CAPSTONE PROJECT-3
Enhancing Customer Experience: A Sentiment Analysis Of E-Commerce Product Reviews

Table of Contents

1.Problem Statement

As a data analyst in an e-commerce company, you have been tasked with improving customer experience and boosting sales through understanding customer sentiment towards the company's products. Your goal is to create a comprehensive report that provides insights into customer perceptions and experiences with the products and offers clear and actionable recommendations for the company to improve their products and customer experience.

2.Project Objective

The objective of this project is to:

- Analyze customer reviews data to identify trends and patterns in customer feedback, and generate useful insights that best describe the product quality.
- Perform sentiment analysis on the customer reviews to classify each review based on its sentiment and provide a comprehensive understanding of customer perceptions and experiences with the products.

3.Data Description

This dataset online_retail.csv contains sales information for multiple retail stores across the country. **The data includes the following variables:**

1.ID: Record ID (numeric)

2.ProductID: Product ID (Categorical)

3.UserID: User ID who posted the review (Categorical)

4.**ProfileName:** Profile name of the User (Text)

5.**HelpfullnessNumerator**: Numerator of the helpfulness of the review (Numeric)

6.**HelpfullnessDenominator**: Denominator of the helpfulness of the review (Numeric)

7.**Score**: Product Rating (Numeric)

8.**Time Review**: time in timestamp (Numeric)

9.**Summary**: Summary of the review (Text)

10.**Text**: Actual text of the review (Text)

The following initial insights from the data:

The dataset consists of 568454 rows observations and 10 variables. There are 16 missing values and 0 duplicate values in the data. The distribution of variables is as follows:

1.ID: It’s an index

2.ProductID: alphanumeric and having 75k distinct values

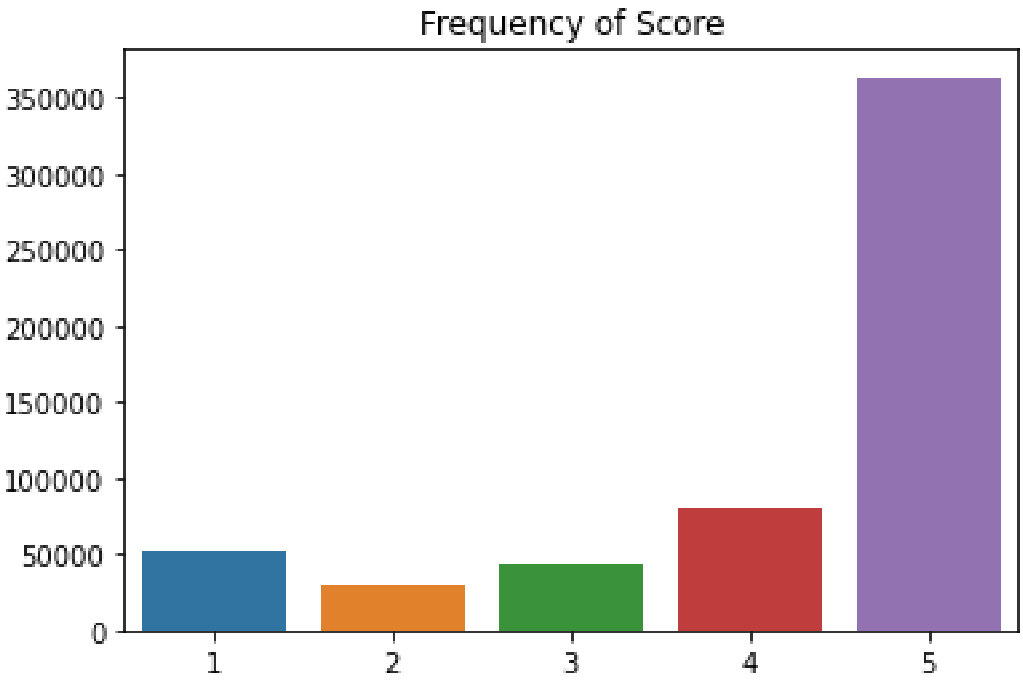
3.UserID: 256k distinct values nearly 45%

4.**ProfileName:** 218k distinct values (35%)

