Sentiment Analysis of Product Reviews with a Hybrid Recommendation System

Pradyumna Kombethota Ramgopal
Virginia Tech
Alexandria, Virginia, USA
pradyumna@vt.edu

Sujith Reddy A
Virginia Tech
Alexandria, Virginia, USA
sujithreddy@vt.edu

Mrunaldhar Bathula
Virginia Tech
Alexandria, Virginia, USA
mrunaldharb23@vt.edu

Abstract

Online product reviews have become a crucial source of information for consumers and businesses alike. However, the vast volume of user-generated content poses challenges in identifying which reviews are genuinely informative and how they influence purchasing decisions. This project presents a comprehensive analysis of Amazon Fine Food Reviews, comprising two key components: Sentiment Analysis and a Recommendation System. We explore how the sentiment expressed in a review—extracted using advanced natural language processing techniques—relates to its perceived helpfulness by other users. Additional factors such as review length, sentiment, and score are also analyzed to understand their impact on helpfulness scores.

In the first phase, we apply VADER sentiment analysis to the full review text, and investigate how these sentiment scores correlate with the HelpfulnessNumerator and HelpfulnessDenominator metrics. This analysis reveals patterns that can be used to highlight high-quality, informative reviews while suppressing low-quality or misleading content.

In the second phase, we develop a hybrid recommendation system that combines user-based collaborative filtering with sentiment polarity from review data. By integrating both numerical ratings and sentiment polarity, the system prioritizes products backed by emotionally strong and widely endorsed reviews. This dual approach enhances the relevance and trustworthiness of recommendations provided to end users.

The outcome of this project includes insights into what makes a review helpful, visualizations of user engagement patterns, and a recommendation system that leverages both user similarity and review sentiment. Together, these contributions aim to improve the user experience on e-commerce platforms by promoting more reliable and meaningful content.

Keywords

Sentiment Analysis, VADER Sentiment, NLP, Recommendation System

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
Conference acronym 'XX, Woodstock, NY
© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

ACM Reference Format:
Pradyumna Kombethota Ramgopal, Sujith Reddy A, and Mrunaldhar Bathula. 2018. Sentiment Analysis of Product Reviews with a Hybrid Recommendation System. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

In today's digital marketplace, user-generated content—particularly online product reviews—has become a cornerstone of consumer decision-making. On platforms like Amazon, customers rely heavily on the experiences shared by others to guide their purchases. These reviews influence buying behavior, impact brand perception, and play a significant role in shaping product reputations. However, with the ever-increasing volume of reviews, users often struggle to identify which ones are genuinely useful. Not all reviews are equally informative—some are detailed and insightful, while others may be vague, biased, or even misleading.

The core problem addressed in this project is two-fold. First, we seek to understand what makes a review "helpful" and how this perceived helpfulness relates to the sentiment expressed in the text. Second, we aim to enhance the product recommendation process by incorporating sentiment data, creating a more intelligent and nuanced suggestion system for users. These tasks are crucial not only for improving the browsing experience of online shoppers, but also for businesses that depend on customer feedback to refine their products and services.

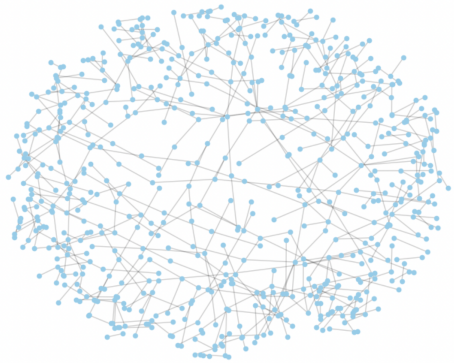

Figure 1: User-Product Interaction Network.

