QTM 340 Project Report

Winci Liang, Angela Liu , Louis Lu

Abstract

This study analyzes Amazon Beauty product reviews to identify characteristics of helpful reviews using machine learning. The dataset comprises 370,000 reviews, examined through techniques like term frequency-inverse document frequency (TF-IDF) and Latent Dirichlet Allocation (LDA). A range of classifiers including Multi-Layer Perceptron (MLP) and Random Forest are employed with static and contextual embeddings to predict review helpfulness. Findings reveal that detailed product discussions correlate with helpfulness, rather than mere positive sentiments. The Random Forest model with contextual embeddings shows the best predictive performance. Results offer insights into e-commerce consumer behavior, suggesting further research opportunities in model optimization and interpretability in review analysis.

1 Introduction

Online product reviews are a critical aspect of e-commerce platforms, influencing consumer behavior and decision-making and they also act as a valuable source of feedback for businesses. Understanding the nuances of these reviews, especially what makes them "helpful," is crucial for both consumers and sellers. This study focuses on Amazon product reviews from the Beauty category, exploring how reviews are perceived and what characteristics make a review helpful. Besides the motivation of helping businesses to understand the dynamics of product reviews for better marketing strategies and for consumers to effectively communicate their experiences and opinions, we hope to explore previous findings on how machine learning can be applied to online shopping reviews and compare the popular models to assess their performance. After the comparisons, meaningful error analyses will provide us with possible explanations for each models' shortcomings.

2 Related work

Previous research has primarily focused on sentiment analysis using machine learning techniques on Amazon review datasets. Rain (2013) employed Naive Bayes and decision list classifiers to tag reviews as positive or negative, using star ratings as training data. Rathor, Agarwal, and Dimri (2018) furthered this approach by evaluating the effectiveness of Support Vector Machines (SVM), Naive Bayes, and Maximum Entropy classifiers in sentiment classification. Haque, Saber, and Shah (2018) and Dey et al. (2020) highlighted the superior performance of Linear Support Vector Machine and SVM, respectively, in polarizing Amazon product reviews. Hawlader et al. (2021) expanded this research by applying various machine learning techniques, including the Multi-Layer Perceptron (MLP), and comparing preprocessing techniques like TF-IDF, Bag of Words, and Word2Vec. The study conducted by Kim et al. that builds a research paper classification system that can categorize research papers into meaningful classes based on subjects, the researchers used K-means clustering to classify the entire dataset of papers into y subject similarity based on the TF-IDF values of each paper (Kim et al., 2019). Furthermore, in a study by Zhou, he used TF-IDF to remove features with low weights and to extract key features and the corresponding word vector through the Word2Vec, and then input it into a convolutional neural networks model (CNN). His purpose of utilizing TF-IDF before inputting data into the neural network model is to reduce model parameters and training time. However, Zhou also notes that the model will lose a certain amount of accuracy in short texts if some features are removed by TF-IDF (Zhou, 2022).Our research leverages insights from previous studies, particularly those by Haque et al. (2018) and Hawlader et al. (2021), which identified SVM and MLP as effective clas-

sifiers in sentiment analysis. We aim to apply and compare these methodologies, along with other useful methodologies such as Logistic Regression, Random Forest, TF-IDF to Amazon product reviews from functionally distinct categories, enhancing machine learning literature with insights into the popularity of classifiers and resilience of established supervised research models. This study contributes to understanding how reviews are valued across e-commerce platforms, influencing consumer and business decision-making.

3 Data

We accessed and downloaded the Amazon reviews dataset compiled by Ni, Li, and McAuley (2019), opting for the "all beauty" category due to its manageable size (with 370k reviews) for practical analysis compared to other categories. The dataset features one-review-per-line in JSON format, encompassing essential details such as the unique reviewer identifier (”reviewerID”), the product identifier (”asin”), reviewer name (”reviewerName”), number of helpful votes (”vote”), and the overall product rating (”overall”). Additionally, comprehensive metadata, including style details such as size and color, is provided in a dictionary format (”style”). Each review entry contains the reviewer’s detailed opinions (”reviewText”), with a concise summary or title (”summary”). Temporal aspects include Unix timestamp format (”unixReviewTime”) and raw format (”reviewTime”). Users can augment their reviews with images (”image”), represented as URLs, providing a visual component.

