**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Number** | **Particular** | **Page Number** |
| **1** | **Introduction** | **1** |
| **2** | **Methodology** | **1** |
| **3** | **Data cleaning and pre-processing** | **3** |
| 3.1 | Data Sourcing | **3** |
| 3.2 | Data cleaning and Normalization | **3** |
| **4** | **Bag of Words Analysis** | **3** |
| **5** | **Sentiment Analysis** | **3 – 4** |
| **6** | **Regression Analysis** | **4 – 5** |
| 6.1 | Correlation heatmap | 4 |
| 6.2 | Feature Importance | 4 |
| 6.3 | Model Evaluation | 4 – 5 |
| **7** | **Topic Modelling** | **5 – 7** |
| **8** | **Conclusions** | **7** |
|  | **APPENDICES** | **8** |
|  | **Appendix 1: Data visualization** | **8 – 9** |
|  | **Appendix 2: N-gram Analysis (BoW)** | **10 – 13** |
|  | **Appendix 3: Comparative N-gram Analysis** | **14 – 16** |
|  | **Appendix 4: Sentiment Analysis**   1. Results of Sentiment analysis 2. Distribution of sentiment in each star rating and each lexicon   B1. Before Sparsity Reduction  B2. After Sparsity Reduction | **17 – 26** |
|  | **Appendix 5: Correlation Heatmaps**   1. Before Sparsity Reduction 2. After Sparsity Reduction | **27 – 30** |
|  | **Appendix 6: Feature Importance**   1. Before Sparsity Reduction 2. After Sparsity reduction | **31 – 32** |
|  | **Appendix 7: Results of Regression Analysis**   1. Before Sparsity Reduction 2. After Sparsity reduction 3. Cross-validation | **33 – 36** |
|  | **Appendix 8: Exploratory Data Analysis** | **37 – 47** |
|  | **References** | **48 – 49** |
|  |  |  |

1. **Introduction**

Online product reviews have become increasingly influential in shaping consumer buying decisions in the era of e-commerce. Scholars from various disciplines have recognized their importance and conducted extensive research (Mudambi & Schuff, 2010; Chevalier & Mayzlin, 2006). In this study, we focus on consumer reviews in the 'Digital Music' category on Amazon. Using natural language processing techniques, we analyse sentiment and extract dominant topics. We examine the correlation between text-derived features and review scores and explore the predictability of ratings using regression models (Wang et al., 2010; Liu, 2012; Pang & Lee, 2008; Hutto & Gilbert, 2015; Ghose & Ipeirotis, 2011). Our analysis aims to generate actionable insights for businesses to improve customer satisfaction and product development (Blei et al., 2003).

1. **Methodology**

The methodology, detailed in Figure 2.1, involves sourcing and cleaning Amazon's 'Digital Music' category data, with subsequent preprocessing and exploratory data analysis (EDA). Feature extraction used Bag-of-Words, TF-IDF, and sentiment analysis techniques. Ratings prediction utilized regression models, applied before and after sparsity reduction. Imbalanced data was addressed, model evaluation was based on MSE, RMSE, and R2 score metrics. Cross-validation and topic modelling were conducted for broad data insights.

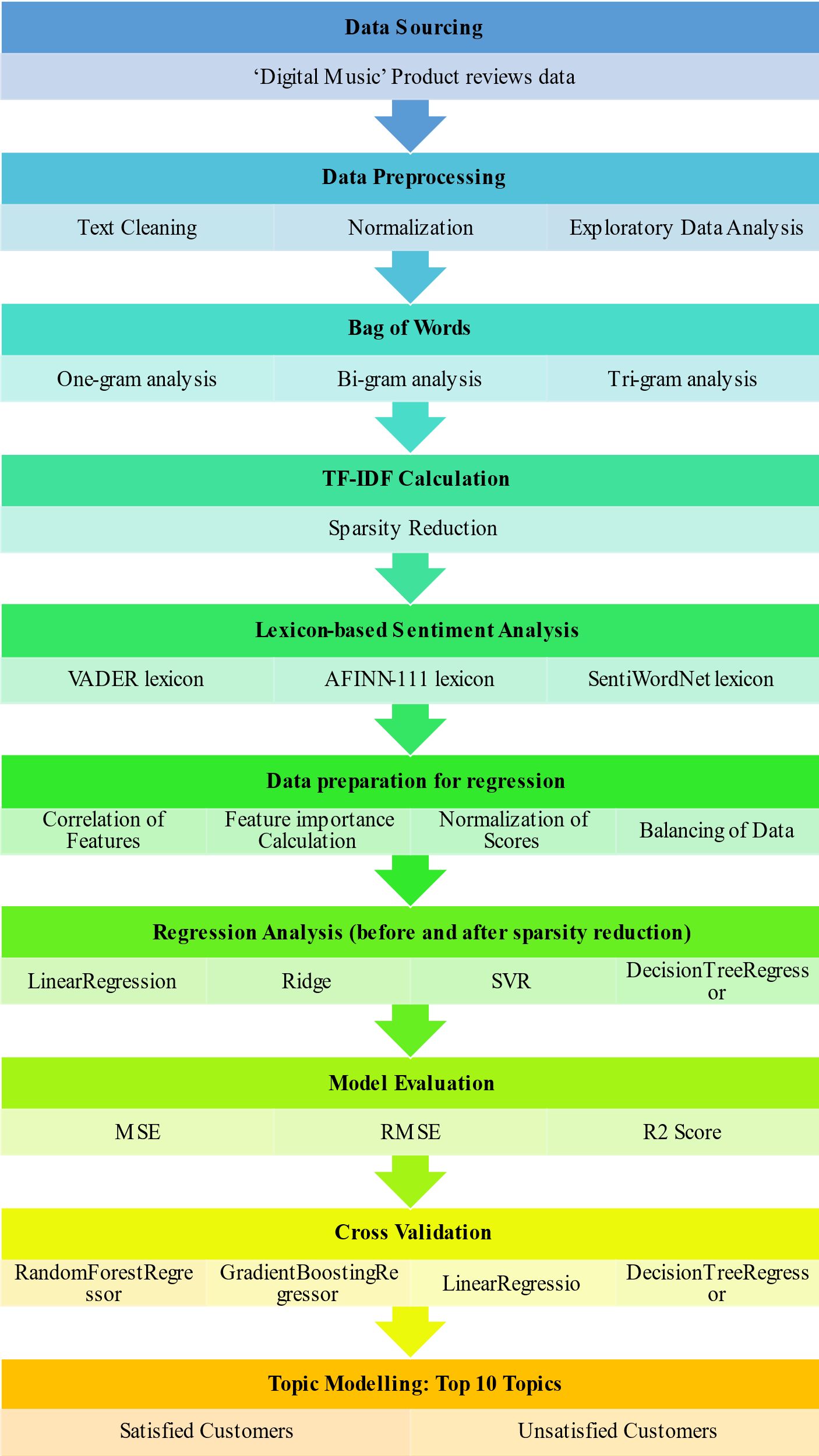


Figure 2.1. Adopted Methodology

1. **Data cleaning and pre-processing**
   1. **Data sourcing**

Data from Ni et al. (2019) was used for this report, focusing on 5-core product reviews in the 'Digital Music' category. Amazon, top player in 2021 (Statista), ranks fifth in subscribers compared to other platforms Spotify, Apple Music, etc. The music industry has experienced a transformative impact due to digitalization (Wilson, 2022), with a 26% revenue increase in the first half of 2021 (RIAA). Exploring the 'Digital Music' category offers relevant insights into contemporary consumer preferences.

