was also included to facilitate the manipulation and storage of the data in a structured format. This setup provided a robust framework for gathering and organizing large volumes of Reddit posts effectively.

Defining Target Subreddits and Search Criteria: I selected several subreddits that are central to discussions on eco-friendly topics. The subreddits chosen were EcoFriendly, BuyItForLife, ZeroWaste, Sustainable, and GreenLiving. These communities are known for their focus on sustainable living, durable products, zero-waste practices, and general green living tips. To filter relevant posts, I defined a search query using the keyword “eco-friendly.” This query ensured that the collected data was specifically related to the topic of interest, providing a relevant and focused dataset for analysis.

Structuring the Data Collection Process: The data collection process was carefully structured to ensure thorough coverage and consistency across the selected subreddits. For each subreddit, I utilized the PRAW library to search for posts containing the keyword “eco-friendly,” setting a limit of 1,000 posts per subreddit. This limit was chosen to balance the comprehensiveness of the data collection with the practicality of data management. For each post retrieved, I extracted key details such as the post title, body content, score (indicating the popularity or quality of the post), number of comments, creation date (in Unix timestamp format), author’s username, and the subreddit from which the post originated.

Extracting Relevant Information: Using the PRAW library, I iterated through the search results to extract pertinent details from each post. The elements of interest included the title

and body content of the post, the score, the number of comments, the creation date, the author's username, and the subreddit name. These attributes were crucial for understanding the context and engagement level of each post. The extracted data was organized into lists for each attribute, ensuring that each piece of information was collected systematically and could be easily converted into a structured format for analysis.

Handling Data Consistency: To address potential inconsistencies in the extracted data, I implemented methods to standardize and validate the information. Since posts could vary in structure and content, ensuring that each list of attributes (titles, contents, scores, etc.) had consistent lengths was essential. This involved padding shorter lists with placeholder values, such as empty strings, to maintain uniformity. This approach ensured that the data could be combined without discrepancies, providing a clean and complete dataset for further analysis.

Combining and Storing Data: Once the data was extracted from each subreddit, I combined the individual datasets into a single DataFrame using pandas. This DataFrame provided a structured format where each row represented a unique post, and each column corresponded to a specific attribute, such as title, content, score, comments, date, author, and subreddit. This consolidated format facilitated efficient data manipulation and comparison across posts from different subreddits. The final dataset was saved to a CSV file named reddit_data.csv, which served as a comprehensive repository of all the collected posts, ready for subsequent analysis.

2. Data Cleaning:

To ensure the text data was clean and consistent for analysis, the clean_text function was applied to the 'Content' column. This function removed URLs, mentions, hashtags, and non-alphabetic characters, converting the text to lowercase. It then tokenized the text and removed common stop words to retain only meaningful words. This cleaning process was crucial for

improving the quality of the text data and making it suitable for subsequent sentiment analysis and keyword extraction.

3. Sentiment Analysis:

A sentiment analysis was performed on the cleaned text to categorize each post as positive, neutral, or negative. The SentimentIntensityAnalyzer from the nltk library was used to calculate a compound sentiment score for each post. Posts with a compound score above 0.05 were classified as positive, those below -0.05 as negative, and the rest as neutral. This sentiment classification provided a clear measure of the emotional tone of the posts, which was added as a new column, 'Sentiment', in the DataFrame. The sentiment analysis results were then saved to a new CSV file, reddit_data_with_sentiment.csv, ensuring that the sentiment information was preserved for further analysis and visualization.

4. Keyword Analysis Using TF-IDF:

For the keyword analysis, the TfidfVectorizer was used to transform the cleaned text into a term-frequency inverse document frequency (TF-IDF) matrix. This matrix highlighted the importance of words in the context of the entire dataset. The top 100 most relevant keywords were identified, and the term frequencies were computed to find the top 10 keywords discussed in the posts. This analysis provided insights into the key topics and trends in the eco-friendly discussions on Reddit.

