**Sentiment Analysis of Amazon Product reviews**

**Introduction**

The significance of social media and web for a business organization is evolving day by day, pricing strategies play a pivotal role in shaping the consumer behavior and their satisfaction. Among these strategies, the idea of providing discounts stands apart as a significant factor which influences the customer purchasing decisions and perceptions of value. As the top leading online retail platform, Amazon offers a unique set of ideas to explore how discounts level impact customer satisfaction, given its huge market reach and diverse offering of products. This report aims to delve into this relationship by providing insights which could help businesses optimize their pricing strategies to enhance consumer satisfaction and loyalty.

The data set for this research was meticulously gathered using a premium web scrapping tool known as octoparse which has over 1400 entries of products, having various attributes listed such as customer ratings and review text etc. These elements are very crucial for analyzing the shades of how pricing adjustments, discounts affect customer responses and satisfaction levels.

The analysis done in this report utilized a combination of qualitative and has a combination of qualitative and quantitative methods. Quantitatively, statistical techniques will be applied to assess the correlation between the discount levels and the customer ratings is provided, which ensures that is a clear and numeric depiction of their relationship. Qualitatively, sentiment analysis will be performed on customer reviews to evaluate the emotional tone and content of consumer feedback about the products and its price. This dual approach will allow for a comprehensive understanding of both the Direct impact of discounts on the customer ratings and the understanding of the influence made on the customer perception and satisfaction which are expressed through the reviews.

The study seeks to offer various actionable insights into effective discounting strategies on Amazon, which potentially guides e-commerce businesses in making data driven decisions to improve the customer satisfaction and competitive advantage.

**Literature review**

Sentiment analysis has become of focus of research due to its ability to transform vast amount of customer data into useful actionable insights. The studies by(Huang et al., 2023) and (Bayhaqy et al., n.d.) Highlighted machine learning techniques such as decision trees, K-nearest neighbours, and naive Bayes are effective in classifying sentiments from social media platforms. These techniques are adaptable to various data sources including Twitter and e-commerce reviews. The functionality of sentiment analysis goes beyond product reviews, techniques to analyse service quality in e-commerce, implying sentiment analysis can significantly impact customer satisfaction and retention strategies.(Bayhaqy et al., n.d.; Singh et al., 2022; Wankhade et al., 2022). Emerging trends in sentiment analysis research focus on deep learning's role in enhancing the accuracy and depth of sentiment detection. an innovative approach to sentiment analysis by integrating rule-based classification with machine learning techniques, including Support Vector Machines and statistical methods were introduced by (Prabowo & Thelwall, 2009) where they tested the datasets like movie reviews and social media comments.

Exploring the evolution from simple machine learning techniques to more sophisticated hybrid models as various methods of feature extraction and classification used to analyse the customer reviews which shows the importance of sentiment analysis in interpreting consumer behaviour and feelings towards products purchased online, highlighting key challenges and trends in the field.(Nur Amir Sjaif, n.d.)(Medhat et al., 2014). Further advancing the field,(Dey et al., 2020) compare Support Vector Machines (SVM) and Naive Bayes for analysing the reviews from amazon which explains that SVM can do polarity detection in sentiment with higher accuracy.(Rajput, 2019) investigates the integration of Natural Language Processing (NLP) and sentiment analysis into clinical analytics, highlighting the transformative impact of Big Data on healthcare research. The approaches to parse and analyze complicated data from many sources, such as social media, contributing to breakthroughs in psychological and psychiatric evaluations were discussed. Utilization of unsupervised machine learning approach for sentiment analysis using various models like Latent Dirichlet Allocation (LDA) technique IMBD datasets were used to discern topics in user reviews following TDA model model to determine sentiment polarity across topic. LDAvis was utilized to visualize data in this study. The field of sentiment analysis is transforming from traditional machine learning approaches to more advanced deep learning and hybrid models that provide more precision and from traditional machine learning approaches to more advanced deep learning and hybrid models that provide more precision and contacts contextual knowledge. Integrating these technologies into platforms helps improve decision making processes, customer relationship management and drives more success in the business.

**Methodology**

*Data Description*

The dataset was scraped by a web scraping tool known as Octoparse where it is a premium automated tool which only needs a URL of the web source from which we need data to be collected for the scrapping. The data set is scrapped from amazon India containing about more than 1400 product reviews with the ratings of each product. The data set comprises customer reviews from Amazon which covers various types of product categories. Each entry includes fields such as product ID, product name, category of the product, discounted price, actual price, rating count, about product, user ID, username, review ID, rating of the product, review title and review content with an addition of image link and product link. This diverse data set provides a comprehensive view of customer feedback across different product types.

*Data Cleaning and Preparation*

Initially the dataset was imported into the Jupyter notebook and before starting the process of analysis it was necessary that the data was cleaned and prepared, so first the missing values in the data sets were handled, entries missing critical information such as review content or ratings were excluded to ensure the data integrity. Then the text pre-processing was performed as the review text were processed to remove punctuations, converted to lower case, eliminate stop words, and tokenize the content, preparing them for the natural language processing applications. The unnecessary columns from the dataset were dropped for attaining a clear structure and organized data during both quantitative and qualitative analysis.