* 1. **Data cleaning and Normalization**

Sentiment analysis and topic modelling rely on data cleaning and normalization to eliminate irrelevant noise and standardize text for accurate, meaningful insights. Initial steps included removing null text rows, and filtering verified reviews for credibility, yielding 148,787 reviews. Html tags, punctuation, special characters, URLs, white spaces were stripped; numbers were converted to words and text lower-cased. The analysis highlighted a Q3 2014 review peak, declining due to competition. Hence, 2017-2018 data with 16,250 reviews was deemed critical for understanding recent market shifts (Appendix 1). Stopwords were removed, part of speech tags identified, text lemmatized, spelling corrected, and meaningless words eliminated for further analysis. (See Appendix 8 for Exploratory data analysis)

1. **Bag of Words (BoW) Analysis**

The BoW technique or Bag of Features or attribute selection (Li, Li and Liu, 2017) represents text by counting individual word frequency, disregarding grammar, and word order, which aids textual data analysis. In this study, it helped identify dominant words across different star ratings. Positive terms like 'great,' 'love,' and 'good' predominated in 4 and 5-star reviews, reflecting satisfaction and positive sentiment. Neutral terms such as 'ok,' 'sound,' and 'like' dominated 2 and 3-star reviews, indicating mixed sentiments, while 1-star reviews were marked by negative words like 'order,' 'get,' 'go,' pointing towards dissatisfaction (Annexure 2).

Subsequently, Term Frequency-Inverse Document Frequency (TF-IDF) was used to pinpoint relevant, distinctive words. This approach considers both word frequency in a document (TF) and its rarity across all documents (IDF). Post TF-IDF, words appearing in more than 75% or less than 1% of documents were removed to reduce sparsity (Gopal, 2023). N-gram analysis indicated preference for classical songs with high-quality production. Reviews exceeding 20 words were considered outliers and excluded.

1. **Sentiment Analysis**

Sentiment analysis is the process of analysing text to determine the sentiment expressed by utilizing VADER, AFINN-111, and SentiWordNet lexicons, with a focus on VADER due to its high accuracy in social media language analysis (Bonta, Kumaresh and Janardhan, 2019).

The table provides a comparison of sentiment analysis results using multiple lexicons, showing the count of positive, negative, and neutral sentiments in the text data.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Lexicon** | | | | | | **Polarity** | |
| **Sentiment** | VADER | | AFINN | | SentiWordNet | |
|  | Before | After | Before | After | Before | After | Before | After |
| Positive | 13170 | 12028 | 12784 | 11736 | 9820 | 8898 | 12674 | 11684 |
| Negative | 298 | 25 | 253 | 0 | 951 | 326 | 272 | 12 |
| Neutral | 1911 | 3326 | 2342 | 3643 | 4608 | 6155 | 2433 | 3683 |
| **Total** |  | | | | | | **15379** | |

Table 1: Sentiment Analysis Results using Multiple Lexicons

Table 1 summarizes sentiment analysis results using multiple lexicons, revealing sentiment count before and after processing. Most notable is the substantial positive sentiment in data (~75%), possibly due to high occurrence of 4 and 5-star ratings (Annexure 4a).

Visual distribution across each star rating (Annexure 4b2) corroborates this dominance of positive sentiment, with an upsurge of neutral sentiment for lower ratings and minimal negative sentiment. These findings provide a valuable understanding of customer sentiment, aiding in strategy formulation to improve product and service quality.

1. **Regression Analysis**

**6.1 Correlation**

Before and after sparsity reduction, regression analysis and heatmaps demonstrated strongest correlations between ratings and VADER score (0.23 pre, 0.21 post) and polarity (0.19 pre, 0.17 post). AFINN-111 labels also correlated, with increased significance post-reduction (0.1 pre, 0.18 post). (Annexure 5).

**6.2 Feature importance**

Feature importance in regression aids model interpretation and understanding by identifying influential variables in predicting the target variable (Breiman, 2001). Length and VADER score were significant predictors for rating in regression, even after sparsity reduction (Annexure 6).

**6.3 Model evaluation**

SMOTENN was used for data balancing (Nishat et al., 2022). Models like Linear Regression (LR) (Montgomery et al., 2012), Ridge Regression (RR) (Hoerl and Kennard, 1970), Support Vector Regression (SVR) (Drucker et al., 1997), Decision Tree Regression (DTR) (Breiman et al., 1984), Random Forest Regression (RFR) (Breiman, 2001), Gradient Boosting Regression (GBR) (Friedman, 2001), and Ordinal Logistic Regression (OLR) (Agresti, 2010) were utilized. Models' performance was gauged using MSE, RMSE, and R2 (Montgomery et al., 2012). Ratings were predicted via VADER score, label, polarity, word count, and text length before/after sparsity reduction.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Before Sparsity Reduction** | | | **After Sparsity Reduction** | | |
| **MSE** | **RMSE** | **R2** | **MSE** | **RMSE** | **R2** |
| **LR** | 1.5240 | 1.2345 | 0.3856 | 2.0761 | 1.4408 | 0.2165 |
| **RR** | 1.5241 | 1.2345 | 0.3855 | 2.0761 | 1.4408 | 0.2166 |
| **SVR** | 1.7481 | 1.3221 | 0.2952 | 2.0032 | 1.4153 | 0.2440 |
| **DTR** | 0.2048 | 0.4525 | 0.9174 | 0.2156 | 0.4643 | 0.9186 |
| **RFR** | 0.1336 | 0.3655 | 0.9461 | 0.1758 | 0.4193 | 0.9336 |
| **GBR** | 0.7159 | 0.8461 | 0.7113 | 0.7288 | 0.8537 | 0.7249 |
| **OLR** | 2.6765 | 1.6360 | 0.3595 | 3.1257 | 1.7679 | 0.0889 |

Table 2: Evaluation of Regression models

The results, shown in Table 2, indicate that RFR performs best among the evaluated models, exhibiting the lowest MSE, RMSE, and highest R2 score, indicating superior predictive power. DTR also performs well but slightly underperforms compared to RFR. LR, RR, and SVR perform poorly, while GBR falls in between. The models show better performance on features derived before sparsity reduction. VADER score, polarity, and length have a positive effect on rating, while word count and label have a negative effect (Annexure 7). Positive sentiment, strong emotions, and longer reviews generally led to higher ratings. However, more words and certain labels, possibly indicating negative sentiment, lowered the ratings. These insights can help businesses improve customer satisfaction and product ratings by addressing concerns raised in negative reviews and encouraging detailed feedback.