5.**HelpfullnessNumerator**: (0-866) range and average of 2 and having 231 distinct values. Actually, it’s significant with 0 to 7 values only. 0 is awarded by 53%

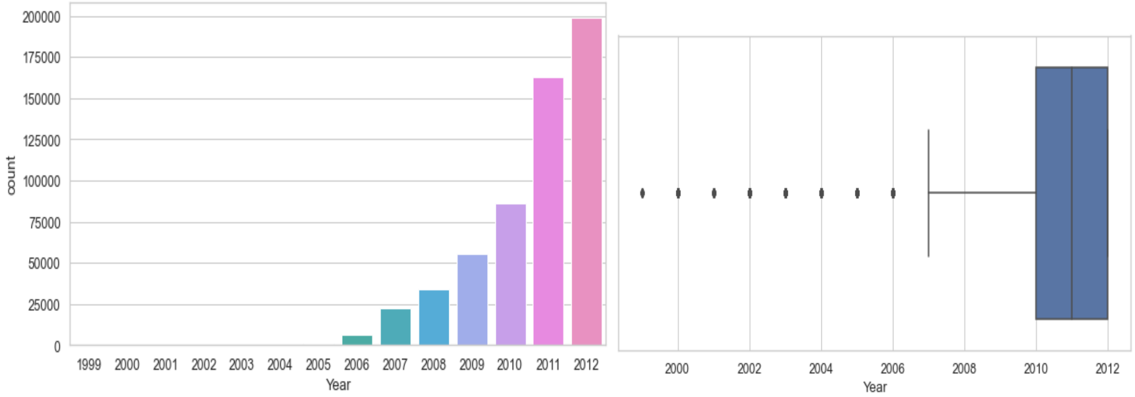
6.**HelpfullnessDenominator**: (0-923) range and average of 2 and having 234 distinct values. Actually, its significant with 0 to 7 values only 0 is awarded by 48%

7.**Score**: 5 distinct values and 5 is awarded by 64% of users



8.**Time**: 3168 unique values written in timestamp form

So its pre-processed and year data says most of reviews from 2006-2012 remaining are outliers



9.**Summary:** 295.75k distinct summaries and Delicious, Yummy are frequently used words

10.**Text**: 340k distinct texts and are lengthy paragraphs.

4. Data Pre-processing Steps and Inspiration

The pre-processing of the data included the following steps:

1. Data Cleaning:
   - Handling Missing data:
     Check null values
     No need to Drop null values as there are 16 from ProfileName variable as UserID is sufficient
   - Handling Duplicate values:
     Check duplicated values 0

2. Data Transformation:
   - Convert the datatype of the variables:
     Previously Time variable is Object it is then converted as DateTime respectively.
   - Splitting Date variable:
     Date variable is spited into week, month and year variables
   - Adding new column Helpfullness which is division of HelpfulnessNumerator an HelpfulnessDenominator.
   - Storing the cleaned dataframe as df_clean.csv

5.Choosing the Algorithm for the Project

I chose the SentimentIntensityAnalyzer which uses a pre-trained model that takes into account various linguistic rules, word associations, and sentence structure to determine the sentiment score. This is useful for performing sentiment analysis in natural language processing tasks and can be applied to a wide range of text data, such as social media posts, reviews, or customer feedback.

The steps in the algorithm include:

1. Import the SentimentIntensityAnalyzer from the nltk.sentiment module.
2. Initialize the sentiment analyzer.
3. Define a function get_sentiment that takes a sentence as input and returns the compound sentiment score.
4. Use the polarity_scores method of the sentiment analyzer to obtain a dictionary of sentiment metrics for the input sentence.
5. Extract the compound score from the sentiment metrics dictionary.
6. Return the compound score as the sentiment score of the sentence.

Here is a brief explanation of the score that is used in this code:

The sentiment score is a value between -1 and 1 that indicates the overall sentiment of the text, with values closer to 1 being more positive, values closer to -1 being more negative, and values close to 0 being more neutral.

6. Motivation and Reasons for Choosing the Algorithm

The SentimentIntensityAnalyzer is a popular choice for sentiment analysis due to several reasons:

- Robustness: The SentimentIntensityAnalyzer takes into account various linguistic rules, word associations, and sentence structure to determine the sentiment score, making it a robust tool for sentiment analysis. This allows it to handle complex linguistic phenomena such as sarcasm and negation, which are often challenging for rule-based sentiment analysis methods.

