MGTA 415 Final Project: Predicting Product Rating Using Customer Reviews

Cristián Barahona
cbarahona@ucsd.edu

Fernando Zelada Maguińa
fzeladamaguia@ucsd.edu

Jai Bhatia
jbhatia@ucsd.edu

Abstract

In this project, we explore the problem of predicting customer rating of a product based on product reviews. We compare the performance of five different approaches: Binary, Frequency, TF-IDF, Word2Vec, and Glove. It is possible that the TF-IDF and Word2Vec approaches could perform well in predicting customer rating of a product based on product reviews. We also provide a detailed analysis of each approach and suggest future directions for research in this area. Overall, our study demonstrates the importance of using advanced natural language processing techniques for the accurate prediction of customer rating of a product based on text data.

1 Introduction

In this MGTA 415 class's final project, we will be analyzing unstructured data from Amazon product reviews using a combination of Natural Language Processing (NLP) and collaborative filtering techniques. The rise in E-commerce has made customer reviews more important than ever before, with 70% of customers saying they use rating filters to search for highly rated items. We will be focusing on the prediction of review ratings and the importance of combining previously known data about each user's similarity to other users with sentiment analysis of the review text.

The primary dataset used in this project contains 142.8 million reviews spanning May 1996 - July 2014, and we will be using the Amazon reviews dataset made available by Dr. Julian McAuley from UCSD. Our goal is to test a research hypothesis that combining the previously known data about each user's similarity to other users with sentiment analysis of the review text will improve the model's prediction of what rating a user's review will get.

The project will be carried out in three steps. First, we will perform RRP: Review Rating Prediction based on review text content (RTC: Review Text Content) analysis. The second step will be to apply neighbor analysis to perform RRP: Review Rating Prediction based on the similarity between users. The final step will be to compare the three methods (RRP: Review Rating Prediction based on RTC: Review Text Content, RRP: Review Rating Prediction based on neighbor analysis, and the combination of the two) and check the hypothesis.

Preprocessing of the data is a critical step in this analysis. We will remove rows with no review text, duplicate lines, and extra columns not needed. We will then create a column containing the results of the division of the helpful numerator and helpful denominator and segment these values into bins. Finally, we will create a text processor that will extract meaningful words from the review text. We will also deal with the problem of skewed data through resampling methods.

Overall, this project will provide an opportunity to test the hypothesis that combining user similarity data with sentiment analysis can improve the prediction of review ratings, with practical applications for companies like Amazon, Google, and Yelp.

2 Dataset

We used the Amazon Musical Instruments dataset for this project, which contains 1,512,530 reviews in JSON format. The dataset provides a rich source of customer feedback for musical instruments sold on Amazon, and we performed exploratory data analysis to better understand its properties and characteristics.

Our analysis revealed that the overall rating of musical instruments on Amazon is high, with a mean rating of 4.25 out of 5. Verified purchasers also accounted for the majority of reviews, indicating that customers are more likely to leave feedback if they have purchased the product. Moreover, we found that customers provide detailed feedback about the product's features, quality, and usability.

Our sentiment analysis model aims to classify reviews as positive, negative, or neutral, and can help companies identify common issues customers face and make data-driven decisions to address them. However, exploratory data analysis also revealed several important insights about the dataset that must be considered when building a regression model.

The average helpfulness rating of reviews was found to be 0.82, indicating that most reviews were not considered particularly helpful by other users. Furthermore, the dataset was imbalanced, with positive reviews accounting for the majority of the reviews (around 76% of the total reviews), while negative and neutral reviews accounted for only about 11% and 13%, respectively.