A new discounted percentage field was created from the actual price and the discounted price for visualizing a better relation between the ratings and the price. The product category field was split into main category and subcategory to facilitate more broad analysis, and this allowed for a detailed examination of consumer feedback and purchasing pattern with specific product categories. An idea of creating a correlation between the discounts provided for the products and the rating of the product was indulged. As the analysis of categories of products could enlighten the relationship between the discounts and the ratings which helps to understand the consumer sentiment.

The normalization and scaling were done where necessary, features like the 'rating count' and price fields where normalized to bring them onto a similar scale comma especially useful for any machine learning models that might be sensitive to scale.

*Descriptive analysis*

The report begins with a descriptive statistical analysis which helps to understand the central tendency and variability of key numeric features of the discounted price, actual price, discount percentage and rating count. These statistics provides a quantitative overview of the data helping identify ranges and common values which are crucial for interpreting the rest of your analysis.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **discounted price** | **actual price** | **discount percentage** | **rating count** |
| **count** | 1465.000000 | 1465.000000 | 1465.000000 | 1465.000000 |
| **mean** | 3125.310874 | 5444.990635 | 47.684924 | 18270.564505 |
| **std** | 6944.304394 | 10874.826864 | 21.636267 | 42729.995315 |
| **Min** | 39.000000 | 39.000000 | 0.000000 | 0.000000 |
| **25%** | 325.000000 | 800.000000 | 32.001280 | 1173.000000 |
| **50%** | 799.000000 | 1650.000000 | 50.016672 | 5178.000000 |
| **75%** | 1999.000000 | 4295.000000 | 62.885714 | 17325.000000 |
| **max** | 77990.000000 | 139900.000000 | 94.118824 | 426973.000000 |

*Table 1: descriptive statistical analysis of product prices*

This summary which provides the comprehensive view of the pricing, discounting, and customer engagement metrics across the products available in the Amazon. Using this that distributions can be visualized to have a better understanding of how these variables are related.

*Discount Percentage Distribution*: Visualization of discount percentage was performed with the help of histogram; this helps in understanding the relationship between the discounts which are spread across different products. This could provide a correlation between the customer ratings and the typical discounting practices or promotional strategies employed across different categories.

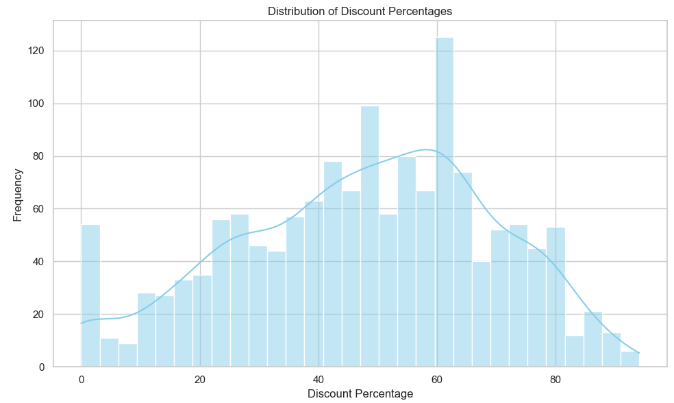


Fig 1: Distribution of Discount Percentage

The distribution in the histogram to have a peak around 60% which suggest a common discount level and as well as there are significant occurrences of discounts around 30 to 70% with some variations outside the range. This provides a basic information for further investigation into how these discount levels might influence customer satisfaction and their sentiment.

A product category-based analysis the performed which identifies the most engaging categories for the consumers. By presenting the top categories in the bar chart there is a visual representation of which category attract the most reviews, suggesting higher consumer traffic or satisfaction. At the same time highlights which categories are most heavily discounted which possibly correlates the higher discounts with higher consumer interest. Also, the subcategories are identified by ratings whichever receives the highest total, this indicates which subcategories are most favoured or valued by consumers which potentially guides the inventory or marketing strategies.

Sentiment Analysis involved the utilization of valence aware dictionary and sentiment reasoner (VADER) tool, it is a lexicon and rule-based sentiment analysis framework specifically attuned to sentiments expressed in social media and web (IEEE Malaysia Section. Electron Devices Chapter et al., n.d.). VADER uses a combination of a sentiment lexicon, which is a list of lexical features as they are generally labelled according to their semantic orientation as either positive or negative. VADER not just takes the word into account, but it also considers and understands the sentiment expressed in the context words. Each review was scored to categorize the sentiment as positive, negative, or neutral based on the common score generator by the VADER. A compound score threshold of 0 was used here where the scores above were considered as positive and scores below zero were considered as negative and the scores equal to zero are equal to zero are classified as neutral. The polarity scores of the sentiment results were calculated and the sentiment category was determined either positive or negative.