Analyzing the rating distribution (Figure 1), we observe that 61.82% of reviews have a 5.0 rating, followed by 15.15% with a 4.0 rating, 14.54% with a 1.0 rating, 5.90% with a 3.0 rating, and 4.14% with a 2.0 rating. Plotting ratings against the number of votes (Figure 2) reveals that 5.0 reviews attract some of the reviews with the highest number of helpful votes and span a broad range in terms of helpful votes. 1.0 reviews show a similar but smaller range with a few outliers of exceptionally helpful reviews, followed by 4.0, 2.0, and 3.0 ratings. This pattern suggests that extremely positive or negative reviews might offer more insightful content, potentially due to their detailed descriptions and engaging writing styles, which would be explored further in the Exploratory Data Analysis section.

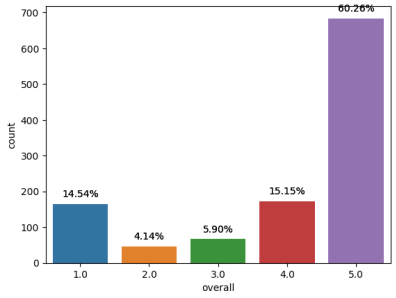


Figure 1: Distribution of Overall Ratings in Beauty Product Reviews

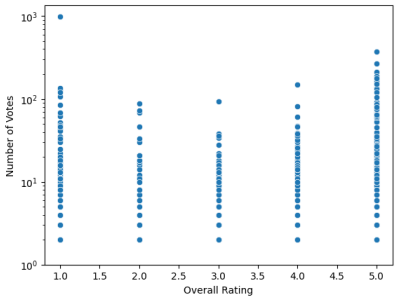


Figure 2: Relationship between Helpful Votes and Overall Ratings for Beauty Products

4 Exploratory Data Analysis

In the exploratory data analysis, we examined the characteristics of words and phrases used in helpful reviews, compared to unhelpful reviews. Here, helpfulness is classified as whether a review falls within the 90th percentile of votes, surpassing 30 “helpful” votes. Specifically, we scrutinized TF-IDF scores to identify the most frequently used meaningful words and employed Latent Dirichlet Allocation (LDA) to unveil prevalent topics within this category.

4.1 Word and Phrase Frequency Analysis

We conducted a comparative analysis of term frequency-inverse document frequency (TF-IDF) scores in helpful and unhelpful review corpuses. This involved utilizing separate instances of TfidfVectorizer for each category, configured to consider a maximum of 1000 features and exclude English stop words. The vectorizers transformed the reviews into TF-IDF matrices. For helpful reviews, the code extracted feature names, calculated IDF (Inverse Document Frequency) weights, and organized the results in a DataFrame. The same process was repeated for unhelpful reviews. Both DataFrames were then sorted based on IDF weights, providing insights into the most influen-

tial terms in each category. We discovered that although there are some overlaps in the terms with the highest IDF weights between the two corpora, such as “hair” and “product,” the unhelpful reviews exhibit more sentiments, as noted in Table 1. Words such as “love,” “great,” and “like”are highly frequent in unhelpful reviews, but they do not appear prominently in helpful reviews. This suggests that expressing strong positive sentiments alone is often insufficient for a review to be deemed "helpful" or to garner numerous "found helpful" feedback in the Amazon beauty product review section. Rather, helpful reviews tend to have more detailed discussions of the products.