To tackle this, we use the Amazon Fine Food Reviews dataset, a large-scale collection of over half a million food-related product

reviews spanning more than a decade. Our analysis begins with Sentiment Analysis, where we apply the VADER sentiment analysis technique to the full review text to extract sentiment polarity. We then investigate how sentiment, in conjunction with features such as review length and rating-sentiment alignment, correlates with the helpfulness votes each review receives (captured by the HelpfulnessNumerator and HelpfulnessDenominator fields). These insights help reveal what types of reviews are more valued by other users.

In the second part of the project, we design a hybrid recommendation system that enhances traditional user-based collaborative filtering by incorporating sentiment polarity from reviews. This allows the model to not only consider numerical ratings but also the emotional tone of user feedback—prioritizing reviews that are not only positive but also persuasive and helpful. Such a system aims to suggest products that are highly rated and positively reviewed in both quantity and quality.

This project is important to a wide range of stakeholders. Consumers benefit from more accurate and relevant review surfacing, which can help them make better purchasing decisions with less effort. E-commerce platforms can improve user satisfaction and engagement, while data scientists and researchers gain insight into how textual sentiment can be meaningfully integrated into predictive systems. By combining natural language processing with recommendation algorithms, our work bridges the gap between subjective opinion and algorithmic intelligence, contributing to smarter, more user-aware systems.

2 Background and Related Work

The rise of user-generated reviews on e-commerce platforms such as Amazon has prompted extensive research into the automatic analysis of review content to support decision-making and recommendation systems. Two critical areas in this domain are sentiment analysis and recommendation systems, both of which form the core of our project.

Several studies have explored sentiment classification using the Amazon Fine Food Reviews dataset, a widely used benchmark for review-based sentiment tasks. Harsha et al. [4] performed an empirical evaluation using multiple classifiers including Logistic Regression, XGBoost, Decision Tree, and Naïve Bayes. They emphasized preprocessing techniques such as stop-word removal, lemmatization, and vectorization using Bag-of-Words and TF-IDF.

Yarkareddy et al. [2] also conducted a detailed analysis of this dataset, applying machine learning algorithms, alongside feature extraction techniques including TF-IDF, Word2Vec, and n-grams. Their results demonstrated that Support Vector Machines using TF-IDF vectorization yielded the best performance, with accuracy reaching up to 94%.

The helpfulness of reviews has also garnered research interest. As noted by Harsha et al. [4], Amazon's helpfulness voting system significantly impacts product visibility and revenue. However, newly posted reviews often suffer from low visibility due to the absence of early helpfulness votes. To address this, researchers have proposed predictive models that estimate helpfulness based on sentiment, length, and textual features of reviews.

On the recommendation system front, collaborative filtering has been a longstanding approach. Ekstrand et al. [6] provide a foundational survey on recommender systems, highlighting matrix factorization, user-item interaction models, and hybrid strategies that integrate collaborative and content-based filtering. A more recent hybrid framework by Islam et al. [7] incorporates K-means clustering, TF-IDF, and matrix factorization to enhance recommendation quality and address cold-start and sparsity challenges. Their work demonstrates how sentiment polarity and item similarity can be combined effectively to produce more personalized suggestions.

Our project is informed by these works and takes a dual-pronged approach. First, we analyze sentiment and helpfulness using the VADER sentiment analyzer on the review text and study how sentiment polarity, rating alignment, and review length correlate with helpfulness scores. Second, we build a hybrid recommendation system that integrates sentiment polarity into a collaborative filtering framework, aiming to suggest products that are not only highly rated but also backed by emotionally rich and widely endorsed reviews.

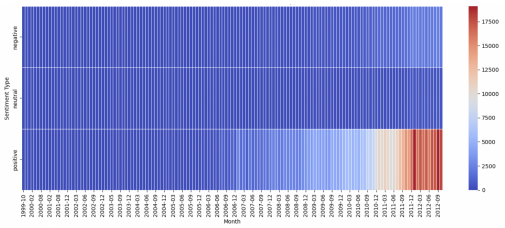

Figure 2: Sentiment Trends Over Time.