- Speed: The SentimentIntensityAnalyzer is a fast and efficient tool for sentiment analysis. It is optimized for real-time analysis and can handle large volumes of text data, making it suitable for applications where speed and scalability are important.

- Integration with other NLP tools: The SentimentIntensityAnalyzer is part of the nltk library, which is a widely used NLP library in Python. This makes it easy to integrate with other NLP tools and techniques, such as text preprocessing, named entity recognition, and text classification.

In summary, the SentimentIntensityAnalyzer provides a fast, robust, and accessible tool for sentiment analysis that can be easily integrated with other NLP techniques, making it a popular choice for sentiment analysis tasks.

7. Assumptions

The SentimentIntensityAnalyzer assumes that sentiment scores are:

- context-independent
- based on word associations
- determined using sentiment lexicons
- performed as binary positive/negative analysis.

8. Model Evaluation and Techniques

Here are some common evaluation techniques for sentiment analysis models:

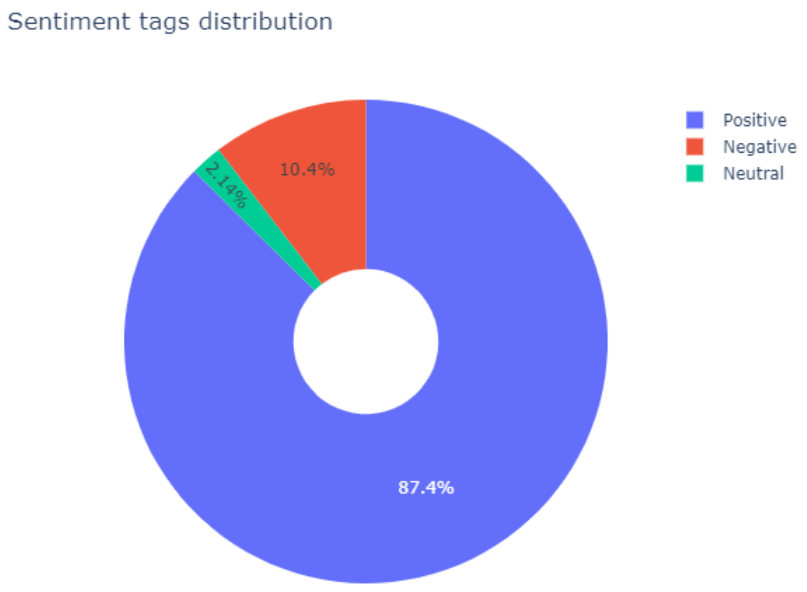
1. Accuracy: accuracy is the most commonly used metric for evaluating sentiment analysis models. It measures the proportion of correct predictions made by the model.
   Accuracy : 0.59

2. Precision, Recall, and F1 Score: Precision measures the proportion of correct positive predictions, Recall measures the proportion of positive examples that were correctly classified, and the F1 Score is the harmonic mean of precision and recall. These metrics provide a more nuanced evaluation of the model's performance, especially in imbalanced datasets.
   Precision: 0.55
   Recall: 0.59
   F1 Score: 0.56

3. Confusion Matrix: A confusion matrix is a table that summarizes the performance of a classification algorithm by comparing the true labels to the predicted labels. It helps to identify the false positive and false negative rates, which can be useful in understanding the limitations of the model.

```
[[ 12577   8999   7118   9085  14489]
 [  3899   4021   4028   5371  12450]
 [  2484   3866   4451   7525  24314]
 [  1536   2736   4108   9904  62371]
 [  4056   6951  11714  33788 306613]]
```