Table 1: Top Terms in Helpful and Unhelpful Reviews

Helpful Review Terms	**idf_weights**
hair	0.065259
product	0.046645
color	0.041413
just	0.037051
skin	0.036904

Unhelpful Review Terms	**idf_weights**
hair	0.040323
love	0.040056
product	0.038671
great	0.035074
like	0.034670

4.2 Topic Modeling

To implement LDA, we first tokenized the review text, removed stopwords, and filtered out non-alphabetic characters using the nltk library. We also created a dictionary in gensim to convert a stream of tokens to a stream of ids and, subsequently, a corpus in the format that gensim expects. We then run the LDA Multicore model for both corpuses to get ten topics, with the number of passes over the data set to 50 and the number of iterations set to 500.

The differences in topic distributions between the two corpuses indicate variations in the predominant themes and interests expressed by the reviewers in each dataset. The unhelpful review corpus encompasses a broader spectrum of beauty and grooming topics, while the helpful review corpus appears to have a more concentrated emphasis on hair-related discussions.

Both types of reviews revolve around positive product experiences, with terms like ’love,’ ’nice,’ ’great,’ and ’good’ across various topics found in both corpuses. However, the differences in topic distributions between the two corpuses indicate variations in the product types and aspects of experiences. The unhelpful review corpus encompasses a broader spectrum of beauty and grooming topics with general descriptions of experiences, while the helpful review corpus appears to have a more concentrated emphasis on hair-related discussions with respect to duration of products and the specific uses.

In the unhelpful review corpus, topics are characterized by specific beauty and grooming-related terms such as ’razor,’ ’stand,’ ’brush,’ ’blades,’ ’soap,’ ’polish,’ ’wig,’ and ’serum.’ The associated weightings suggest a focus on diverse aspects of personal care and beauty products.

On the other hand, the helpful review corpus exhibits topics revolving around hair care, with prominent terms like ’hair,’ ’color,’ ’product,’ ’day,’ ’dry,’ ’wear,’ ’weeks,’ ’long,’ ’black,’ ’natural,’ and ’smell.’ These weightings emphasize discussions related to hair health, color, and maintenance, with attention to duration and the use of specific products.

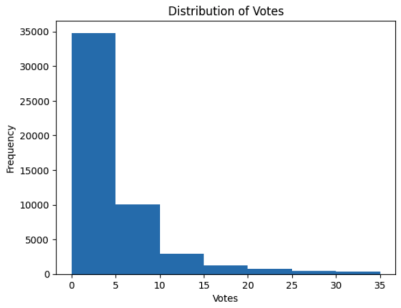
5 Helpful Review Predictions

5.1 Problem Overview

To address the challenge of distinguishing helpful from unhelpful product reviews based on textual features, we conceptualize this task as a binary classification problem. We define a threshold to separate helpful and unhelpful reviews, employing the number of upvotes as a criterion. Specifically, we use the 90th percentile of upvotes as our threshold. As illustrated in Figure 3, the distribution of votes is right-skewed, characterized by a predominance of reviews with few votes and a smaller fraction receiving a high number of votes. This skew reflects the reality that most reviews are not deemed helpful and are rarely referenced.

However, utilizing the 90th percentile as the threshold yields an imbalanced distribution of labels: approximately 10 percent of the reviews are categorized as positive (helpful), and 90 percent as negative (unhelpful). To rectify this imbalance and prevent model bias towards the majority class, we adopted a stratified sampling strategy that ensures a 50-50 split of positive and negative classes. Ad-

ditionally, we applied a stratified train-test split to maintain this equal representation in both our training and testing datasets. This balanced approach is crucial for effective classifier training and for evaluating the model's performance in a manner that is representative of real-world scenarios.



5.2 Methods

We adopted a structured approach to address the problem, starting from simpler models and advancing to more complex ones. Initially, we employed basic linear classifiers, namely Logistic Regression and Support Vector Classifier (SVC), leveraging fundamental text features such as Bag of Words and TFIDF (Term Frequency-Inverse Document Frequency).

Further, to potentially enhance performance, we integrated static embeddings in conjunction with a Multi-Layer Perceptron (MLP). This combination aimed to capture more nuanced patterns in the data that basic linear models might miss.