Cross-validation is crucial for obtaining reliable and generalizable sentiment-based star rating predictions by evaluating the model's performance on unseen data and avoiding overfitting. It was evident that RF model was the best at predicting ratings based on text-derived features. It has both lowest RMSE best ability to explain variance in the ratings (highest Adjusted R2). The Decision Tree Regressor is also a good option with the smallest average absolute error (lowest MAE) (Annexure 7c).

1. **Topic Modelling**

Major topics for satisfied (star rating >= 4) and dissatisfied (star rating <= 4) customers were identified using Latent Dirichlet Allocation (LDA), a common method used for modelling and visualisation using Word Clouds.



Figure 7.1: Word cloud of Topics Identified for Satisfied Customers.

|  |  |  |
| --- | --- | --- |
| Topic\_Number | **Topic Name** | |
| **Satisfied Customers** | **Dissatisfied Customers** |
| **0** | Sound and Track Selection | Sound & CD |
| **1** | Song Preference | Song Knowledge & Preferences |
| **2** | Lyrics and Fan Love | Time, Lyrics & Love |
| **3** | Classic Songs and Timeless Appeal | Daily Play & Classic |
| **4** | Purchase Experience - Emotional Connection | Purchase & Enjoyment |
| **5** | Preferences, Listening Habits | Likability & Listening |
| **6** | Enjoyment in Songs | Song Enjoyment & Work |
| **7** | Music Quality | Best Beats & Music |

Table 3: Nomenclature of identified topics

Satisfied customers in the digital music product review dataset value sound quality, song preferences, buying convenience, enjoyable experiences, and overall satisfaction. They appreciate good lyrics, a variety of tracks, and easy purchases and downloads. These insights emphasize the significance of delivering high-quality sound, diverse song selections, and user-friendly purchasing options to enhance customer satisfaction in the digital music industry.



Figure 7.2: Word cloud of Topics Identified for Unsatisfied customers.

Dissatisfied customers in the digital music product review dataset have concerns regarding sound quality, song preferences, delivery time, lyrics, enjoyment, and overall satisfaction with the purchased music product. They express dissatisfaction with factors such as song versions, beats, and the overall experience. These insights provide valuable understanding of the specific areas of dissatisfaction and can guide improvements in product quality, song selection, delivery process, and overall customer satisfaction.

1. **Conclusions**

Analysis of 'Digital Music' product reviews on Amazon yielded valuable insights:

1. Higher-rated reviews had more positive terms, while lower-rated reviews contained negative words.
2. Most reviews expressed positive sentiment, likely influenced by high 4 and 5-star ratings.
3. RFR effectively predicted ratings, highlighting machine learning's potential in sentiment analysis.
4. Topic modelling identified satisfied customers emphasizing sound quality and song selection, while unsatisfied customers criticized song versions and overall experience.

These insights aid businesses in improving customer satisfaction and ratings by addressing negative feedback and encouraging positive reviews. However, limitations include data imbalance, accuracy of lexicon, and potential cultural/language differences among reviewers. Future work could involve better quality data, advanced models (e.g., deep learning), and exploring temporal sentiment and topic trends to understand evolving consumer opinions.

**APPENDICES**

**Appendix 1: Data visualization**

**A picture containing line, sketch, handwriting

Description automatically generated**

Figure 1. Number of reviews against year

Figure 1 above depicts a plot of reviews from 1999 to 2018, showing increased reviews over time and then a decline after the third quarter of 2014.

**A picture containing line

Description automatically generated**

Figure 2: Number of reviews in each quarter from year 2012 to 2018

Above Figure 2 number of reviews show a decline after 2015. The number of reviews may be attributed to the emergence of competitors such as Spotify and Apple Music. To gain more meaningful insights and improve services, it is essential to focus on studying more recent data. Analysing the sentiment and conducting topic modelling on the current demand can provide valuable insights into customer preferences and expectations. By understanding the evolving landscape and addressing the needs of the present, it becomes possible to adapt and enhance the services offered to meet the changing market dynamics effectively.

A blue line graph on a white background

Description automatically generated with low confidence

Figure 3: Number of reviews in the study sample

The previous graph analysis underscores a noticeable trend, the peak of reviews in Q3 2014, with a subsequent drop possibly due to competitors like Spotify and Apple Music. Studying recent data, through sentiment analysis and topic modelling, is vital for discerning contemporary customer needs and preferences, thus enabling effective adaptation to shifting market dynamics.

**Appendix 2: N-gram Analysis (BoW)**

This section highlights the Bag of Words Analysis conducted in the study. The provided text analyzes the dominant words in Amazon Digital Music product reviews based on different star ratings (5.0, 3.0, 4.0, 2.0, 1.0).

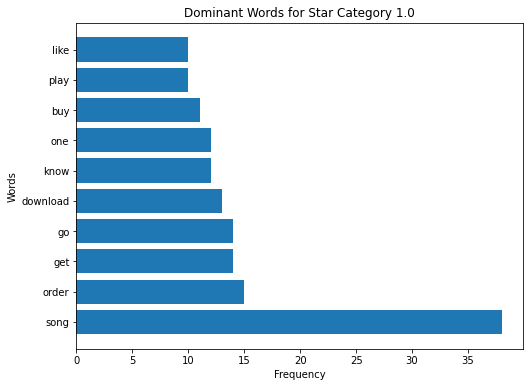
****

Figure 4: Top 10 common words in star rating 1.0

The dominant words in reviews with a 1.0 rating include 'song', 'order', 'get', 'go', and 'download'.