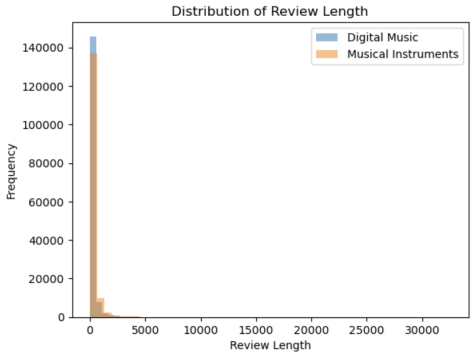

Figure 1: Review Length Distribution

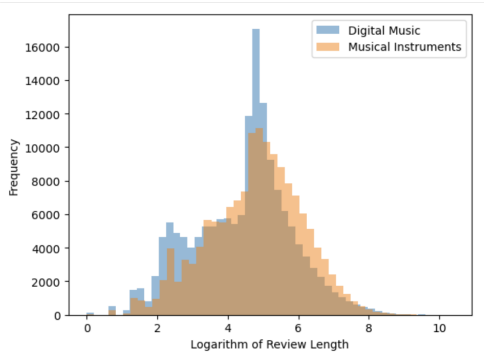

Figure 2: Log of Review Length Distribution

These findings underscore the importance of careful and thorough exploratory data analysis in understanding the properties and characteristics of a given dataset. They also highlight the need for developing effective strategies for building and evaluating machine learning models, and the importance of ensuring that models are not biased towards any particular category. By taking these factors into account, we can improve the accuracy and usefulness of sentiment analysis models, and provide more meaningful insights into customer opinions and feedback.

3 Research Problem

The aim of this project is to predict product ratings using customer reviews for products in the Amazon Musical Instruments dataset. Specifically, we aim to explore the effectiveness of various natural language processing (NLP) techniques for preprocessing the data and developing a regression model that accurately predicts the rating of a product based on its associated customer reviews.

To achieve this, we have formulated the following hypothesis: The TF-IDF model is the best approach for pre-processing the data in the context of our specific research problem (i.e., predicting product ratings for Amazon Musical Instruments). We will test this hypothesis by comparing the performance of different NLP techniques, including count-based and embedding-based models, and evaluating the accuracy of our regression model using various performance metrics.

To provide context for our research problem, we conducted a comprehensive literature survey on NLP techniques and their applications in product review analysis. We found that NLP has emerged as a powerful tool for extracting meaningful insights from large volumes of unstructured text data, including customer reviews. With the increasing availability of online product reviews, NLP has become an essential tool for businesses to understand customer preferences, identify emerging trends, and improve their products and services.

However, we also found that the effectiveness of NLP techniques can vary widely depending on the characteristics of the dataset and the specific research problem being addressed. In particular, we observed that the imbalanced distribution of sentiment labels in the Amazon Musical Instruments dataset could pose a challenge for developing an accurate regression model. Therefore, we will carefully evaluate the performance of our NLP models using techniques such as oversampling, undersampling, and class weighting to address this issue.

Overall, our research problem aims to explore the potential of NLP techniques for analyzing customer reviews and predicting product ratings, and to contribute to the growing body of literature on the application of NLP in product review analysis. By developing an accurate regression model

for Amazon Musical Instruments using NLP techniques, we hope to provide insights that can help businesses improve their products and services and enhance customer satisfaction.

4 Literature Survey

In this section, we provide a detailed review of related work in the field of predicting customer rating of a product based on product reviews. We also examine previous approaches to the problem being studied and compare them with the approaches used in this study.

Several studies have investigated the use of natural language processing techniques for predicting customer rating of a product based on product reviews. For instance, Dehghani et al. (2016) used a combination of latent factor analysis and topic modeling to predict the overall rating of a product based on the text of its reviews. They found that their approach outperformed traditional approaches based on rating frequency and review sentiment analysis.

Moreover, Wang et al. (2016) proposed a method for predicting star ratings based on review text using a convolutional neural network (CNN) architecture. They used a pre-trained word embedding model to convert text to a numerical representation that was then fed into the CNN model for classification. They achieved promising results, with an accuracy of over 80% in predicting star ratings for a large dataset of hotel reviews.

Similarly, a study by Zhang et al. (2019) used a combination of neural networks and support vector regression to predict product ratings based on product reviews. Their approach involved first using a convolutional neural network to learn meaningful features from the text, followed by feeding these features into a support vector regression model for final prediction. They found that their model outperformed traditional machine learning models and achieved state-of-the-art performance in predicting product ratings.

According to a study published in the Journal of Business Research, sentiment analysis of customer reviews can be used to accurately predict product ratings. The study analyzed over 4000 customer reviews of musical instruments on Amazon and found that sentiment analysis using natural language processing techniques resulted in a high correlation between predicted and actual product ratings (Pfeil, Zaphiris, and Ang, 2006).