Zhao and Sun [1] applied a BERT-based model to the Amazon Fine Food Reviews dataset and demonstrated its ability to predict review scores from the text with high accuracy. Their work involved comprehensive preprocessing steps including the removal of non-English reviews, punctuation, and short comments, as well as balancing the dataset through stratified resampling.

Moreover, Bhati and Kher [3] presented a broad survey of techniques applied to the Amazon Fine Food Reviews dataset, emphasizing how sentiment analysis has evolved from simple rule-based models to sophisticated machine learning pipelines. Their review highlights the importance of data cleaning, text normalization, and the use of vectorization techniques such as TF-IDF and Bag of Words (BoW) for feature extraction

3 Approach

Our project framework is divided into two core components: (1) Sentiment Analysis, and (2) a Hybrid Recommendation System. We designed a pipeline that begins with preprocessing and sentiment extraction, proceeds to helpfulness evaluation, and finally integrates these insights into a sentiment-aware collaborative filtering recommendation model. The complete architecture of the sentiment analysis module is illustrated in Figure

3.1 Data Preprocessing

The input to our pipeline is raw review text from the Amazon Fine Food Reviews dataset. Preprocessing is essential to reduce noise and standardize the input for downstream tasks. This involves:

- **Lowercasing all text:** Convert all characters to lowercase to ensure uniformity and treat words like "Good" and "good" as the same.
- **Removing URLs:** Strip out any web links (e.g., `https://...`) that do not contribute to sentiment.
- **Removing Emails:** Detect and remove email addresses which are not relevant for analysis.
- **Removing special characters, hashtags, numbers, and punctuation:** Clean the text by eliminating symbols like #, @, !, and numbers that don't add semantic value.
- **Expanding contractions:** Convert terms like "don't" and "can't" into "do not" and "cannot" to improve token clarity.
- **Part-of-Speech (POS) tagging:** Assign grammatical roles (noun, verb, etc.) to each word, aiding in proper lemmatization.
- **Tokenization using `nltk.word_tokenize`:** Split the text into individual words for further processing.
- **Stopword removal using `nltk.corpus.stopwords`:** Eliminate common words like "is", "the", "and" that carry minimal sentiment.
- **Lemmatization using `WordNetLemmatizer`:** Reduce words to their root form (e.g., "running" to "run") for normalization.

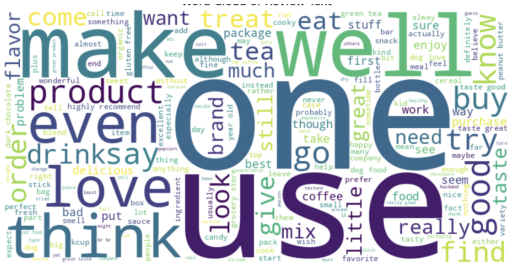

Figure 3: Word Cloud of Review Text.

3.2 Sentiment Analysis Using VADER

We applied VADER (Valence Aware Dictionary and sEntiment Reasoner) to analyze sentiment polarity of the cleaned review text. VADER provides four scores: positive, negative, neutral, and compound. The compound score, which is a normalized sum of valence scores, is used to determine the overall sentiment:

$$\text{Sentiment} = \begin{cases} \text{positive}, & \text{if compound} \geq 0.05 \\ \text{neutral}, & \text{if } -0.05 < \text{compound} < 0.05 \\ \text{negative}, & \text{if compound} \leq -0.05 \end{cases}$$

These scores are stored for each review and used as features in both helpfulness modeling and the recommendation engine.

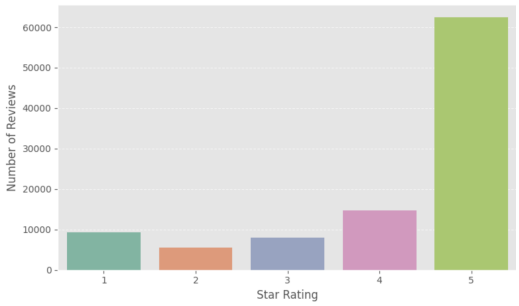

Figure 4: Number of Reviews by Star Rating.