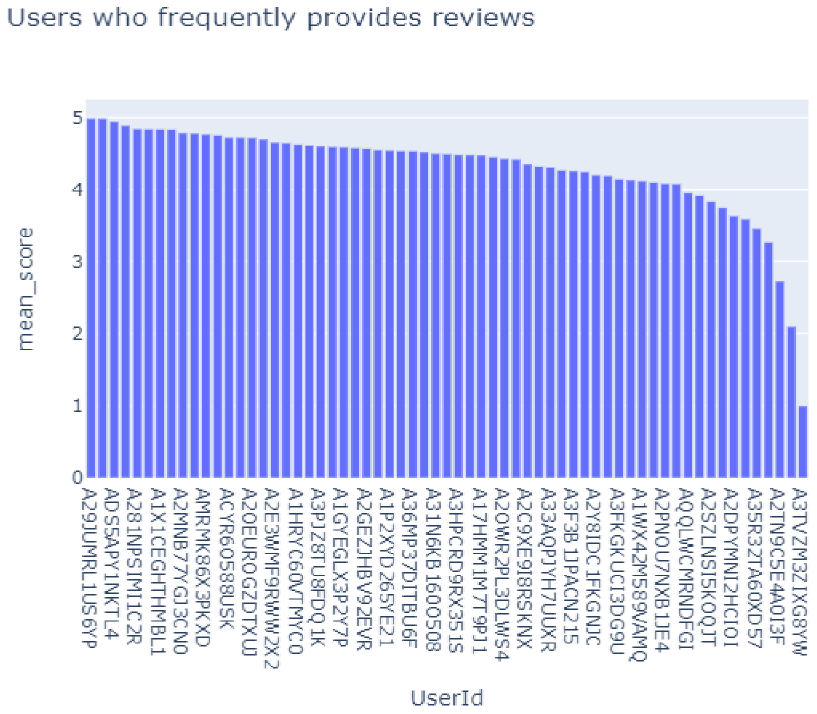
9. Inferences from the Same



Here's a list of insights from the sentiment analysis results:

- Positive sentiment: 87% of the customer reviews are positive, which suggests that the majority of customers have a good perception and experience with the products. This is a positive sign for the company and could lead to increased customer loyalty and word-of-mouth promotion.
- Negative sentiment: 10% of the customer reviews are negative, which provides valuable feedback for the company to identify areas where they need to improve their products and customer service. Addressing these issues can increase customer satisfaction and potentially turn some of these negative reviews into positive ones.
- Neutral sentiment: 2% of the customer reviews are neutral, which can provide useful information for the company. These reviews may highlight areas where customers have mixed feelings about the products or where there is room for improvement.

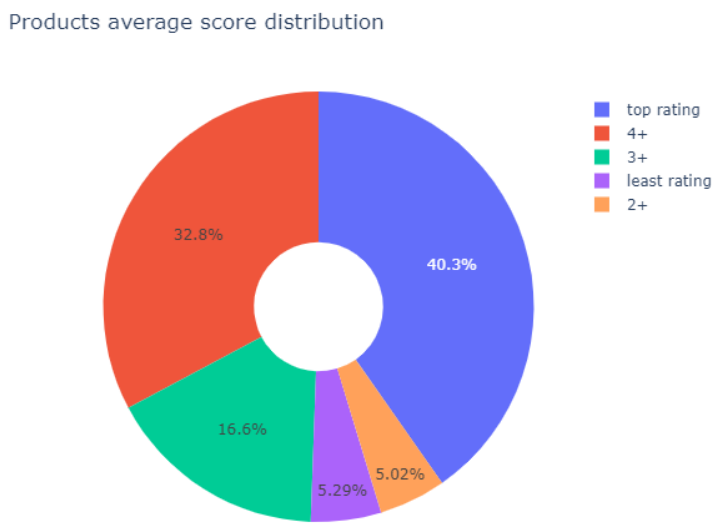
Sentiment analysis of customer reviews provides valuable insights into customer perceptions and experiences with products. By analyzing the sentiment, the company can understand what customers like and dislike and make necessary improvements to enhance the customer experience. Addressing negative reviews and monitoring sentiment over time is important to improve customer satisfaction and increase positive reviews.



Here is a list of insights from the bar graph of users who frequently provide reviews:

- The bar graph represents the distribution of users who frequently provide product reviews.
- The users depicted in this graph have purchased more than 150 products and have given scores to the products they bought.
- The mean score given by these users is greater than 4, indicating that these users generally have a positive experience with the products they purchase.
- The insights from this bar graph can provide valuable information to companies about the preferences and behavior of frequent reviewers.
- It can also give companies an idea of the products that are highly appreciated by these users and the areas that need improvement.
- These insights can help companies make data-driven decisions to enhance customer satisfaction and improve their products and services.