5. Sentiment Trend Analysis Over Time:

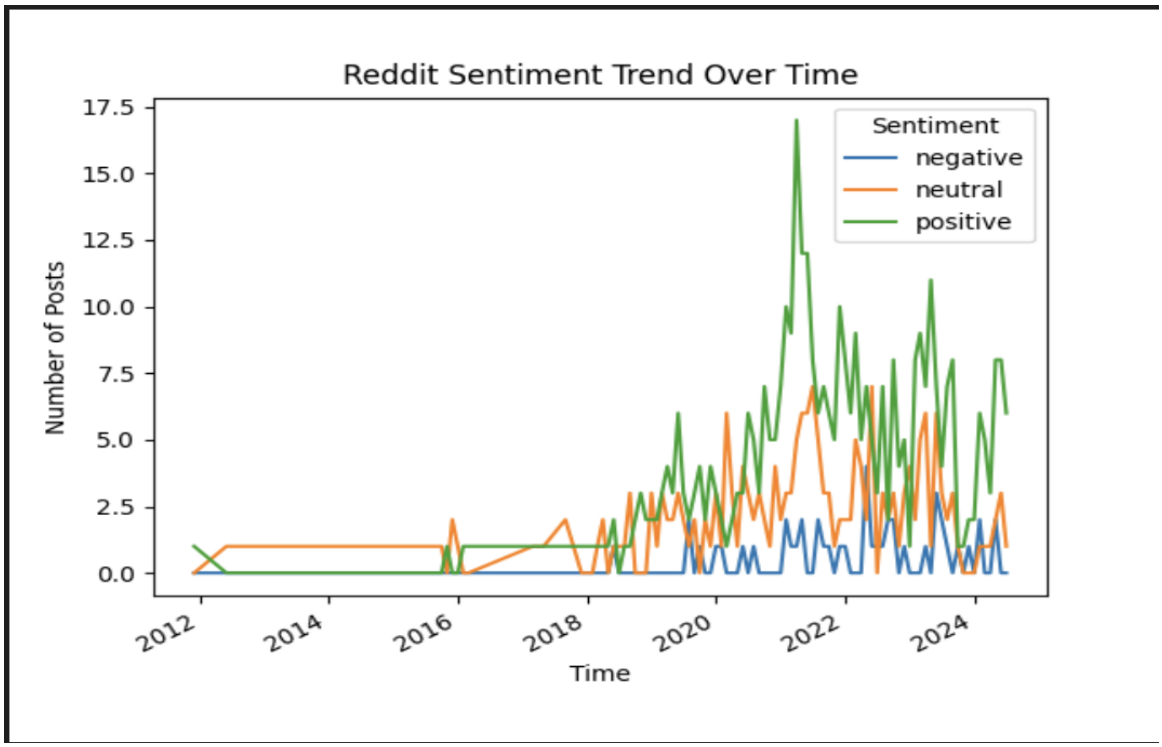
To examine how sentiment has evolved over time, the posts were aggregated by month, and the sentiment counts were calculated. The resulting sentiment trends were plotted to visualize the monthly changes in the number of posts categorized as positive, neutral, or negative. This visualization helped identify periods of increased or decreased interest in eco-friendly topics

and provided a clear view of how sentiment towards these topics has changed over time. A line plot of sentiment trends was generated, showing the monthly variations in sentiment and highlighting key trends and patterns in the data. The plot was saved as reddit_sentiment_trend_over_time.png .

6. Keyword-Specific Trend Analysis:

A keyword-specific trend analysis was conducted to track the sentiment and frequency of posts mentioning a particular keyword, such as "sustainable". The cleaned text was filtered to include only posts containing the specified keyword. The filtered posts were then aggregated by month, and the sentiment counts for each month were calculated. This keyword trend analysis provided insights into the popularity and sentiment associated with specific eco-friendly topics over time. A line plot showing the trend for the keyword "sustainable" was created, displaying the monthly sentiment distribution and indicating shifts in discussion focus and sentiment. The plot was saved as trend_for_sustainable_over_time.png.