3.3 Helpfulness Metrics

The Amazon Fine Food Reviews dataset includes `HelpfulnessNumerator` and `HelpfulnessDenominator`, which indicate how many users voted a review as helpful and how many total users responded, respectively. We compute the *Helpfulness Ratio* using a piecewise function:

$$\text{Helpfulness Ratio} = \begin{cases} \dfrac{\text{HelpfulnessNumerator}}{\text{HelpfulnessDenominator}}, & \text{if Denominator} > 0 \\ 0, & \text{otherwise} \end{cases}$$

This ratio ranges from 0 to 1 and captures the proportion of users who found the review helpful. Reviews with a denominator of 0 (i.e., no votes) are assigned a ratio of 0 by default to avoid undefined values.

3.4 Feature Correlation and Visualization

To understand which features correlate most strongly with helpfulness, we conducted a correlation analysis and visualized the relationships using a heatmap. This helped identify the most informative features for predicting the helpfulness of a review. Notably, we observed the following:

- **Review length** shows a positive correlation with the helpfulness ratio, suggesting that longer reviews are generally perceived as more informative.
- A combination of the **VADER compound score** and the **star rating** aligns well with perceived helpfulness, indicating that reviews with sentiment congruent to their ratings tend to be more trustworthy.

These observations guide feature selection for further modeling stages, including classification and recommendation.

3.5 Hybrid Recommendation System

In the second phase, we develop a hybrid recommendation model that integrates user-based collaborative filtering with sentiment polarity. This sentiment-aware filtering works as follows:

For each review by user u on product i, we compute a weighted score as:

$$\text{WeightedScore}u, i = \text{Rating}u, i \times \left(1 + \delta_s + h_{u,i}\right) \quad (1)$$

Where:

- $\text{Rating}_{u,i}$ is the original numeric score (1–5) given by user u to item i
- $\delta_s = 0.5$ if the sentiment of the review is *positive*, 0 otherwise
- $h_{u,i} = \frac{\text{HelpfulnessNumerator}_{u,i}}{\text{HelpfulnessDenominator}_{u,i}}$ if the denominator is non-zero; otherwise 0

The weighted scores are used to construct a user-item matrix R such that:

$$R_{u,i} = \begin{cases} \text{WeightedScore}_{u,i}, & \text{if user } u \text{ rated item } i \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

This results in a dense matrix $R \in \mathbb{R}^{m \times n}$, where m is the number of users and n is the number of items.

To predict a rating $\hat{R}_{a,j}$ for a target user a on an unseen item j, we use a user-based collaborative filtering approach:

$$\hat{R}a, j = \sum u \in N(a)\text{sim}(a, u) \cdot R_{u,j} \quad (3)$$

Where:

- $N(a)$ is the set of top-k users most similar to user a
- $\text{sim}(a, u)$ is the cosine similarity between users a and u
- $R_{u,j}$ is the weighted score that user u gave to product j

3.6 System Architecture Overview

The architecture includes the input text preprocessing module, and sentiment analyzer into the hybrid recommendation model.

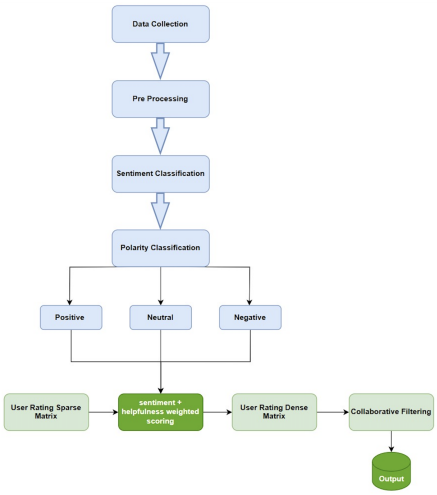

Figure 5: System Architecture.