By analyzing the data from this bar graph, companies can gain a better understanding of the customer journey and tailor their offerings to meet the needs and preferences of their target audience.



The insights based on the product rating distribution:

- The majority of the products have received high ratings, with 40% earning top marks of 4 or 5 stars. This is a strong indication that the majority of customers are satisfied with the products they have purchased.
- A significant portion of the products, approximately a third, have received a rating of 4 or higher (32%). This further underscores the overall positive sentiment towards the products.
- However, a relatively small percentage of products have received a rating of 3 or higher (16.6%). This suggests that there may be some room for improvement for a portion of the products.
- Only 5% of the products have received a rating of 2 or higher, which indicates that a small percentage of the products may be underperforming and not meeting customer expectations.
- The lowest rating category is even smaller, with only 5.3% of the products receiving a rating of 1 or 2 stars. This suggests that the majority of the products are well-received by customers.

To further improve customer satisfaction, it may be useful for the company to conduct surveys or gather feedback from customers who have given low ratings to understand the root cause of the issue and address it in future product development and improvement efforts.

By utilizing these insights, the company can take action to continue delivering high-quality products that meet or exceed customer expectations and maintain a strong reputation in the market.

10. Future Possibilities of the Project

The future possibilities:

- Improving Product Offerings: The insights from the product rating distribution can be used to identify areas where specific products may be underperforming. The company can gather feedback from customers who have given low ratings to understand the root cause of the issue and implement improvements. This will help ensure that the company continues to offer high-quality products that meet or exceed customer expectations.
- Customer Feedback Programs: The company can implement customer feedback programs, such as regular surveys, to gather more in-depth insights into customer preferences and experiences. This information can be used to make data-driven decisions and further improve the customer experience. The company can also use this feedback to identify areas for improvement and develop targeted marketing campaigns.

By regularly monitoring customer sentiment and feedback, the company can maintain customer satisfaction, improve their offerings, and stay ahead in the market. The results of this project provide a valuable tool for the company to drive business growth and success in the future.

11. Conclusion

In conclusion, the sentiment analysis, bar graph of frequent reviewers, and product rating distribution provide a comprehensive understanding of customer perceptions and experiences with the products. The sentiment analysis results show that 87% of customer reviews are positive, which is a positive sign for the company and could lead to increased customer loyalty and word-of-mouth promotion. This high positive sentiment is further supported by the bar graph of frequent reviewers, which shows that these users generally have a positive experience with the products they purchase, with a mean score greater than 4.

The product rating distribution also provides valuable insights into the overall customer satisfaction with the products. With 40% of the products receiving a rating of 4 or 5 stars, it is clear that the majority of customers are satisfied with their purchases. However, the lower ratings for some products indicate that there is still room for improvement. By utilizing these insights, the company can take necessary actions to improve their offerings, address any negative feedback, and continue to meet or exceed customer expectations. The insights from this analysis can help the company make data-driven decisions to enhance customer satisfaction, maintain a positive reputation in the market, and drive business growth.

12. References

1. Mejova, Yelena. "Sentiment analysis: An overview." University of Iowa, Computer Science Department (2009).
2. Medhat, Walaa, Ahmed Hassan, and Hoda Korashy. "Sentiment analysis algorithms and applications: A survey." Ain Shams engineering journal 5.4 (2014): 1093-1113.
3. Feldman, Ronen. "Techniques and applications for sentiment analysis." Communications of the ACM 56.4 (2013): 82-89.
4. Elbagir, Shihab, and Jing Yang. "Sentiment analysis on Twitter with Python's natural language toolkit and VADER sentiment analyzer." IAENG Transactions on Engineering Sciences: Special Issue for the International Association of Engineers Conferences 2019. 2020.
5. Cox, Grace, and Anuska Acharya. "Sentiment Analysis and NLP models for Identifying Biases of Online News Stations." (2021).
6. Podo, Luca, and Paola Velardi. "Plotly. plus, an Improved Dataset for Visualization Recommendation." Proceedings of the 31st ACM International Conference on Information & Knowledge Management. 2022.