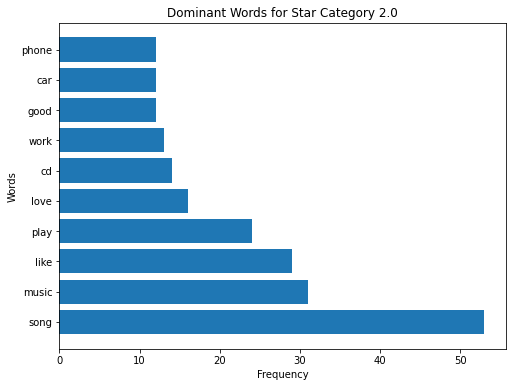
****

Figure 5: Top 10 common words in star rating 2.0

Negative reviews with a 2.0 rating mention words like 'song', 'music', 'like', 'play', and 'love'.

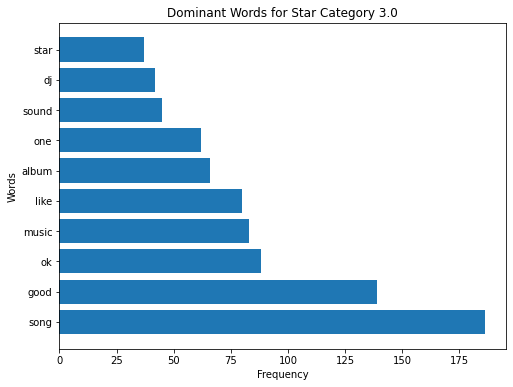
****

Figure 6: Top 10 common words in star rating 3.0

Neutral reviews with a 3.0 rating feature words such as 'ok', 'sound', and 'dj'.

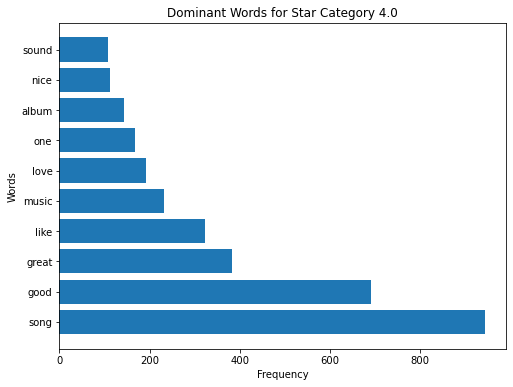
****

Figure 7: Top 10 common words in star rating 4.0

The dominant words in reviews with a 4.0 rating are 'song', 'good', 'great', 'like', and 'music'.

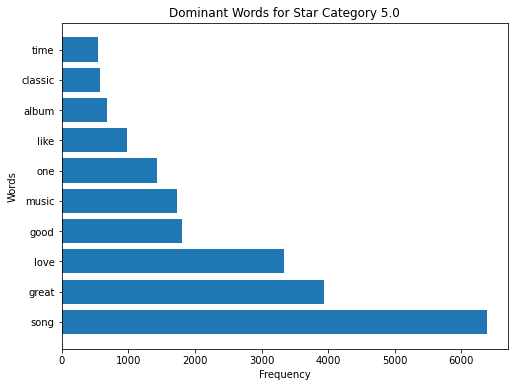
****

Figure 8: Top 10 common words in star rating 5.0

Finally, Positive reviews with a 5.0 rating commonly include words like 'song', 'great', 'love', and 'good'. These findings provide insights into customer sentiment and align with the goal of analysing sentiment in the review data.

**Appendix 3: Comparative N-gram Analysis**

**A picture containing screenshot, design, pixel

Description automatically generated**

Figure 9: Top 10 common words before(left) and after(right) sparsity reduction.

The identification of the top ten common words in the dataset provides insights into the most frequently occurring terms. In the given dataset, the following words are observed as the most common:

1. "song"

2. "great"

3. "love"

4. "good"

5. "music"

6. "one"

7. "like"

8. "album"

9. "sound"

10. "classic"

The presence of these common words in the digital music reviews indicates their high frequency and significant impact on sentiment analysis and topic modelling. Analysing these words helps understand prevalent themes, sentiments, and customer preferences. "Song" emphasizes the importance of specific songs, while "great," "love," and "good" reflect positive sentiments. "Music," "sound," and "album" indicate attention to musical aspects. Overall, these words reveal key aspects mentioned by reviewers, such as song quality, positive sentiments, and the overall music experience. Understanding these common words provides insights into customer preferences and sentiments in digital music reviews.

**A picture containing screenshot, line, graphics, design

Description automatically generated**

Figure 10: Top 10 bi grams before (left) and after(right) sparsity reduction.

The two-gram words found in the dataset include "great song", "love song", "good song", "great music", "thousand nine", "one thousand", "nine hundread", "awesome song", "one favorite", "good music". These word combinations indicate commonly occurring phrases in the reviews, reflecting positive sentiments towards songs, preferences for favorites, and occasional occurrences of numeric references. Analyzing these phrases provides insights into reviewers' sentiments and preferences.

**A picture containing screenshot, line, graphics, design

Description automatically generated**

Figure 11: Top 10 trigrams before(left) and after(right) sparsity reduction.

efore sparsity reduction, the top 10 most common trigrams include phrases like "thousand nine hundred," "one thousand nine," "great song great," "classic song add," and "mp3 sound quality." After sparsity reduction, the top 10 most common trigrams consist of phrases such as "thousand nine one," "great song love," "good quality music," "one song favourite," and "music perfect sound.” This analysis provides insights into the frequently occurring combinations of words within the reviews, both before and after sparsity reduction, showcasing the key trigrams that contribute to the overall sentiment and themes in the reviews.

**Appendix 4: Sentiment Analysis**

1. **Results of sentiment analysis**

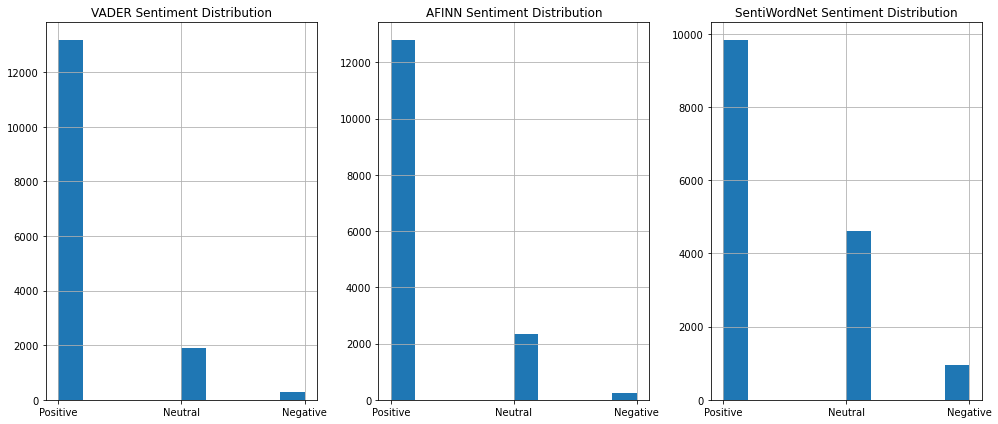
****

Figure 12: Sentiment analysis on multiple lexicons (before sparsity reduction)

* + VADER: Out of the total reviews, 13,170 were classified as positive sentiment using the VADER lexicon.
  + AFINN: The AFINN lexicon identified 12,784 reviews as positive sentiment.
  + SentiWordNet: Based on SentiWordNet, 9,820 reviews were classified as positive sentiment.
  + Polarity: Using a polarity measure, 12,674 reviews were determined to have positive sentiment.
  + Summary\_text: Among the sentiment categories, there were 3831 positive, 298 negative, and 1911 neutral reviews.