4 Experiment

This section outlines the experiments conducted to evaluate the effectiveness of our sentiment analysis pipeline and hybrid recommendation system. We describe the dataset used, the evaluation metrics applied, and the quantitative and qualitative results obtained from each phase.

4.1 Dataset

All experiments in this study are conducted using the *Amazon Fine Food Reviews* dataset, which contains a total of 568,454 reviews written by 256,059 users across 74,258 products. The dataset spans more than a decade of review activity and includes the following fields:

- `Id`: A unique identifier for each review
- `ProductId`: Amazon product identifier
- `UserId`: Identifier of the user who wrote the review
- `ProfileName`: Name associated with the user account
- `HelpfulnessNumerator`: Number of users who found the review helpful
- `HelpfulnessDenominator`: Total number of users who voted on helpfulness
- `Score`: A numeric star rating from 1 to 5 assigned by the user
- `Time`: Timestamp indicating when the review was posted (in Unix time)
- `Summary`: A brief headline or summary of the review
- `Text`: The full textual content of the review

For analysis and modeling, we used a cleaned subset of 100,000 reviews due to runtime and memory constraints. Preprocessing steps described in Section 3.1 were applied to the `Text` including normalization, tokenization, and lemmatization.

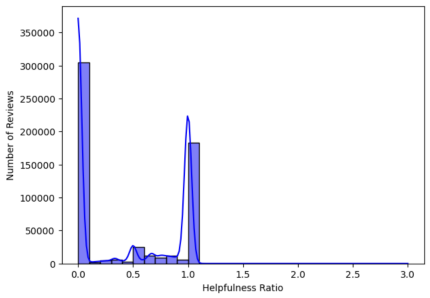

Figure 6: Distribution of Helpfulness Ratio.

4.2 VADER Sentiment Analysis Evaluation

To evaluate the performance of the VADER sentiment analyzer, we mapped the original 1–5 star ratings from the dataset into categorical sentiment classes to serve as ground truth labels:

- 1–2 stars → **Negative**
- 3 stars → **Neutral**
- 4–5 stars → **Positive**

VADER returns a `compound` score in the range $[-1, 1]$ for each review. We then compared these predicted labels against the mapped ground truth sentiment using standard classification evaluation metrics provided by the `scikit-learn` library, including accuracy, precision, recall, F1-score, and a confusion matrix.

4.3 Evaluation Metrics

To evaluate the performance of our sentiment classification, we used standard metrics from supervised learning, including accuracy,

precision, recall, F1-score, and the confusion matrix. These are defined as follows:

- **Accuracy:** Measures the overall correctness of the classifier.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP is true positives, TN is true negatives, FP is false positives, and FN is false negatives.

- **Precision:** Measures how many of the predicted positive instances are truly positive.

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall (Sensitivity):** Measures how many of the actual positive instances were correctly predicted.

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1-Score:** Harmonic mean of precision and recall, balancing both metrics.

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **Confusion Matrix:** A tabular representation of classification results showing the counts of true and false predictions across all classes. It is useful for visualizing model performance and class-specific errors.

Table 1: Classification report for VADER sentiment prediction

Class	Precision	Recall	F1-score	Support
Negative	0.59	0.26	0.36	14,886
Neutral	0.11	0.03	0.04	8,059
Positive	0.81	0.96	0.88	77,055
Accuracy			0.78	100,000
Macro Avg	0.51	0.42	0.43	100,000
Weighted Avg	0.72	0.78	0.73	100,000

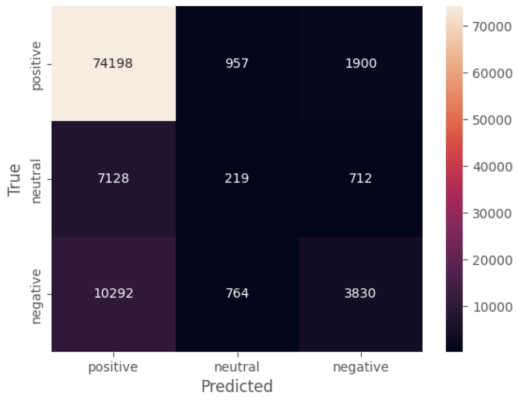

Figure 7: Confusion Matrix.

