**By: Jaslyn Toh Lixuan**

**Sentiment Analysis on Amazon Electronics**

**Introduction**

In today’s digital age, added convenience from technological advancements have transformed the way people operate, from taking online lessons to making online purchases. Amazon is one of the top e-commerce platforms for consumers and the usability of the website and vast amount of options and categories for purchase is no doubt why consumers are flocking towards Amazon. In this report, we will explore Amazon’s quality of products and how Amazon managed to attract buyers to purchase on their online platform, by conducting text and sentiment analysis on its reviews in the Electronics category.

**Motivation**

As a top e-commerce platform driven by consumers, Amazon’s product quality and platform revenue is essential to its growth and reputation. Thus, consumers have to be heard. This can be through analysing reviews so that product performance and bad sellers can be identified. This can help Amazon improve their seller management, as well as product quality. With text mining, reviews can be analysed in a more efficient way, allowing Amazon to change its strategies to cater to consumer’s voice.

**Dataset**

**A picture containing table

Description automatically generated**

(Fig. 1 Datafiniti\_Amazon\_Consumer\_Reviews\_of\_Amazon\_Products.csv)

The amazon dataset used in this paper is an open source dataset which is extracted from the data.world website. The dataset contains consumer reviews for amazon electronic products such as Kindle and Amazon Alexa. There are 5,000 rows and 18 columns. We will be focusing on the **primaryCategories, reviews.numHelpful, reviews.text and reviews.rating** for the sentiment analysis.

**PrimaryCategories:** 4 sub-categories from Electronics

**Reviews.numHelpful:** On every review, there will be a button for other consumers to give ratings to the review. This column represents the number of times a review is marked as ‘helpful’ for other consumers.

**Reviews.text:** This column represents the viewpoint of the user who created the review.

**Reviews.rating:** The user may leave a rating for the review. Ratings are on a scale of 1 to 5.

**Exploratory Data Analysis**

1. **Popularity of words**

In order to find out the trends in the electronic category and the popularity of products, we can examine the popularity of words. By looking at the frequency of each word which is mentioned, we can roughly estimate the overall sentiment of the products. This will be done through extracting the frequency of words using the term document matrix and the final result will be plotted on word cloud and a simple bar plot.

Text

Description automatically generatedChart, bar chart

Description automatically generated

(Fig. 2 Word cloud and Bar plot on text popularity)

As observed in the word cloud and bar plot, the word ‘great’ is the top choice of usage. This shows that the electronics category on Amazon largely garnered positive remarks. It is observed that ‘great’ is the most mentioned word, followed by ‘tablet’. This means that consumers are very satisfied with the tablet being bought on Amazon. Other popular words are descriptive, consisting of ‘love’, ‘use’, ‘can’ and ‘easy’. These words also show that the product is user friendly and it is easy to use in the consumers’ eyes. Amongst the other non-descriptive words, product names such Kindle and Alexa are widely used, indicating the popularity of such products amongst the Electronics category.

1. **Relationship between review length and review ratings**

**A picture containing text

Description automatically generated**

(Fig 3. Summary statistics of review length)

To further enhance our exploratory analysis, we can look at the relationship between the length of the review to the review ratings. From the summary of our review length, the median is about 105 characters, while the maximum characters is 8351.

**Chart, scatter chart

Description automatically generated**

(Fig. 4 Scatter plot of length of review and reviews rating)

As seen from the scatter plot, the longer the review length, the higher the ratings. Also, reviews with rating of 2 have the shortest length in general. This shows that consumers prefer to write more when they are satisfied with the product, and will take time to leave long reviews for good products.

1. **Relationship between review length and helpfulness of ratings**

The helpfulness of rating is another way we can analyse the sentiment of the reviews, as it is a factor users take into consideration when buying the product. By plotting a scatter plot, we can observe the relationship between review length and helpfulness of ratings.

**Chart, scatter chart

Description automatically generated**

(Fig. 5 Scatter plot of length of review against helpfulness)

We can observe that the large majority of the consumers do not leave helpful ratings for shorter length reviews, as they fall under 0 number of helpful ratings. However, the helpfulness of the longest review also has the highest number of ratings. This shows that consumers trust reviews which provide more information, as they help them to make more informed choices during their purchase.

1. **Mean ratings of primary categories**

There are 4 primary categories under Electronics in the dataset – electronics, electronics (hardware), electronics (media), electronics (office supplies). As they belong to different categories under electronics, we can find out each category’s performance in terms of their average review ratings. As observed from the bar plot, all the categories have ratings above 4. Electronics hardware has a slightly higher rating as compared to the rest.