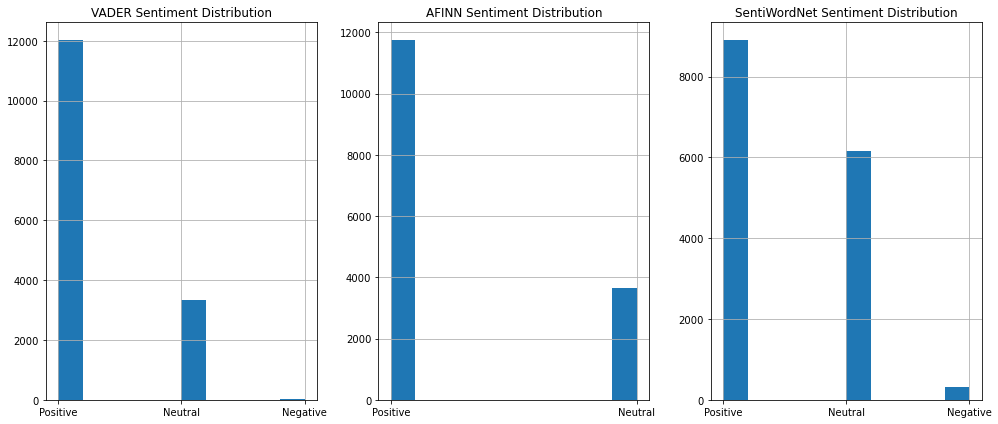
****

Figure 13: Sentiment analysis on multiple lexicons (After sparsity reduction)

- VADER: VADER lexicon identified 12,012 positive sentiments, 40 negative sentiments, and 3,327 neutral sentiments.

- AFINN: AFINN lexicon detected 11,736 positive sentiments, 0 negative sentiments, and 3,643 neutral sentiments.

- SentiWordNet: SentiWordNet lexicon assigned 8,898 positive sentiments, 326 negative sentiments, and 6,155 neutral sentiments.

- Polarity: The polarity column indicates that VADER has the highest number of positive sentiments (11,677), followed by SentiWordNet (8,898), AFINN (11,736), and finally, the polarity calculated (11,677).

1. **Distribution of sentiment in each star rating and each lexicon**

**B1: Before Sparsity reduction**

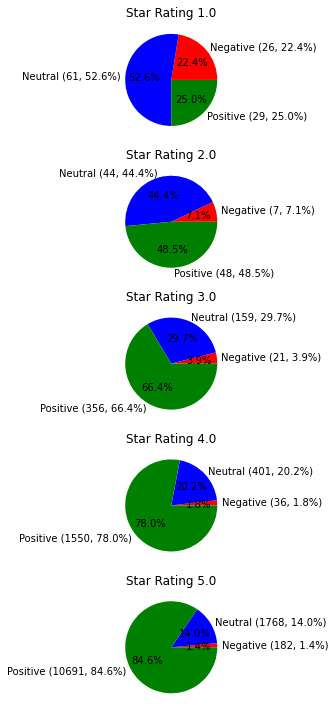
****

Figure 14: Sentiment distribution in each rating using Polarity.

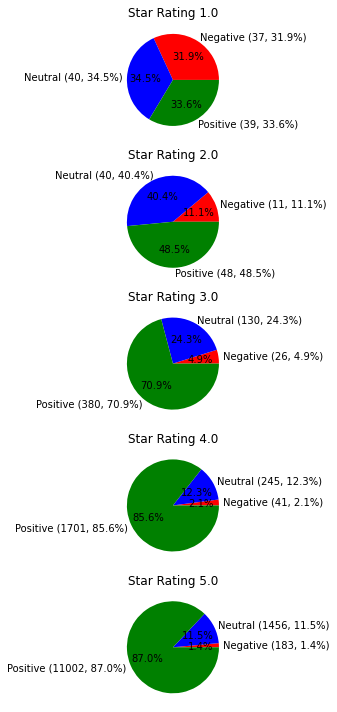


Figure 15: Sentiment distribution of each rating using VADER lexicon.

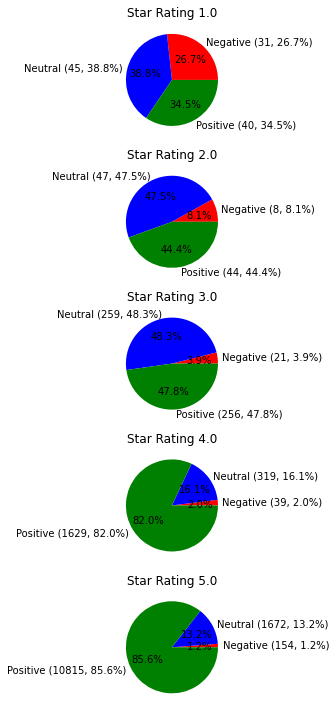


Figure 16: Sentiment distribution of each rating using AFINN-111 lexicon.

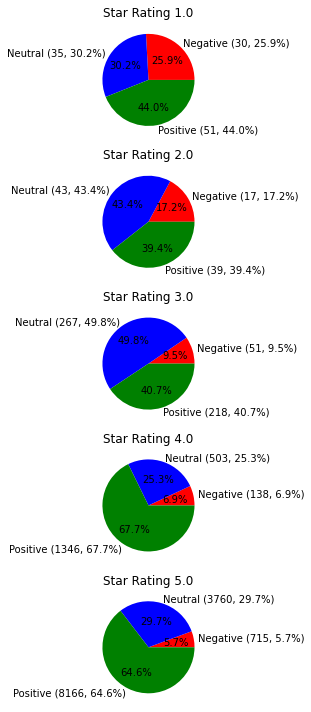


Figure 17: Sentiment distribution of each rating using SentiWordNet lexicon.