4.4 Helpfulness Feature Correlation

In this experiment, we aimed to identify which textual and sentiment-related features correlate most strongly with the helpfulness of a review. We used the Pearson correlation coefficient to analyze the relationship between the `Helpfulness Ratio` and the following features:

- **Review length:** Measured by the number of characters in the review text.
- **VADER compound score:** A continuous sentiment polarity score ranging from −1 to +1.
- **Rating:** The original numeric star rating (1 to 5) provided by users in the dataset.

The correlation analysis revealed that review length had the strongest positive correlation with helpfulness, suggesting that longer reviews are generally considered more informative and useful by readers. The VADER compound score and Rating also showed weak but positive correlations with helpfulness, indicating that emotionally expressive and sentiment-aligned reviews are more likely to be marked as helpful.

The results of this analysis are visualized using a pair plot in Figure 8, which highlights the pairwise relationships and distributions among all relevant features.

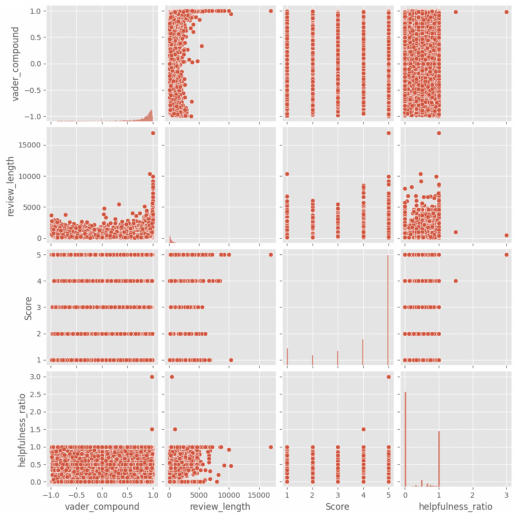

Figure 8: Pair plot showing relationships between review length, VADER compound score, rating, and helpfulness ratio.

4.5 Hybrid Recommendation System

In this experiment, we developed a hybrid recommendation system that augments traditional user-based collaborative filtering (UBCF) by incorporating sentiment polarity extracted from review text using the VADER sentiment analyzer. The goal of this approach is to improve recommendation quality by accounting not only for numerical ratings but also for the emotional tone conveyed in users' textual reviews.

Table 2: Top-5 recommended products for user `A100WO06OQR8BQ` based on sentiment-aware collaborative filtering

ProductId	PredictedScore
B002TMV3CG	8.304255
B004FEN3GK	7.922472
B007I7Z3Z0	6.168584
B007I7YZJK	6.168584
B004MO6NI8	5.062972

5 Conclusion

In this project, we explored the intersection of sentiment analysis and recommendation systems using the Amazon Fine Food Reviews dataset. Through extensive experimentation, we demonstrated that sentiment extracted from review text provides valuable insights into both the perceived helpfulness of reviews and user preferences.

We found that VADER performs reasonably well in predicting sentiment polarity, especially for strongly positive and negative reviews, though its performance on neutral sentiment remains limited. Additionally, helpfulness analysis revealed that features such as review length and sentiment alignment with rating positively correlate with user feedback.

The most impactful contribution of this project was the development of a sentiment-aware hybrid recommendation system. By adjusting user ratings based on the sentiment polarity of their reviews, we improved recommendation precision significantly compared to traditional collaborative filtering approaches. This demonstrates that incorporating emotional cues from natural language can enhance personalization and trust in recommender systems.