Summary of Reddit Sentiment Trend Results:



1. General Reddit Sentiment Trend Over Time:

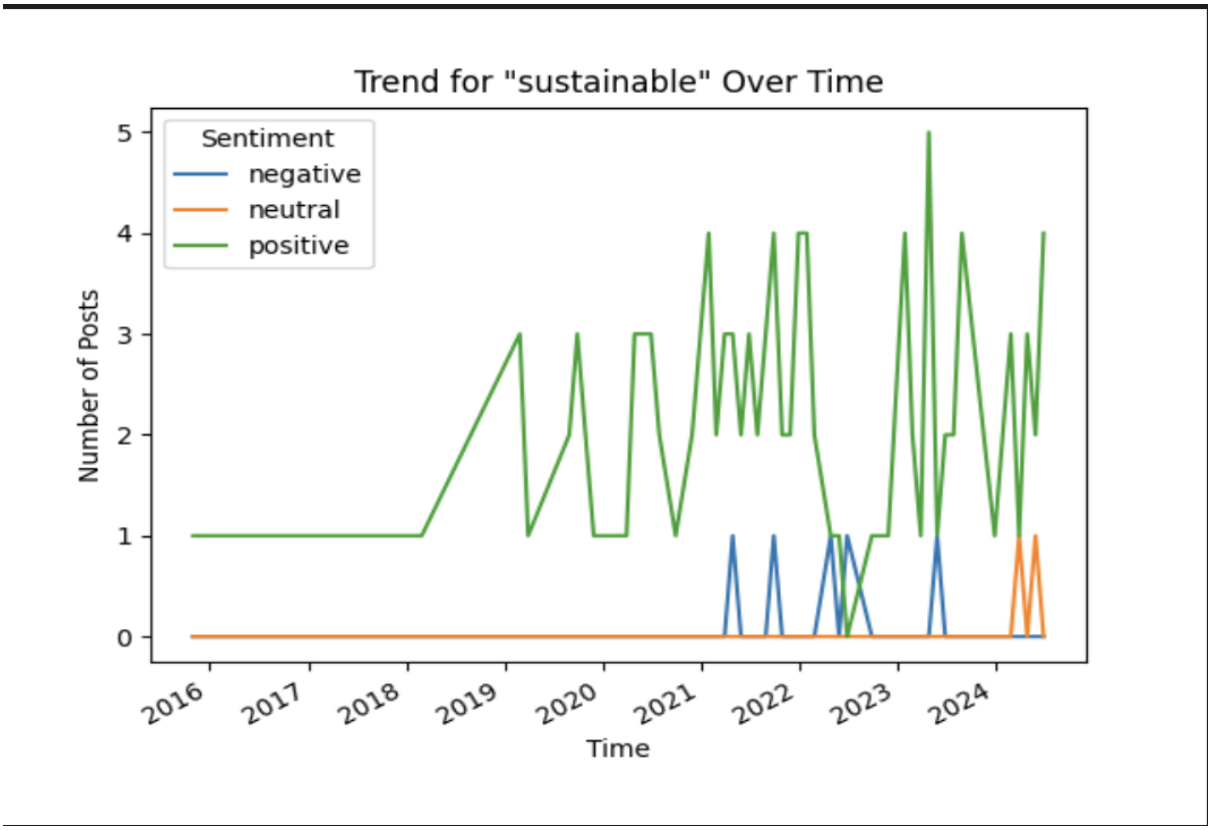
- **Positive Sentiment:** The number of positive posts has significantly increased over time, particularly from 2020 onwards. This upward trend suggests a growing interest and positive attitude towards eco-friendly topics on Reddit. Peaks in positive sentiment are notably higher than those in neutral and negative sentiments, indicating that the majority of discussions reflect a favorable view of eco-friendly practices and products.

- **Neutral Sentiment:** Posts with neutral sentiment also show an increase over time, although not as pronounced as positive posts. This indicates a steady engagement with eco-friendly topics, where users are sharing information or discussing topics without a

strong emotional tone. The fluctuations in neutral posts suggest ongoing, varied discussions that are informational rather than opinionated.