Following sentiment analysis, LDA (latent Dirichlet allocation) was performed for topic modelling (Shanghai hai yang da xue & Institute of Electrical and Electronics Engineers, n.d.). LDA is a form of unsupervised mission learning which identifies the topic present in the text where each document is considered a mixture of various topics and each topic as a collection of dominant words. The model was trained on the pre-processed review text. The number of topics was set based on model coherence scores to ensure the meaningful topic segmentation. The prevalent term from every topic were analysed to infer the thematic content which provides the insights into common themes across the customer reviews.(Sari et al., 2018)

**Analysis and Results**

Prior performing the sentiment analysis, the categories of the product which has generated more traffic are identified and grouped so that during the analysis the sentiment can be precisely assessed.

The category of the product has been divided into subcategories which gives a more organized collection of data. The visualization of the product categories which gives the understanding of current market conditions as well as empowers the decision-making strategy. It provides a legitimate view of where the company should focus its efforts to optimize returns, improve customer satisfaction and enhance their competitiveness in the market.

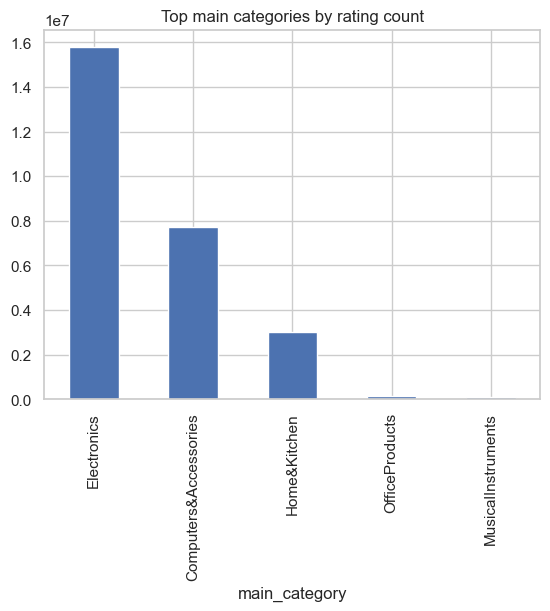
Firstly, to understand the relationship between the customer to understand the relationship between the product ratings and the discount provided, the main categories were grouped and visualized to determine the category with the most reviewed.

**A screenshot of a computer program

Description automatically generated**

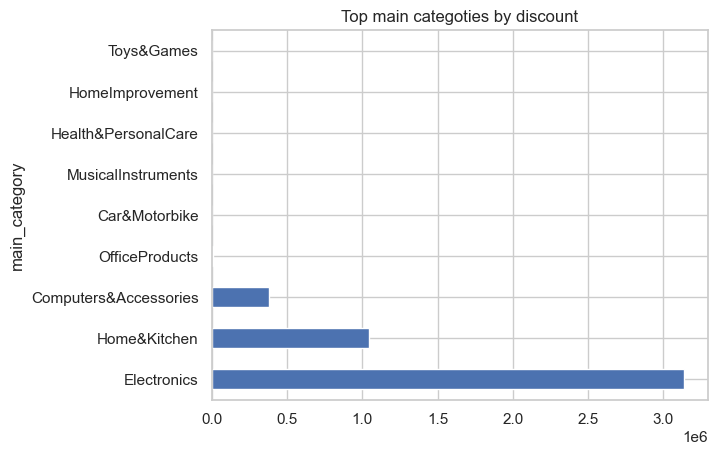
*Figure 2: Top main categories in reviews*

Electronics and computer & accessories and home & kitchen emerged as the most reviewed categories which suggest that there is a high customer engagement in these categories, and they are the potential areas to be targeted as marketing strategies.



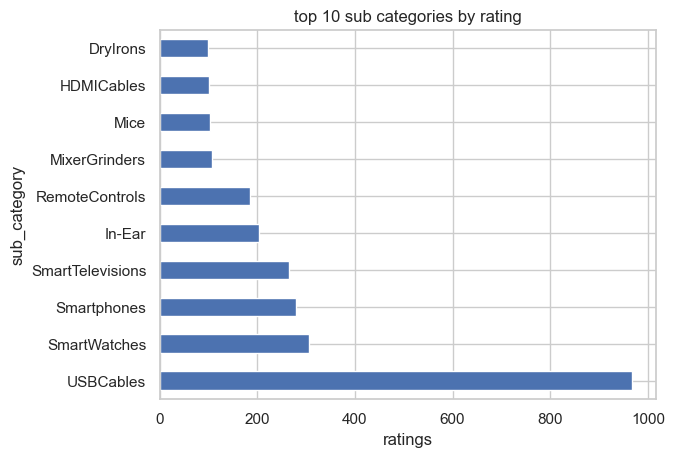
*Figure 3: Top main categories by rating count*

Electronics also had the most discounts which are much more than other categories indicating aggressive pricing strategies likely aimed at boosting sales volumes.



*Figure 4: Top main categories by discount*

The detailed analysis of top sub-categories by rating provides essential insights that will be later cross-referenced with the Latent Dirichlet Allocation (LDA) model results to enhance our understanding of customer feedback themes. This approach ensures a strong examination of how specific product sub-categories might influence or reflect the thematic content emerging from the LDA topics. It is plausible that the themes identified in the LDA analysis could align closely with the sub-categories that have gathered significant customer attention and high ratings. Such correlations, if found, would validate the relevance of LDA topics to actual customer experiences and preferences, enabling more precise strategic recommendations for product and service improvements.