Chart, bar chart

Description automatically generated

(Fig. 6 Bar chart for average ratings by category)

**Sentiment Tagging with Ratings (Positive/Negative)**

**Text

Description automatically generated**

(Fig. 7 Summary statistics of review rating)

The Amazon dataset provided ratings for each product review. Using the summary function, we can observe that the highest rating is 5, while the minimum rating is 1. The ratings of the electronic products seem to be high for Electronics category as seen from the high median score of 5. The classification is done where ratings more than 3 is positive, while ratings less than 3 will be tagged as negative. Although the ratings are largely accurate in identifying the positive sentiments of the reviews, there are mixed reactions where users give a high rating of more than 3 but review the product in a **negative light.**

Graphical user interface, application

Description automatically generated

(Fig. 8 Sentiment Tagging)

**Text Processing & Modelling to predict sentiments**

The reviews.text column is processed by stemming, removing stop words in English, removing punctuation, removing symbols, removing numbers and URLs. The tf-idf is then computed with dfm\_tfidf and training and test sets are created. The dataset is split into Training – 50% and Testing – 50%. The label is ‘Sentiment’, where there are 2 outcomes – Positive (Pos) and Negative (Neg)

1. **Naïve Bayes model**

**Text

Description automatically generatedTable

Description automatically generated**

(Fig. 9 Results of Naïve Bayes)

The Naïve Bayes model applies the Bayes theorem and assumes strong independence between features. This model is said to achieve very high accuracy scores with a fast speed on large datasets. In this report, it is used as a baseline model. After building the model, the final scores are generated on the confusion matrix. As observed, the model achieved high **precision**, where among the labelled positives, 96.40% are actually accurately predicted as positive reviews. The **recall score** is also high at 90.40%, where of the reviews that are truly positive, 90.40% of the time it is labelled. The **accuracy** of the model is 87.84%, showing a high ratio of correctly predicted observation to the total number of observations. However, accuracy is not very useful here as the dataset is very unbalanced. Thus, using an **F1-score** would be more accurate, where it takes the weighted average of precision and recall. The model achieved a high 93.30% F1-score, proving that the model has excellent performance. However, with Naïve Bayes, the main limitation is its assumption of independence between features. This assumption is rarely accurate in real life, as features do depend on one another.

1. **Support Vector Machine (SVM)**

Text

Description automatically generatedTable

Description automatically generated

(Fig. 10 Results of SVM)

The SVM model is a supervised machine learning model which uses classification algorithms for binary classifications. SVM is relevant in text mining as it helps to categorise text. Similarly to NB, this model is said to achieve very high accuracy scores with a fast speed on large datasets. After building the model, the final scores are generated on the confusion matrix. As observed, the model achieved high **precision**, where among the labelled positives, 94.37 % are actually accurately predicted as positive reviews. The **recall score** is extremely high at 99.49%, where of the reviews that are truly positive, 99.49% of the time it is labelled. The **accuracy** of the model is 93.96%, showing a high ratio of correctly predicted observation to the total number of observations. However, accuracy is not very useful here as the dataset is very unbalanced. Thus, using an **F1-score** would be more accurate, where it takes the weighted average of precision and recall. The model achieved a high 96.86% F1-score, proving that the model has excellent performance.

**Comparison between SVM and NB with ROC Curve**

The ROC (receiver operating characteristic curve) Curve represents the performance of the classification models at all classification thresholds.

Chart, line chart

Description automatically generated

Naïve Bayes

SVM

(Fig. 11 Comparison of ROC)

Both models have a high ROC value where they lie above 0.5. Comparing the True positive rate/False positive rates of SVM and Naïve Bayes model, we can observe that the SVM model is closer to 1 and wins as the better model. This means that SVM is better at distinguishing the positive reviews from the negative reviews.

Chart, line chart

Description automatically generated

SVM

Naïve Bayes

(Fig. 12 Comparison of precision)

Both models have a high precision value where they lie above 94%. Comparing the precision scores of SVM and Naïve Bayes model, we can observe that the NB model wins as the better model, with a higher precision score of 96% above. This means that NB is better at predicting positive results when the label is truly positive.

SVM

Chart

Description automatically generated

Naïve Bayes

(Fig 13. Comparison of Recall)

Both models have a high precision value where they lie above 80%. Comparing the recall scores of SVM and Naïve Bayes model, we can observe that the SVM model wins as the better model, with a higher recall score close to 100%. Overall, the SVM model performed better than Naïve Bayes in terms of giving a better recall and accuracy score.

**Limitations**

As mentioned, tagging sentiments with user ratings may not generate the best outcome. As seen in the above, users could give a high rating but the reviews might say otherwise when observed manually. The tagging of sentiments can be improved with other methods, such as using a dictionary of positive and negative words and tag the dataset against the dictionaries.

**Conclusion**

In conclusion, the text analysis and modelling on Amazon electronics dataset has shown how powerful text mining can be in helping companies to improve their sales process. As the top e-commerce company, Amazon should definitely consider using SVM modelling as part of their seller management and product quality review process. The model is highly accurate in helping to predict positives and negatives from reviews, and thus it can be used on new reviews which have to be tagged to sentiments.