**B2: After Sparsity reduction**

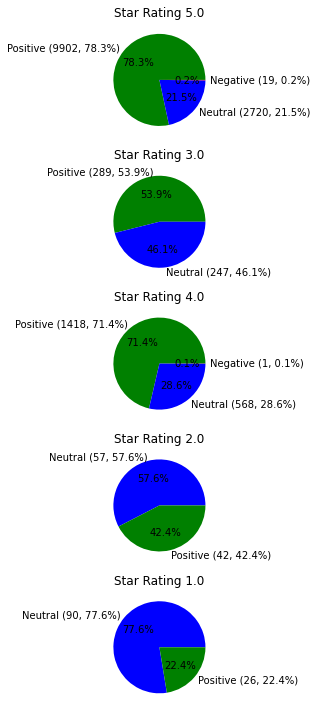
****

Figure 18: Sentiment distribution in each star rating using Polarity.

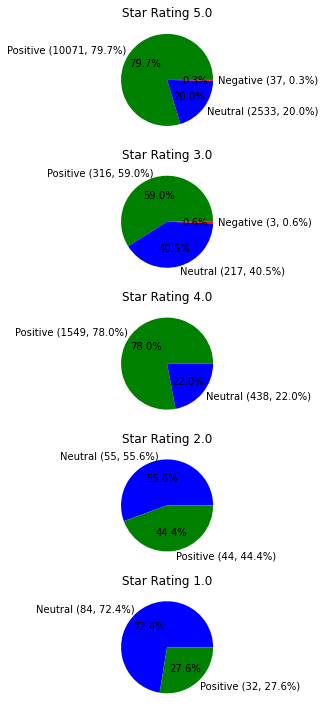


Figure 19: Sentiment distribution of each rating using VADER lexicon.

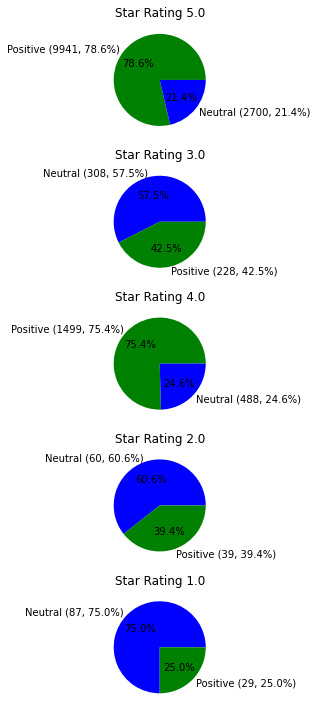


Figure 20: Sentiment distribution of each rating using AFINN-111 lexicon.

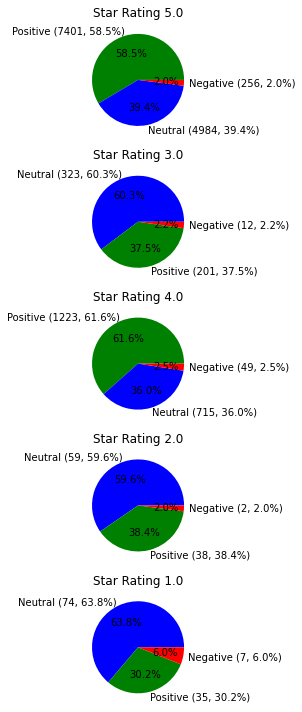


Figure 21: Sentiment distribution of each rating using SentiWordNet lexicon.

**Appendix 5: Correlation heatmaps**

1. **Before Sparsity Reduction**



Figure 22: Sentiment Label and Rating Correlation Heatmap

Figure 22 shows a correlation heatmap of Sentiment labels and rating. It is found that AFINN label (0.19) has higher correlation than VADER (0.13).

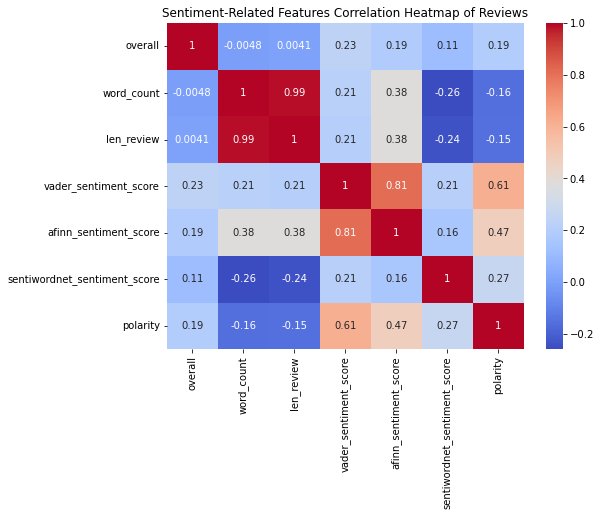


Figure 23: Text-derived features and Rating Correlation Heatmap

In Figure 23, heatmap of scores, polarity and other text derived features was plotted. It is clear that VADER score (0.23) showed higher correlation than other variables.

1. **After Sparsity Reduction**

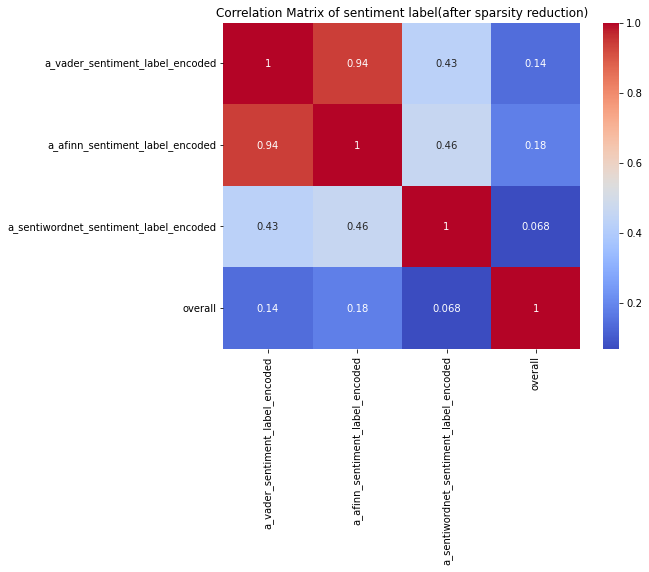


Figure 24: Sentiment Label and Ratings Correlation Heatmap

In Figure 24, again AFINN label (0.18) showed higher correlation than VADER label (0.14) with rating.