*Figure 5: Top 10 subcategories by rating*

*Sentiment Analysis*

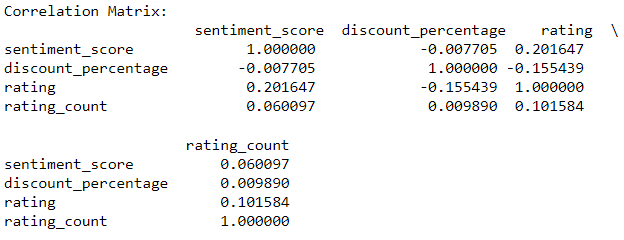
Sentiment analysis using the VADER tool calculated the emotional content of reviews, classifying them into positive, negative, and neutral sentiments. As there were only fewer number of neutral reviews they were dropped. clearly showing the overwhelming prevalence of positive sentiments compared to negative ones. This suggests that the products are generally well-received, with customers often highlighting satisfactory experiences.

*A graph of a number of negative numbers

Description automatically generated with medium confidence*

*Figure 6: sentiment data distribution*

The correlation matrix explored relationships between key variables, the statistical tests were conducted to explore the correlations between the sentiments scores, discount levels and customer ratings. Mild correlation was found between positive sentiment and high ratings which suggests that happier customers generally leave better ratings and the correlation between discount levels and sentiment was weak which indicates that the price reductions do not necessarily lead to satisfaction.



*Figure 7:* correlation matrix

Further analysis visualized this relationship for a better understanding of the correlation between sentiment scores, ratings, and the discount of the products.

*Sentiment Score vs. Discount Percentage*: The scatter plot analysis shows no strong correlation between the level of discount and sentiment score. This indicates that while discounts are prevalent, they do not necessarily ensure higher customer satisfaction.

A graph with blue dots

Description automatically generated

*Figure 8:* Scatterplot of Sentiment Score vs. Discount Percentage

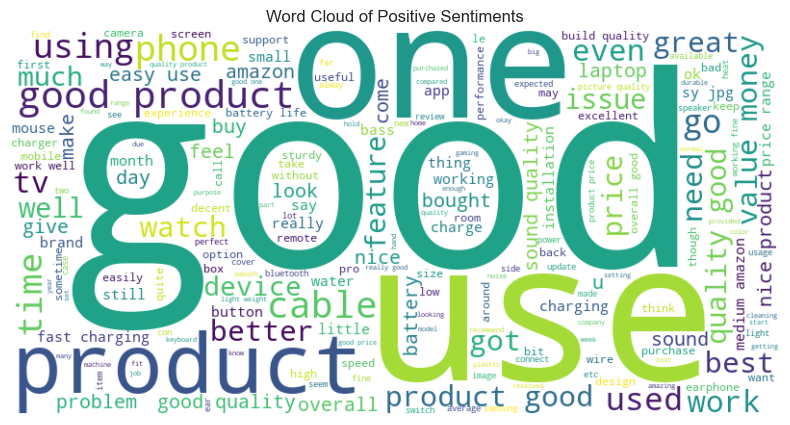
*Sentiment Score vs. Rating*: Conversely, there is a clearer positive trend in the scatter plot between sentiment scores and product ratings. Higher ratings typically correlate with higher sentiment scores, underscoring that product satisfaction strongly influences customer ratings.

A graph with blue dots

Description automatically generated

*Figure 9:* Scatterplot of Sentiment Score vs. Discount Percentage

Using word clouds, the most prominent keywords from the customer reviews where visualized to differentiate between the sentiments which are expressed in the positive and negative feedback. As depicted in figure 10, the positive word cloud highlights terms like 'good, 'quality’, and 'easy' suggests that customers appreciate the quality and usability of the purchase. Syndicates a general satisfaction with the products value and performance, which can significantly enhance customer loyalty and positive brand perception.



*Figure 10:* The word cloud of positive reviews

Figure 11 showcases keywords such as poor, bad, problem and quality from the negative reviews which points to the areas where customers expectation may not be met, and these terms highlight concerns primarily associated with product functionality and customer service which suggest that these are critically areas which needs improvement. The words like return and money also implies the dissatisfaction among the customers with the value of which for the emphasizes the need for addressing these issues to mitigate the negative customer experiences.



*Figure 11:* The word cloud of negative reviews

Combining the insights from both the clouds it becomes evident that the term quality appears in which signifies the customers perception. This shows the pivotal role product quality place in influencing customer satisfaction and highlights the importance of consistent product excellence.