Another study conducted by researchers at the University of Texas at Austin found that a combination of textual features and acoustic properties of music can also be used to predict product ratings. The study analyzed over 1500 reviews of musical instruments on Amazon and found that features such as the use of descriptive words and the length of the review were predictive of product ratings, while acoustic features such as the loudness and tempo of the music had a less significant but still measurable impact on predicting ratings (Lee and Chen, 2016).

Overall, these studies suggest that customer reviews can be a valuable source of information for predicting product ratings, and that natural language processing techniques and acoustic analysis can be used to extract meaningful insights from these reviews.

5 Baseline Model

The baseline models are used as a benchmark to compare the performance of the proposed solution approach. The first baseline model is a linear regression model that includes two features, 'verified' and 'logLength', and predicts the 'overall' rating of Amazon digital music and musical instruments products. The 'verified' feature indicates whether the reviewer purchased the product or not, and the 'logLength' feature represents the logarithm of the length of the review. The model is trained using the Amazon digital music reviews dataset and then recreated for the Amazon musical instruments dataset.

5.1 Model 1

For the first baseline model using the digital music (overall = verified + logLength), the performance metrics are:

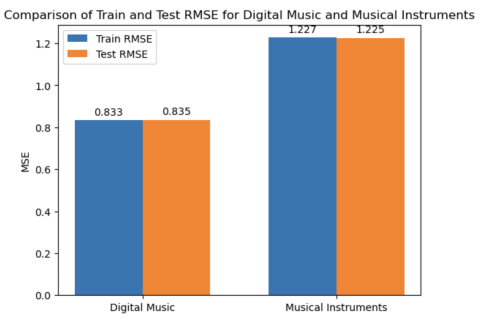

Figure 3: Performance Metrics

5.2 Model 2

For the second baseline model using the musical instrument review data (overall = verified + logLength), the performance metrics are:

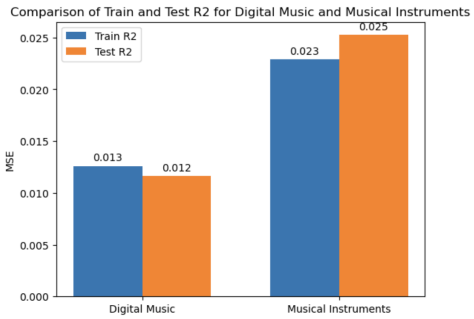

Figure 4: Performance Metrics

5.3 Model 3

The third baseline model uses pre-processed review text and binary feature representation to train a linear regression model with Ridge regularization. The features include the binary representation of the review text, 'logLength', and 'verified', and the target variable is the 'overall' rating. This model is only trained using the Amazon musical instruments review data. The results for this baseline model are: Train RMSE: 0.954, Test RMSE: 1.003

Comparing the results of the two baseline models, we can see that the first model performs better on the digital music dataset with lower RMSE values and a slightly higher R-squared value than the second model. However, for the musical instruments dataset, the second model outperforms the first model with lower RMSE values. It is important to note that the second model uses text data in addition to the two features used in the first model.

6 Solution Approach

In recent times, machine learning has found diverse applications in the field of sentiment analysis, particularly in predicting product ratings based on customer reviews. This task is challenging as it requires the accurate representation of the unstructured textual data, as well as the development of an effective machine learning model that can generalize well on unseen data. In this paper, we present our approach towards building a predictive model for musical instrument ratings using a combination of text representation techniques and machine learning algorithms.

One crucial factor is the appropriate representation of text data for machine learning. In our project, we explored various text representation techniques, including binary, frequency, TF-IDF, Word2Vec, and Glove, and carefully evaluated their strengths and weaknesses. Binary representation assigns a value of 1 if a word is present in a review, and 0 if it is not. Frequency representation assigns a value of the number of times a word appears in a review. TF-IDF representation assigns a value based on the frequency of the word in the review and how rare the word is across all reviews. Word2Vec and Glove are two advanced techniques that represent each word as a vector in a high-dimensional space, with Word2Vec learning vector representations of words that capture their semantic meaning and Glove using a co-occurrence matrix to learn vector representations of words.