In our pursuit of refinement, we experimented with various pooling methods, including Max Pooling. We also innovatively applied TFIDF weighting as a pooling strategy, hoping to leverage the inherent significance of specific terms in the text.

Finally, we incorporated contextual embeddings derived from the BERT (Bidirectional Encoder Representations from Transformers) model. This step was intended to harness the advanced capabilities of BERT in understanding contextual nuances in text, thereby potentially boosting the accuracy and robustness of our classification model. In addition, we tried the ensemble methods like gradient boosting and bagging to see if that could help improve the prediction performance.

5.3 Results

Our evaluation of various models and feature sets for classifying product reviews (Table 1) reveals significant performance differences. The Multi-Layer Perceptron (MLP) models using TFIDF and Bag of Words feature sets underperformed, both achieving an F1 score of 0.4972 and an AUC of 0.0000. In contrast, models employing static embeddings demonstrated better performance, with Logistic Regression achieving the highest F1 score of 0.7166 and an AUC of 0.5987 among these. The use of contextual embeddings further enhanced model performance, with the Random Forest classifier achieving the best overall results—an F1 score of 0.7333 and an AUC of 0.7575. This indicates the effectiveness of contextual embeddings over other feature sets in this classification task.

5.4 Error Analysis & Reflection

- TFIDF

  From the result section, we can see TFIDF, no matter used as features or pooling weights, doesn't offer comparable performance with the best classifier. Since it benefits from a corpus with a wide variation in word usage across documents. In short product reviews within the same genre, the vocabulary used may be very similar, leading to less discriminative TFIDF scores. In product reviews for similar items, common words might actually carry significant meaning. For example, common words found by our EDA, such as "like", "love", and "effective", could be very informative, but TF-IDF will downplay these if they appear across many reviews.

- Static vs Contextual Embeddings:

  Contextual embeddings are designed to capture the meaning of words in context. They are especially powerful in understanding nuanced language and distinguish words that have different meanings in different sentences. However, in short product reviews (in the same genre), there may not be enough context for these models to significantly outperform static embeddings. The short nature of the text might provide limited contextual clues for the model to leverage, potentially leading to marginal gains or even misinterpretations in comparison to static embeddings.

  In addition, contextual embeddings models are generally more computationally expensive to train and use for inference than static models. Since the static embeddings can offer comparable performance, we can consider which one to use based on the the computational cost and

Model	Feature Set	F1 Score	AUC
MLP	TFIDF	0.4972	0.0000
MLP	Bag of Words	0.4972	0.0000
SVC	Static Embeddings	0.6997	0.2007
Logistic Regression	Static Embeddings	0.7166	0.5987
MLP	Static Embeddings	0.6134	0.5930
MLP	Static Embeddings (TFIDF Poolling)	0.5868	0.6466
MLP	Contextual Embeddings	0.6657	0.6667
Random Forest	Contextual Embeddings	0.7333	0.7575
LightGBM	Contextual Embeddings	0.6322	0.6866

Table 2: Comparison of model performance

5.5 Limitations

- Loss of Interpretability

  Unlike simpler bag-of-words models, where the influence of each term on the classification is clear, embeddings encode semantic information across many dimensions, obscuring the contribution of individual words. This makes it challenging to discern specific textual features that signify helpfulness in product reviews, limiting our ability to extract actionable insights.

5.6 Performance Comment

While the Random Forest classifier with contextual embeddings demonstrated the best performance among the tested models, it's important to note that the overall effectiveness remains moderate, as suggested by the F1 score and AUC values. Despite these limitations, this model still provides a useful benchmark for predicting helpful reviews. Its relative success in classification can serve as a foundational reference for further model development and refinement

6 Conclusion

In conclusion, this study has provided valuable insights into the dynamics of Amazon's Beauty product reviews, emphasizing the importance of understanding what constitutes a "helpful" review. Through comprehensive analyses utilizing various machine learning techniques, we've explored the intricate relationship between review content, consumer perception, and their helpfulness.