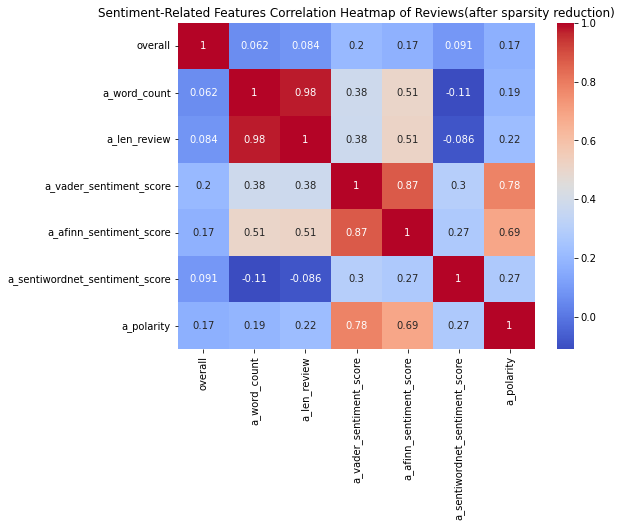


Figure 25: Text-derived features and Ratings Correlation Heatmap

Figure 25 depicts the correlation heatmap of text-derived features and rating. It was evident that, VADER score (0.20) shows higher correlation. However, it is less than that of before sparsity reduction. This might have lead to better regression performance before sparsity reduction.

**Appendix 6: Feature Importance**

1. **Before Sparsity Reduction**

**A picture containing screenshot, text, design

Description automatically generated**

Figure 26: Feature importance of text-derived features w.r.t. ratings

Figure 26 is plot of text-derived features against the target variable rating. It is evident that the Length of review (0.27) and VADER sentiment score (0.18) are important factors in predicting rating. However, other lexicons perform nearly equally. But as decided all models were run on VADER given the literature suggestions that VADER has higher accuracy in social media text.

1. **After Sparsity Reduction**

**A picture containing screenshot, text, display, design

Description automatically generated**

Figure 27: Feature importance of text-derived features w.r.t. ratings

Figure 27 is plot of text-derived features against the target variable rating. It is evident that the Length of review (0.3) and VADER sentiment score (0.21) are important factors in predicting rating. However, other lexicons perform nearly equally. But as decided all models were run on VADER given the literature suggestions that VADER has higher accuracy in social media text.

**Appendix 7: Results of Regression**

1. **Before Sparsity Reduction**

A blue and white rectangles

Description automatically generated with low confidence

Figure 28: Feature importance of features used in the Decision tree regression.

The coefficients were calculated for LR and RR models:

- For the Linear Regression model, the coefficients indicate that the vader\_sentiment\_score has a coefficient of 1.813101, polarity has a coefficient of 3.212788, len\_review has a coefficient of 0.014324, word\_count has a coefficient of -0.127167, and vader\_sentiment\_label has a coefficient of 0.165814.

- Similarly, for the Ridge Regression model, the coefficients are very close to the Linear Regression model, with vader\_sentiment\_score having a coefficient of 1.812706, polarity having a coefficient of 3.207304, len\_review having a coefficient of 0.014318, word\_count having a coefficient of -0.127154, and vader\_sentiment\_label having a coefficient of 0.166555.

1. **After Sparsity Reduction**

A blue and white rectangles

Description automatically generated with low confidence

Figure 29: Feature importance of features used in the Decision tree regression.

Similarly,

- For the Linear Regression model, the coefficients indicate that the vader\_sentiment\_score has a coefficient of 3.518580, polarity has a coefficient of 1.337395, len\_review has a coefficient of 0.094280, word\_count has a coefficient of -0.768266, and vader\_sentiment\_label has a coefficient of -0.152603.

- Similarly, for the Ridge Regression model, the coefficients are slightly different but still similar, with vader\_sentiment\_score having a coefficient of 3.477028, polarity having a coefficient of 1.346908, len\_review having a coefficient of 0.094346, word\_count having a coefficient of -0.768053, and vader\_sentiment\_label having a coefficient of -0.147997.

1. **Cross-Validation**

**A screenshot of a cell phone

Description automatically generated with low confidence**

Figure 30: Cross validation evaluation metrics

The table displays the performance metrics of different regression models on the dataset:

- Random Forest: RMSE = 0.218676, MAE = 0.060720, Adjusted R2 = 0.981684.

- Gradient Boosting: RMSE = 0.799433, MAE = 0.576478, Adjusted R2 = 0.755720.

- Linear Regression: RMSE = 1.398391, MAE = 1.196336, Adjusted R2 = 0.252684.

- Decision Tree Regressor: RMSE = 0.299570, MAE = 0.044379, Adjusted R2 = 0.965540.

- Support Vector Regression: RMSE = 1.264997, MAE = 0.827168, Adjusted R2 = 0.388419.

- Ridge Regression: RMSE = 1.398392, MAE = 1.196397, Adjusted R2 = 0.252683.

The metrics indicate the accuracy and goodness of fit of each model. Lower values of RMSE and MAE indicate better performance, while higher values of Adjusted R2 indicate better explanatory power of the model.

**Appendix 8: Exploratory Data Analysis**

This dataset consisted of 169781 number of reviews. There were 158 null values in the "reviewText" column and no null values in the "overall" column. Null value rows were removed from the dataset. Only verified reviews were further considered as they provided reliability and trustworthiness. Now dataset had 148787 reviews which were plotted against time and then recent data was utilized in the study (Annexure 1).

**A picture containing screenshot, rectangle, square, design

Description automatically generated**

Figure 31: Frequency of the product style

There were in all, eight styles, namelyMP3 Music, Audio CD, Vinyl, DVD, Audio Acsette, Blu-Ray Audio, Amazon Video, Blu-ray, and NaN (rows with no style). MP3 Music was the most reviewed product style.

After the data was cleaned and pre-processing, word count and length were visualized. Using bar chart and box plot.

A picture containing screenshot, rectangle, square, design

Description automatically generated

Figure 32: Frequency of reviews in each star rating

The output displays the count of reviews for each overall rating category. The rating category "5.0" has the highest number of reviews with 13,363, followed by "4.0" with 2,086 reviews, "3.0" with 573 reviews, "1.0" with 121 reviews, and "2.0" with 107 reviews. This information provides an overview of the distribution of reviews across different rating categories.

A picture containing colorfulness, screenshot, rectangle, line

Description automatically generated

Figure 33: Frequency of reviews

Reviewer ID was plotted against Review count and thus, it was found that lot of customers purchased frequently. There were 3525 unique reviewers in the dataset of 16250 reviews, indicating loyal customer base formed by Amazon’s Digital Music. However, new customers are not reviewing the product or are not created in the process.

A picture containing screenshot, rectangle, square, design

Description automatically generated

Figure 34: Frequency of length of reviews

The code analyses the length of each review by counting the number of characters. It reveals that most reviews have fewer than 500 characters. Additionally, the word count of the reviews is visualized. The analysis provides insights into the length and word count distribution of the reviews, highlighting the typical length of customer feedback.