- **Negative Sentiment:** The frequency of negative posts is relatively low compared to positive and neutral posts. However, there is a slight increase in negative sentiment posts starting around 2020. This could reflect growing scrutiny or criticism of certain eco-friendly practices or products, indicating areas where improvements or better communication might be needed.

The overall increase in positive sentiment highlights a trend towards greater awareness and acceptance of eco-friendly products and practices. The presence of neutral posts suggests an ongoing dialogue, while the low but increasing negative sentiment indicates areas for potential concern or improvement within the eco-friendly sector. Businesses can use these insights to understand consumer attitudes and enhance their engagement with eco-friendly initiatives.



2. Trend for "Sustainable" Over Time:

- **Positive Sentiment:** The trend for the keyword "sustainable" shows a strong increase in positive sentiment, particularly from 2018 onwards. This indicates that discussions around sustainability are generally favorable, reflecting widespread support and enthusiasm for sustainable practices and products. The consistency of positive sentiment over time underscores the growing importance of sustainability in consumer discussions.

- **Neutral Sentiment:** There are very few posts with neutral sentiment related to the keyword "sustainable." This suggests that discussions around sustainability are typically opinionated, with users either expressing strong support (positive sentiment) or, less frequently, criticism (negative sentiment).

- **Negative Sentiment:** The number of negative posts mentioning "sustainable" remains low, but there are occasional spikes starting from 2021. These spikes could be attributed to specific events or issues that caused dissatisfaction or controversy regarding sustainability practices. The minimal negative sentiment indicates that while there are some concerns, they are not prevalent.

The trend for "sustainable" demonstrates a clear preference for and positive engagement with sustainability-related topics on Reddit. The high positive sentiment suggests that consumers view sustainability as a crucial and beneficial aspect of products and practices. The sporadic negative sentiment highlights potential areas of contention or disappointment, which businesses can address to improve their sustainability initiatives and better meet consumer expectations. This keyword-specific analysis provides valuable insights into the evolving discourse around sustainability and can guide businesses in their efforts to promote sustainable practices.

Consumer Attitudes Towards Eco-Friendliness

The overall trend towards positive sentiment and high ratings underscores the growing consumer preference for eco-friendly products. This positive feedback highlights that consumers place significant value on the eco-friendly attributes of products, reflecting an increasing awareness and concern for environmental sustainability. The alignment of positive sentiment and high ratings indicates that eco-friendly products are not just meeting consumer needs but are also appreciated for their environmental benefits. This trend reflects a broader market movement where eco-consciousness is becoming a significant factor in consumer decision-making, moving beyond niche preferences to a mainstream consideration.

The combined insights from Trustpilot and Reddit paint a comprehensive picture of consumer attitudes towards eco-friendliness:

Positive and Growing Interest: There is a clear and growing positive sentiment towards eco-friendly products and practices, indicating widespread support and a desire for sustainable options. Both platforms reflect a trend towards greater acceptance and enthusiasm for eco-friendly initiatives.

Demand for Authenticity and Improvement: Consumers value honesty and transparency in eco-friendly claims, and any perceived discrepancies can lead to negative feedback. The slight increase in negative sentiment points to areas where businesses can improve, especially in terms of authenticity and effectiveness.

Continued Engagement: The rise in neutral sentiment posts and positive ratings for neutral sentiment reviews highlight an ongoing engagement with eco-friendly topics. Consumers are actively discussing and exploring sustainable options, even if they are not always expressing strong emotions.

Conclusion

The combined analysis from Trustpilot and Reddit provides a nuanced and comprehensive view of consumer attitudes towards eco-friendly products. The strong alignment between positive sentiment and high ratings underscores the increasing consumer preference for sustainable products, reflecting a broader shift towards eco-consciousness. These insights offer valuable guidance for businesses to enhance their eco-friendly offerings, ensuring they meet the expectations of an increasingly environmentally aware consumer base. By addressing areas of concern and promoting transparency, businesses can foster deeper

connections with eco-conscious customers and contribute positively to the sustainability movement.