Our exploration of term frequencies and topic modeling revealed that helpful reviews often contain more detailed and specific discussions about the product, rather than just expressing positive sentiments. This distinction is critical for businesses and consumers as it underscores the importance of detailed, informative reviews in influencing consumer decisions and providing actionable feedback to businesses.

In the realm of predictive modeling, our findings highlight the potential of machine learning in classifying helpful reviews. While the Random Forest model with contextual embeddings emerged as the most effective in our tests, the moderate F1 score and AUC values indicate there is room for improvement. This finding allows room of further research, particularly in refining these models or exploring new methodologies to enhance predictive accuracy.

Our study has several limitation and should be taken into consideration; for exampel The loss of interpretability with complex models and the challenge of balancing computational costs with performance. Future work could focus on improving the interpretability of these models or finding a more efficient balance between computational resources and predictive power.

Overall, this research contributes to a deeper understanding of online consumer behavior and the evolving field of e-commerce as it becomes an indispensable part of people's lives especially after the pandemic. It not only aids businesses in leveraging customer feedback more effectively but also empowers consumers to make more informed decisions based on the quality of product reviews. As e-commerce and the power of machine learning, especially NLP continue to grow, we expect to see further investigation into this topic by other researchers.

References

[1] Dey, S., Wasif, S., Tonmoy, D. S., Sultana, S., Sarkar, J., & Dey, M. (2020, February). A comparative study of support vector machine and Naive Bayes classifier for sentiment analysis on Amazon product reviews. In *2020 International Conference on Contemporary Computing and Applications (IC3A)* (pp. 217-220). IEEE.

[2] Haque, T. U., Saber, N. N., & Shah, F. M. (2018, May). Sentiment analysis on large scale Amazon product reviews. In *2018 IEEE International Conference on Innovative Research and Development (ICIRD)* (pp. 1-6). IEEE.

[3] Hawlader, M., Ghosh, A., Raad, Z. K., Chowdhury, W. A., Shehan, M. S. H., & Ashraf, F. B. (2021, September). Amazon product reviews: Sentiment analysis using supervised learning algorithms. In *2021 International Conference on Electronics, Communications and Information Technology (ICECIT)* (pp. 1-6). IEEE.

[4] Kim, SW., & Gil, JM. (2019). Research paper classification systems based on TF-IDF and LDA schemes. *Human-Centered Computing and Information Sciences*, 9, 30. https://doi.org/10.1186/s13673-019-0192-7

[5] Meyer-Baese, A., & Schmid, V. (2014). Foundations of Neural Networks. In *Pattern Recognition and Signal Analysis in Medical Imaging* (pp. 197–243). https://doi.org/10.1016/b978-0-12-409545-8.00007-8

[6] Ni, J., Li, J., & McAuley, J. (2019). Empirical Methods in Natural Language Processing (EMNLP), pdf.

[7] Rathor, A. S., Agarwal, A., & Dimri, P. (2018). Comparative study of machine learning approaches for Amazon reviews. *Procedia Computer Science*, 132, 1552-1561.

[8] Rain, C. (2013). Sentiment analysis in amazon reviews using probabilistic machine learning. Swarthmore College, 42.

[9] Shargabi, B., et al. (2019). Enhancing multi-layer perceptron for breast cancer prediction. *International Journal of Advanced Science and Technology*, 130, 11–20. https://doi.org/10.33832/ijast.2019.130.02

[10] Tamouridou, A., et al. (2018). Spectral identification of disease in weeds using multilayer perceptron with automatic relevance determination. *Sensors*, 18(9), 2770. https://doi.org/10.3390/s18092770

[11] Zhou, H. (2022). Research of text classification based on TF-IDF and CNN-LSTM. *Journal of Physics: Conference Series*, 2171(1), 012021. https://doi.org/10.1088/1742-6596/2171/1/012021