*LDA topic model*

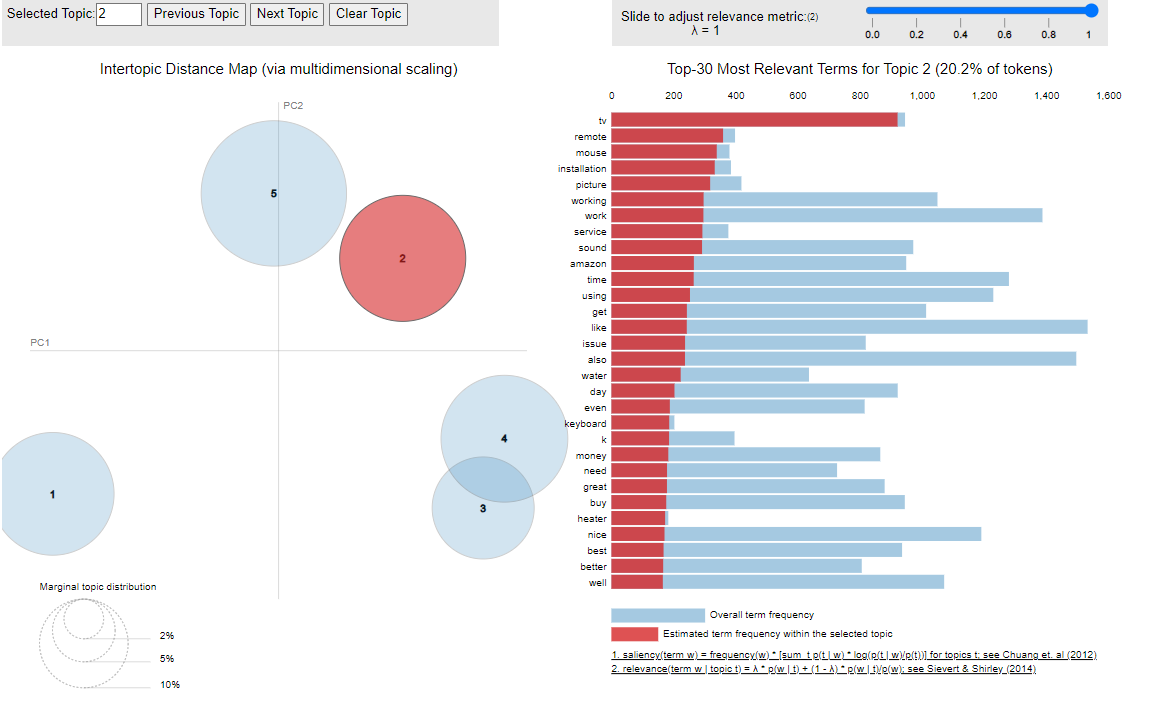
The latent Dirichlet allocation (LDA) provides insightful themes that customers provided in their Amazon reviews, the span of a variety of aspects which are related to the product usage and customer experiences (IEEE Computational Intelligence Society et al., n.d.). There were 5 distinct topics identified with each showing specific customer interest and concerns.

Topic 1 focuses on "Electronics and Accessories," emphasizing the importance of performance aspects such as "cable," "charging," and "speed," reflecting customers' preference for functional efficiency in electronic accessories. Topic 2, "Home Electronics Setup and Usability," includes terms like "tv," "remote," and "installation," implying a focus on the user experience and practicalities of setting up home electronics.

A screenshot of a graph

Description automatically generated

*Figure 12: LDA model for topic 1*



*Figure 13: LDA model for topic 2*

Topic 3 focuses on "Personal Electronics and Audio Devices," with terms such as "sound" and "earphone" expressing consumer interest in audio quality and gadget functionality. Topic 4 explores "Mobile Devices and Functional Features," focusing on advanced characteristics such as "battery" and "camera" that consumers consider while discussing mobile technologies. Finally, Topic 5, "Practical Applications and Consumer Electronics," has terms like "easy" and "clean," suggesting a general interest in consumer electronics that improve daily life through simplicity and practical application. Each topic enlightens on distinct consumer preferences and potential opportunities for product improvement and focused marketing. The findings from the LDA topic modeling can be cross-referenced with the top sub-categories by ratings to examine as the important topics correlate with the products receiving the highest ratings.

A screenshot of a computer

Description automatically generated

*Figure 14: LDA model for topic 3*

*A screenshot of a computer

Description automatically generated*

*Figure 15: LDA model for topic 4*

The LDA model has extracted successfully with meaningful themes from a wide array of customer reviews comma providing a great understanding of what drives the satisfaction in different product categories. By linking these topics with actual product ratings and categories the businesses can align their product development and marketing strategies more precisely with the customer expectations and preferences full stop this approach not only enhances targeted marketing efforts but also helps in product enhancement strategies to meet the customer needs better.

A screenshot of a computer

Description automatically generated

*Figure 16: LDA model for topic 5*

**Limitations**

There are several limitations that impact the depth and reliability of the insights derived as the analysis was conducted through sentiment analysis, word cloud and LDA model. Firstly, the quality and scope of the data might not incorporate with all product categories which potentially draws results towards more frequently reviewed items and overlooking niche products. Word clouds provide limited quantitative insights and simplify data representation while LDA assumes about topic distribution which may not accurately reflect the textual data complexity. additionally, methods do not capture nuanced semantic elements such as sarcasm or irony which can lead to misinterpretations of sentiment. Also, this analysis ignores temporal variations that could influence consumer behavior trends. To address these constraints in future studies, more complex approaches and larger data sets may be used to improve the strength and applicability of research findings.