A picture containing screenshot, rectangle, square, design

Description automatically generated

Figure 35: Frequency of word count in each review

The code counts the number of words in each review and creates a plot to visualize the distribution. The majority of the reviews had a word count ranging from 0 to 100 words. To investigate the outliers, box plots were used.

A picture containing white, rectangle, line, black

Description automatically generated

Figure 36: Boxplot of length and word count frequency

Visualization spots outliers in the data. There is one review with length of more than 6000, and a few having a length of more than 2000. Word count of some reviews was around 1000 too. Thus, word count was reduced to 20, to aid analysis.

|  |  |  |
| --- | --- | --- |
| **Word Count Statistics:**  count 15379.000000  mean 3.728786  std 3.603583  min 0.000000  25% 1.000000  50% 2.000000  75% 5.000000  max 20.000000  Name: word\_count, dtype: float64 | **Length Statistics:**  count 15379.000000  mean 21.566162  std 22.613738  min 0.000000  25% 8.000000  50% 13.000000  75% 27.000000  max 142.000000  Name: len\_review, dtype: float64 | **Summary of Rating:**  count 15379.000000  mean 4.751609  std 0.620935  min 1.000000  25% 5.000000  50% 5.000000  75% 5.000000  max 5.000000  Name: overall, dtype: float64 |

Table 4: Summary of dataset

The total size of dataset used further for all analysis was 15379. Mean word count was 3.72, mean length 21.56 and mean rating was 4.75.

A close up of words

Description automatically generated with low confidence

Figure 37: Word Cloud of the dataset.

Dominant words included great, love, song, excellent, music, good, classic, awesome, nice, best, etc.

After conducting the TF-IDF and sparsity reduction dataset was visualized again.

A picture containing screenshot, pixel, silhouette, design

Description automatically generated

Figure 38: Length of reviews

Review lengths were visualized and the majority of the reviews had a length of approximately 0-20 characters. Additionally, the word count was observed to be up to 10 words. The application of sparsity reduction techniques resulted in a significant reduction in the data size.

A picture containing screenshot, pixel, design

Description automatically generated

Figure 39: Word Count of reviews

A picture containing rectangle, line, screenshot, design

Description automatically generated

Figure 40: Bixplot of length and word count

Visualization reveals that there are no significant outliers and thus data was ready for further analysis.

A close up of words

Description automatically generated with low confidence

Figure 41: Word Cloud after sparsity reduction

Most common and dominant words were great, song, love, best, really, classic, album, excellent, version, etc.

**REFERENCES**

Blei, D.M., Ng, A.Y. and Jordan, M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, pp.993–1022.

Bonta, V., Kumaresh, N. and Janardhan, N. (2019). A Comprehensive Study on Lexicon Based Approaches for Sentiment Analysis. Asian Journal of Computer Science and Technology, 8(S2), pp.1–6. doi:http://dx.doi.org/10.51983/ajcst-2019.8.S2.2037.

Breiman, L. (2001). Random Forests. *Machine Learning*, [online] 45(1), pp.5–32. doi:https://doi.org/10.1023/a:1010933404324.

Chen, T., Samaranayake, P., Cen, X., Qi, M. and Lan, Y.-C. (2022). The Impact of Online Reviews on Consumers’ Purchasing Decisions: Evidence From an Eye-Tracking Study. *Frontiers in Psychology*, [online] 13. doi:https://doi.org/10.3389/fpsyg.2022.865702.

‌Chevalier, J.A. and Mayzlin, D. (2006). The Effect of Word of Mouth on Sales: Online Book Reviews. *Journal of Marketing Research*, 43(3), pp.345–354. doi:https://doi.org/10.1509/jmkr. 43.3.345.

Ghose, A. and Ipeirotis, P.G. (2011). Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics. *IEEE Transactions on Knowledge and Data Engineering*, 23(10), pp.1498–1512. doi:https://doi.org/10.1109/tkde.2010.188.

Gopal, R, 2023, Text Data Structures, lecture notes, Text Analytics IB9CW0, University of Warwick, Warwick Business School, delivered 3 May 2023.

Hoerl, A.E. and Kennard, R.W. (1970). Ridge Regression: Applications to Nonorthogonal Problems. *Technometrics*, 12(1), pp.69–82. doi:https://doi.org/10.1080/00401706.1970.10488635.

Hutto, C.J. and Gilbert, E. (2015). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. In: *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*. Ann Arbor, MI.

Li, Y., Li, T. and Liu, H. (2017). Recent advances in feature selection and its applications. *Knowledge and Information Systems*, 53(3), pp.551–577. doi:https://doi.org/10.1007/s10115-017-1059-8.

Liu, B. (2012). *Sentiment analysis and opinion mining :Synthesis lectures on human language technologies*. 1st ed. [online] Morgan & Claypool Publishers. Available at: https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf [Accessed 15 Jun. 2023].

Montgomery, D.C., Peck, E.A. and Vining, G.G. (2015). *Introduction to Linear Regression Analysis*. 6th ed. [online] New York, NY: John Wiley & Sons. Available at: https://content.e-bookshelf.de/media/reading/L-16125104-1a3a7c5bd1.pdf [Accessed 17 Jun. 2023].

Mudambi, S. and Schuff, D. (2010). Research Note: What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com. *MIS Quarterly*, [online] 34(1), p.185. doi:https://doi.org/10.2307/20721420.

Ni, J., Li, J., McAuley, J. (2019) Empirical Methods in Natural Language Processing (EMNLP)

Nishat, M.M., Faisal, F., Jahan Ratul, I., Al-Monsur, A., Ar-Rafi, A.M., Nasrullah, S.M., Reza, M.T. and Khan, M.R.H. (2022). A Comprehensive Investigation of the Performances of Different Machine Learning Classifiers with SMOTE-ENN Oversampling Technique and Hyperparameter Optimization for Imbalanced Heart Failure Dataset. *Scientific Programming*, 2022, pp.1–17. doi:https://doi.org/10.1155/2022/3649406.

Pang, B. and Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval*, [online] 2(1–2), pp.1–135. doi:https://doi.org/10.1561/1500000011.

Statista (2022). *Music Streaming - Worldwide | Statista Market Forecast*. [online] Statista. Available at: https://www.statista.com/outlook/dmo/digital-media/digital-music/music-streaming/worldwide [Accessed 22 Jun. 2023].

Wang, H., Lu, Y. and Zhai, C. (2010). Latent aspect rating analysis on review text data. In: *Knowledge Discovery and Data Mining*. doi:https://doi.org/10.1145/1835804.1835903.