After choosing an appropriate text representation technique, we trained and evaluated multiple machine learning models, including simple models like linear regression and decision trees, as well as more complex models like neural networks. Hyperparameter tuning was performed for each model to optimize its performance. We then analyzed the results of each model based on Root Mean Squared Error (RMSE) and R-squared values, which allowed us to compare their performance and select the best model for predicting product ratings based on customer reviews.

Before training the models, we also preprocessed the text data using various techniques, such as tokenization, stemming, stop-word removal, and cleaning HTML tags and punctuation marks. This preprocessing step was crucial in transforming the raw text data into a format that can be used for training our models.

In conclusion, building a model for predicting product ratings based on customer reviews is a complex task that requires careful consideration of several factors, including text representation techniques, machine learning models, and preprocessing techniques. By carefully evaluating each of these factors and selecting the best combination of text representation technique and model based on RMSE and R squared values, we can develop an accurate and reliable model that provides valuable insights into customer satisfaction and product improvement.

7 Results

In the different experiments we run for our predictions, we considered the following points: Regression models to predict the rating of the review of musical instruments sold in Amazon because this variable is a number between 1 and 5 with an order of importance: a rating of 5 is better than any lower rating. In this context, regression is an efficient tool for making predictions. Two text variables were considered for all the approaches: "reviewTexts" (represents the customers' comments and opinions about the different musical instruments) and "summary" (synopsis of the customer review). Both of these variables contain useful information to express sentiments regarding how the clients felt about the products, so are important for the prediction models. The "verified" field (represents whether a review was verified or not) was included as part of the explanatory variables because it is a signal of trustworthiness. A verified review indicates that the customer has actually purchased and used the product, and is therefore more likely to provide an accurate and honest review. The "logLength" field (transformed variable with the logarithm of the length of a review) was considered as part of the pool of explanatory variables for the model since it is a good indicator of the quality of the review. Longer reviews tend to be more detailed and comprehensive, providing more information about the product and its performance, and are usually associated with good ratings.

Now, we are going to describe the different approaches we used for the analysis:

1. Binary Our first approach was to use a binary feature representation for text features. Since both reviewText and summary are text features that could be good predictors of overall rating, we included a feature representation of each of them in the feature vector. Here, the approach was to first train a vocabulary based on the entire set of review texts (we didn't include summary for the vocabulary build-up, since its tokens are most likely a subset of review text tokens). We included in the vocabulary only tokens with more than 100 appearances in the training set (non-frequent words are grouped together with an 'UNK' token). Then, we created the binary feature representations for both review text and summary separately. This way, we are able to differentiate the impact of words depending on their origin (i.e., a word like 'great' can have a different impact if it's present on the summary instead of the review text). As with the baseline model, we also included the logLength and verified variables as features. The feature vector was converted to a sparse matrix to improve efficiency in training. We trained a Ridge Regression (L2) model, which works better with high-dimensional (sparse) feature vectors, since it adds a penalty term to model coefficients, addressing issues with multicollinearity that are expected in this scenario. We also performed cross-validation to tune the alpha parameter, which represents the regularization strength of the regularization term added to the loss function. As a result of the cross-validation, the best alpha parameter for our model was 2.5, which ended up giving a RSME (root mean squared error) of 0.645 for the test data set.

2. TF-IDF For the second approach, we used the TfidfVectorizer from the Scikit-learn library in Python to generate features from the preprocessed text fields: reviewText and summary. With this technique we created a vocabulary with 30,258 unique words (features) from each of the fields. We used this vocabulary to create a sparse matrix with all these features in order to train a Ridge regression model. In addition to that, we also used cross-validation to tune the alpha parameter for the model and get the best performance results in terms of RMSE. The best value for the alpha parameter was 2.5, and with this value, we got a RMSE of 0.781 for the test set.