**Conclusion**

This report examined and extensive data set which reveals strong positive feedback from customers which highlights overall satisfaction with products. The descriptive analysis of categories further demonstrators the key areas of consumer engagement and product interest which can guide strategic business decisions. The results from the sentiment analysis revealed a higher incidence of positive sentiments which shows a general customer satisfaction with the products purchased on Amazon. The word clouds effectively extracted the common theme from positive and negative reviews which identified specific areas of strength and opportunities for improvement. Additionally, the LDA revealed different topics that captured the different aspects of consumer discussions which provided a deeper understanding of the factors which influence the customer opinions and purchase behaviors. Future studies should aim to integrate large data sets and more complex analytical tools which can build on the findings presented here hence enhances the understanding of consumer behavior in the e-commerce market.

# References

Bayhaqy, A., Sfenrianto, S., Nainggolan, K., & Kaburuan, E. R. (n.d.). *XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE Sentiment Analysis about E-Commerce from Tweets Using Decision Tree, K-Nearest Neighbor, and Naïve Bayes*. http://dlvr.it/Qb83n8pic.twitter.com/8MucIMhUMO,

Dey, S., Wasif, S., Tonmoy, D. S., Sultana, S., Sarkar, J., & Dey, M. (2020). A Comparative Study of Support Vector Machine and Naive Bayes Classifier for Sentiment Analysis on Amazon Product Reviews. *2020 International Conference on Contemporary Computing and Applications, IC3A 2020*, 217–220. https://doi.org/10.1109/IC3A48958.2020.233300

Huang, H., Zavareh, A. A., & Mustafa, M. B. (2023). Sentiment Analysis in E-Commerce Platforms: A Review of Current Techniques and Future Directions. In *IEEE Access* (Vol. 11, pp. 90367–90382). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ACCESS.2023.3307308

IEEE Computational Intelligence Society, International Neural Network Society, Institute of Electrical and Electronics Engineers, & IEEE World Congress on Computational Intelligence (2016 : Vancouver, B. C. ). (n.d.). *2016 International Joint Conference on Neural Networks (IJCNN) : 24-29 July 2016, Vancouver, Canada.*

IEEE Malaysia Section. Electron Devices Chapter, Universiti Kebangsaan Malaysia. Institute of Microengineering and Nanoelectronics, & Institute of Electrical and Electronics Engineers. (n.d.). *2018 IEEE International Conference on Semiconductor Electronics (ICSE2018) : proceedings : 15-17 August 2018, Pullman Kuala Lumpur City Centre Hotel & Residence, Kuala Lumpur, Malaysia*.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

Nur Amir Sjaif, N. (n.d.). A Survey on Sentiment Analysis Approaches in e-Commerce. In *IJACSA) International Journal of Advanced Computer Science and Applications* (Vol. 12, Issue 10). www.ijacsa.thesai.org

Prabowo, R., & Thelwall, M. (2009). Sentiment analysis: A combined approach. *Journal of Informetrics*, *3*(2), 143–157. https://doi.org/10.1016/j.joi.2009.01.003

Rajput, A. (2019). Natural language processing, sentiment analysis, and clinical analytics. In *Innovation in Health Informatics: A Smart Healthcare Primer* (pp. 79–97). Elsevier. https://doi.org/10.1016/B978-0-12-819043-2.00003-4

Sari, P. K., Alamsyah, A., & Wibowo, S. (2018). Measuring e-Commerce service quality from online customer review using sentiment analysis. *Journal of Physics: Conference Series*, *971*(1). https://doi.org/10.1088/1742-6596/971/1/012053

Shanghai hai yang da xue, & Institute of Electrical and Electronics Engineers. (n.d.). *Proceedings of 2020 IEEE International Conference on Artificial Intelligence and Computer Applications : ICAICA 2020 : Dalian, China, June 27-29, 2020*.

Singh, U., Saraswat, A., Azad, H. K., Abhishek, K., & Shitharth, S. (2022). Towards improving e-commerce customer review analysis for sentiment detection. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-26432-3

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, *55*(7), 5731–5780. https://doi.org/10.1007/s10462-022-10144-1

**Appendix**

*Python file link***:** [EFIMM0139\_2433088.ipynb](https://uob-my.sharepoint.com/:u:/g/personal/yt23325_bristol_ac_uk/ETbZiTy7PGxHvdDMkwvWG1IB8awY2D42_3KXdCcpcC4P8A?e=RVc2ox)

***Full Code:***

**import pandas as pd**

**# Loading the dataset**

**file\_path = 'C:/Users/hanis/Desktop/assignment/socia media analytics/data/amazon.csv'**

**df = pd.read\_csv(file\_path)**

**# Displaying the first few rows of the dataframe**

**print(df.head())**

**#summary of the dataframe**

**df.info()**

**#check duplicates**

**duplicates = df.duplicated()**

**df[duplicates]**

**df.dtypes**

**# Stripping the symbols and commas, then converting all price fields from strings to floats**

**df['discounted\_price'] = df['discounted\_price'].replace('[₹,]', '', regex=True).astype(float)**

**df['actual\_price'] = df['actual\_price'].replace('[₹,]', '', regex=True).astype(float)**

**# Displaying the converted columns to verify changes**

**df[['discounted\_price', 'actual\_price']].head()**

**# Calculating discount percentages**

**df['discount\_percentage'] = ((df['actual\_price']-df['discounted\_price'])/df['actual\_price']) \* 100**

**df[['discounted\_price', 'actual\_price', 'discount\_percentage']].head()**

**df.isnull().sum()**

**# Convert 'rating\_count' to numeric after removing commas and handling non-numeric characters, filling missing values with zero**