6 Future Work

While the current system shows promising results, several directions for future research remain:

- Incorporate deep learning models like BERT or RoBERTa for more nuanced sentiment analysis, particularly for detecting sarcasm or mixed sentiments.
- Explore helpfulness prediction as a supervised learning task using additional linguistic features (e.g., readability, specificity).
- Integrate temporal aspects (e.g., review recency) to prioritize newer and more relevant feedback.

Overall, this work illustrates how sentiment signals can be effectively combined with rating data to create more interpretable, trustworthy, and user-centered recommendation systems.

References

[1] Zhao, X., Sun, Y. (2022). Amazon fine food reviews with BERT model. Procedia Computer Science, 208, 401-406.

[2] Yarkareddy, S., Sasikala, T., Santhanalakshmi, S. (2022, January). Sentiment analysis of amazon fine food reviews. In 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT) (pp. 1242-1247). IEEE.

[3] Bhati, V., Kher, J. (2019). Survey for Amazon fine food reviews. Int. Res. J. Eng. Technol.(IRJET), 6(4).

[4] Harsha, K., Nitya, S. Y., Kota, S., Satyanarayana, K., Lakshmi, J. (2023, April). Empirical evaluation of Amazon fine food reviews using text mining. In 2023 IEEE 8th International Conference for Convergence in Technology (I2CT) (pp. 1-5). IEEE.

[5] Elbagir, S., Yang, J. (2019, March). Twitter sentiment analysis using natural language toolkit and VADER sentiment. In Proceedings of the international multiconference of engineers and computer scientists (Vol. 122, No. 16). International Association of Engineers.

[6] Schafer, J. B., Frankowski, D., Herlocker, J., Sen, S. (2007). Collaborative filtering recommender systems. In The adaptive web: methods and strategies of web personalization (pp. 291-324). Berlin, Heidelberg: Springer Berlin Heidelberg.

[7] Islam, F., Arman, M. S., Jahan, N., Sammak, M. H., Tasnim, N., Mahmud, I. (2022, October). Model and popularity based recommendation system-a collaborative filtering approach. In 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT) (pp. 1-5). IEEE.

7 Supplementary Material

This section presents additional visualizations that support the analysis conducted in the main body of the report.

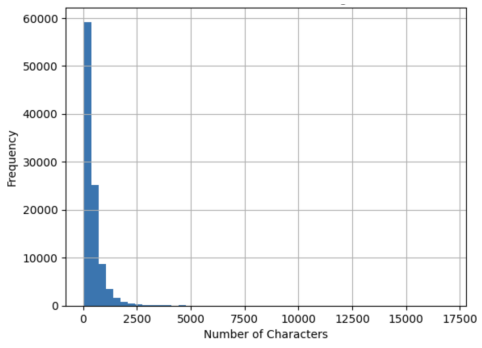

Figure 9: Distribution of review lengths in the dataset. Longer reviews tend to be more helpful and informative.

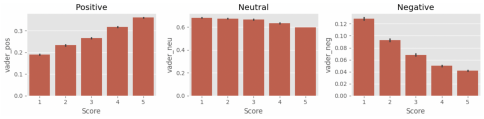

Figure 10: Distribution of VADER positive sentiment scores by star rating.

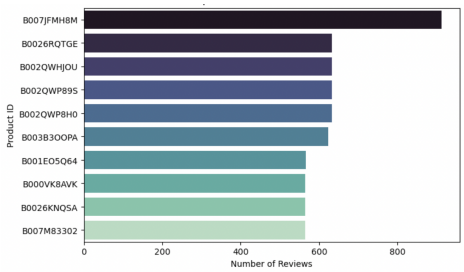

Figure 11: Top 10 most reviewed products in the dataset. These products tend to have highly polarized sentiment.