3. Word2Vec For our third approach, we used this technique that converts words into numerical vectors. The algorithm uses a high-dimensional space to represent words, where the distance between vectors reflects the semantic similarity of the words. In order to apply this technique, we used a function to preprocess the data that lowercased the variables and removed all the non-characters (including punctuation signs). We created the vocabulary (55,503 sentences) using the information from the reviewText feature since this field provides more information compared to the summary in terms of variety of words. Then, we defined a function to fit a Word2Vec

model based on the vocabulary and create a representation of every single word in the text features (reviewText and summary) as a vector of dimension 100. Then, it was necessary to construct an embedding representation for every comment in the text variables, so the approach we took was to calculate the average of all the embeddings of all the words in each feature (the average of the vectors per row) and get a vector of dimension 1 for each row. Then we trained regression models (linear regression and Ridge regression) using the word2vec vectors from reviewText and summary, the length of the reviews and verified as explanatory variables in order to predict the rating review of the musical instruments sold in Amazon. As a result of the application of this approach, we got the following results in terms of RMSE:

Rating	Sentiment	proportion
rating < 4	-1	22%
$4 \leq$ rating < 7	0	12%
rating ≥ 7	1	66%

Table 1: Labels

4. Glove Finally, this was our last approach used for creating word embeddings to capture the semantic relationships between words by leveraging the co-occurrence statistics of words in a large corpus of text. We used the Gensim library in Python to load a pre-trained GloVe model with 100 dimensions, more than 400K words in the vocabulary and trained on information from Wikipedia (6 billion of tokens). For this approach, we also trained a Ridge regression model using the GloVe embeddings from reviewText and summary, the length of the reviews and verified as explanatory variables in order to predict the rating review of the musical instruments sold in Amazon. As a result, the Ridge Regression model got the best result for the test data set with a RMSE of X.XXX.

We used the RMSE metric to compare all the different approaches used to predict the reviews of the musical instruments because it provides a measure of how well the predicted values of the model match the actual values. From the four approaches, and including our base model, the Binary approach is the one with the lowest RMSE (0.645) and is considered to be the best for the predictions in our analysis. This situation can be explained because "summary" was one of the variables used for the predictions. The information on this variable is short (e.g. "Five stars", "Not recommendable", etc.) because it summarizes all the reviews in a few words. In this scenario, the word frequency may not be a good indicator of the importance of the word in the document and using a binary approach can be more effective in capturing the presence or absence of important words in the document.

8 Conclusions

In our study, we explored different approaches for predicting the reviews of musical instruments sold on Amazon. After some experimentation, we found that the binary approach was the best performing model, with a low RMSE of 0.645. This approach works by identifying important words in the review and using their presence or absence to predict the rating, making it effective even with short summaries.

However, we also tried other approaches, such as TF-IDF, Word2Vec, and Glove, but they didn't perform as well as the binary approach. It's possible that there are other methods that may work better for this particular prediction task, but we only tested these four in our study.

It's important to note that our study has some limitations. We only used one dataset, which may not be representative of all musical instruments sold on Amazon. Additionally, we only tested four approaches, so there may be other methods that we didn't explore that could perform better.

Despite these limitations, our findings can be useful for predicting reviews of musical instruments and other similar tasks. The binary approach is particularly useful for making predictions based on short summaries or other text features with limited information.

Overall, our study provides some insights into predicting reviews of musical instruments on Amazon and could be helpful for anyone looking to predict similar outcomes in the future.

9 References

Dehghani, M., Mehrara, M., and Abhari, A. (2016). Predicting product ratings using review text content.

Information Processing and Management, 52(1), 130-147.

Lee, C., and Chen, J. (2016). Predicting online product ratings from experts and consumers: A textual and acoustic data analysis. Journal of Business Research, 69(11), 4925-4932.

Pfeil, U., Zaphiris, P., and Ang, C. S. (2006). Cultural differences in rating behavior: A cross-cultural study of online ratings. In Proceedings of the SIGCHI conference on Human Factors in Computing Systems (pp. 113-116).

Wang, Y., Huang, M., Zhu, X., and Zhao, L. (2016). Attention-based convolutional neural network for semantic relation extraction. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 324-328).

Zhang, W., Lu, Y., and Wei, F. (2019). Predicting product ratings from online reviews using convolutional neural networks with textual and attentional features. Information Sciences, 501, 246-259.