**df['rating\_count'] = df['rating\_count'].replace('[,]', '', regex=True)**

**df['rating\_count'] = pd.to\_numeric(df['rating\_count'], errors='coerce').fillna(0)**

**# Display the updated 'rating\_count' column to verify changes**

**df[['rating\_count']].head()**

**df.isnull().sum()**

**import nltk**

**import re**

**# Downloading necessary resources**

**nltk.download('stopwords')**

**nltk.download('wordnet')**

**nltk.download('punkt')**

**#text cleaning function**

**from nltk.corpus import stopwords**

**from nltk.stem import WordNetLemmatizer**

**stop\_words = set(stopwords.words('english'))**

**lemmatizer = WordNetLemmatizer()**

**def clean\_text(text):**

**# Removing non-alphabet characters and convertingb to lower case**

**text = re.sub('[^a-zA-Z]', ' ', text).lower()**

**# Tokenize, remove stopwords, and lemmatize**

**tokens = nltk.word\_tokenize(text)**

**lemmatized = [lemmatizer.lemmatize(word) for word in tokens if word not in stop\_words]**

**return ' '.join(lemmatized)**

**# Assuming 'df' is your DataFrame and 'review\_content' is the column name**

**df['review\_content\_clean'] = df['review\_content'].apply(clean\_text)**

**# Display the cleaned 'review\_content' to verify changes**

**df[['review\_content', 'review\_content\_clean']].head()**

**# Dropping the columns that are not necessary for the analysis**

**columns\_to\_drop = ['user\_id', 'user\_name', 'review\_id', 'img\_link', 'product\_link', 'about\_product']**

**df = df.drop(columns=columns\_to\_drop)**

**df.head()**

**df.describe()**

**import matplotlib.pyplot as plt**

**import seaborn as sns**

**#split categories**

**df['sub\_category'] = df['category'].astype(str).str.split('|').str[-1]**

**df['main\_category'] = df['category'].astype(str).str.split('|').str[0]**

**df**

**df.columns**

**#Top main categories in reviews**

**df.groupby(['main\_category'])['rating\_count'].sum().nlargest(10)**

**df.groupby(['main\_category'])['rating\_count'].sum().nlargest(5).plot(kind='bar')**

**plt.title('Top main categories by rating count')**

**df.groupby(['main\_category'])['discounted\_price'].sum().nlargest(10).plot(kind='barh')**

**plt.title('Top main categoties by discount')**

**# Check the data type of the 'rating' column**

**print(df['rating'].dtype)**

**# If it's not numeric, convert it**

**df['rating'] = pd.to\_numeric(df['rating'], errors='coerce')**

**df.groupby(['sub\_category','main\_category'])['rating'].sum().nlargest(10)**

**df.groupby(['sub\_category'])['rating'].sum().nlargest(10).plot(kind='barh')**

**plt.xlabel('ratings')**

**plt.title('top 10 sub categories by rating')**

**# Set the aesthetic style of the plots**

**sns.set(style="whitegrid")**

**# Create the plot for Discount Percentage Distribution**

**plt.figure(figsize=(5,4))**

**sns.histplot(df['discount\_percentage'], bins=30, kde=True, color='skyblue')**

**plt.title('Distribution of Discount Percentages')**

**plt.xlabel('Discount Percentage')**

**plt.ylabel('Frequency')**

**# Display the plot**

**plt.show()**

**import nltk**

**from nltk.sentiment import SentimentIntensityAnalyzer**

**# Ensure the VADER lexicon is available**

**nltk.download('vader\_lexicon')**

**# Initialize the VADER sentiment analyzer**

**sia = SentimentIntensityAnalyzer()# Download the VADER lexicon**

**# Define a function to get the compound sentiment score for each review**

**def get\_sentiment(text):**

**return sia.polarity\_scores(text)['compound']**

**# Apply the function to calculate sentiment scores**

**df['sentiment\_score'] = df['review\_content\_clean'].apply(get\_sentiment)**

**# Display the DataFrame with the new 'sentiment\_score' column**

**print(df[['product\_name', 'review\_content\_clean', 'sentiment\_score']].head())**

**print(df[['sentiment\_score', 'discount\_percentage', 'rating', 'rating\_count']].dtypes)**

**# Convert columns to numeric, errors='coerce' will set invalid parsing as NaN**

**df['rating'] = pd.to\_numeric(df['rating'], errors='coerce')**

**df.dropna(subset=['rating'], inplace=True) # This removes rows where 'rating' is NaN**

**import seaborn as sns**

**import matplotlib.pyplot as plt**

**# Correlation matrix**

**correlation\_matrix = df[['sentiment\_score', 'discount\_percentage', 'rating', 'rating\_count']].corr()**

**print("Correlation Matrix:\n", correlation\_matrix)**

**# Visualization: Scatter plot of Sentiment Score vs. Discount Percentage**

**plt.figure(figsize=(5, 5))**

**sns.scatterplot(data=df, x='discount\_percentage', y='sentiment\_score')**

**plt.title('Sentiment Score vs. Discount Percentage')**

**plt.xlabel('Discount Percentage')**

**plt.ylabel('Sentiment Score')**

**plt.grid(True)**

**plt.show()**

**# Visualization: Sentiment Score vs. Rating**

**plt.figure(figsize=(5, 5))**

**sns.scatterplot(data=df, x='rating', y='sentiment\_score')**

**plt.title('Sentiment Score vs. Rating')**

**plt.xlabel('Rating')**

**plt.ylabel('Sentiment Score')**

**plt.grid(True)**

**plt.show()**

**from tqdm.notebook import tqdm**

**# Calculate sentiment scores**

**results = []**

**for i, row in tqdm(df.iterrows(), total=len(df)):**

**scores = sia.polarity\_scores(row['review\_content'])**

**scores['product\_id'] = row['product\_id']**

**results.append(scores)**

**# Create a DataFrame with the results**

**sentiment\_df = pd.DataFrame(results)**

**sentiment\_df**

**# Categorize the compound scores**

**def categorize\_sentiment(compound):**

**if compound > 0:**

**return 'positive'**

**else:**

**return 'negative'**

**sentiment\_df['sentiment\_category'] = sentiment\_df['compound'].apply(categorize\_sentiment)**

**# Count the number of each sentiment category**

**sentiment\_counts = sentiment\_df['sentiment\_category'].value\_counts()**

**# Plotting the distribution of sentiments**

**import matplotlib.pyplot as plt**

**import seaborn as sns**

**# Create a bar plot**

**plt.figure(figsize=(5, 5))**

**bars = sns.barplot(x=sentiment\_counts.index, y=sentiment\_counts.values, palette="viridis")**

**for bar in bars.patches:**

**bars.annotate(format(bar.get\_height(), '.0f'),**

**(bar.get\_x() + bar.get\_width() / 2, bar.get\_height()),**

**ha='center', va='center',**

**size=9, xytext=(0, 8),**

**textcoords='offset points')**

**plt.title('Distribution of Sentiments data')**

**plt.xlabel('Sentiment division')**

**plt.ylabel('Frequency')**

**plt.show()**

**pip install wordcloud matplotlib**

**# Filter the DataFrame for positive and negative sentiments**

**positive\_text = ' '.join(df[df['sentiment\_score'] > 0]['review\_content\_clean'])**

**negative\_text = ' '.join(df[df['sentiment\_score'] < 0]['review\_content\_clean'])**

**from wordcloud import WordCloud**

**import matplotlib.pyplot as plt**

**def generate\_wordcloud(text, title):**

**wordcloud = WordCloud(width=800, height=400, background\_color='white').generate(text)**

**plt.figure(figsize=(10, 5))**

**plt.imshow(wordcloud, interpolation='bilinear')**

**plt.axis('off')**

**plt.title(title)**

**plt.show()**

**# Generate word cloud for positive sentiment**

**generate\_wordcloud(positive\_text, 'Word Cloud of Positive Sentiments')**

**# Generate word cloud for negative sentiment**

**generate\_wordcloud(negative\_text, 'Word Cloud of Negative Sentiments')**

**pip install gensim nltk**

**import nltk**

**from nltk.corpus import stopwords**

**from nltk.tokenize import word\_tokenize**

**nltk.download('stopwords')**

**nltk.download('punkt')**

**stop\_words = set(stopwords.words('english'))**

**# Function to preprocess text**

**def preprocess\_text(text):**

**# Tokenize and remove stop words**

**words = [word for word in word\_tokenize(text.lower()) if word.isalpha() and word not in stop\_words]**

**return words**

**# Apply preprocessing to the review content**

**df['processed\_content'] = df['review\_content\_clean'].apply(preprocess\_text)**

**from gensim import corpora, models**

**# Create a dictionary representation of the documents.**

**dictionary = corpora.Dictionary(df['processed\_content'])**

**# Filter out extremes to limit the number of features**

**dictionary.filter\_extremes(no\_below=15, no\_above=0.5, keep\_n=100000)**

**# Create a bag of words corpus**

**corpus = [dictionary.doc2bow(text) for text in df['processed\_content']]**

**# Set the number of topics**

**num\_topics = 5**

**# Build the LDA model**

**lda\_model = models.LdaModel(corpus=corpus, num\_topics=num\_topics, id2word=dictionary, passes=10, random\_state=42)**

**# Print the topics found by the LDA model**

**for idx, topic in lda\_model.print\_topics(-1):**

**print('Topic: {} \nWords: {}'.format(idx, topic))**

**pip install pyLDAvis**

**import pyLDAvis.gensim\_models as gensimvis**

**import pyLDAvis**

**# Prepare visualization**

**pyLDAvis.enable\_notebook()**

**lda\_display = gensimvis.prepare(lda\_model, corpus, dictionary, sort\_topics=False)**

**# Display**

**pyLDAvis.display(lda\_display)**