Text Mining Final Project

Group 20

Text Mining on Amazon Musical Instruments Reviews

Gorden Li 489716, Yang Shen 490854, Ye Wang 491861,

Justin Jia 491837, Chuling Chen 489695

1 Executive Summary

This project aims to construct a model that can utilize text comments on a product to estimate both the product rating score and comment helpfulness score. The model is expected to have two kinds of applications.

1) Categorize musical instrument consumers into three groups (entry-level music lovers, professional instruments users, and guitar players) in order to have better precise marketing on different segments.

2) Estimate product scores and comments helpfulness estimates. The estimation can be applied to e-commerce website product score estimation, evaluation on the helpfulness of customer comments, and sentiment analysis of video website reviews.

The training dataset comes from Kaggle containing comments on Amazon musical instruments. Before import to the model, the dataset has been repossessed with upsampling and downsampling to balance the value distribution.

The model uses topic modeling to identify customers into 3 different segments. With the test results of 20 different combinations of tokenizers and classifiers in the supervised model, the final method is selected as a combination of TF-IDF Uni+BiGram and Poly SVC, which reaches an accuracy of 90% without high overfitting issues. Detailed accuracy of different combinations of tokenizers and classifiers are visualized as a colored cross table in the appendix.

2 Data Description

2.1 Dataset Info

The dataset we chose was Amazon musical instruments reviews, which were retrieved from Kaggle. The dataset was collected and uploaded two years ago by Eswar Chand. The dataset contained attributes like reviewer ID, product ID, product rating, the text of the review, helpfulness rating of review, review summary, and review time, including 10,255 unique reviews from 1,429 unique reviews.

The weblink of dataset is:

https://www.kaggle.com/eswarchandt/amazon-music-reviews?select=Musical_instruments_reviews.csv.

2.2 Possible Shortcomings and Solutions

2.2.1 Sample Size

Considering that the 10,255 unique reviews might be insufficient for high statistical power. The solution for this issue is using a k-fold validation to improve the performance. By using the k-fold method, we can maximize the use of our available

data in this data set.

2.2.2 Bias in the Score Rating

Also, the distribution of product rating scores seems to be biased. In the subsequent data processing process, we will introduce how the usage of both up-sampling and down-sampling helps rebalance the data. Besides, in order to minimize the effect of bias, we also included the helpfulness rating of reviews as compensation.

3 Project Objectives

There are two main objectives focused on current situation analysis and future use respectively. The first one is to find out typical features for purchased customers based on all reviews received, which might shed light on customers' behavior patterns and potential business opportunities. This result will provide direct guidance to the next-step advertising and promotion strategy.

The second one is to build a model that can estimate the customer's product rating score and helpfulness score in a diverse scenario without actual scores or labels:

1) Comment value evaluation. This is commonly used in E-commerce websites (Amazon, eBay, etc.). The platform could prioritize popular comments by making adjustments in online display and recommendation, which helps better understand the product and then increase sales.

2) Sentiment Prediction, which is mainly used on video platforms (YouTube, TikTok, etc.). Based on comments in a non-survey environment, the prediction results allow us to have a better command of a product property.

3) Establish label data storage. For analytic companies, the model can generate text labels for no-rating reviews for further use.

4 Methodology

4.1 Data Preprocess

As mentioned, the Amazon Musical Instrument dataset includes 10,261 samples, recording 10,261 comments in text toward customers' shopping experiences. The rating ranges from 1~5. The higher the number is, the more pleasant the customers' experience is.

Sample imbalance is an important problem in data preprocessing. According to the sample distribution in Figure 1, the dataset is highly imbalanced since the **comments with a rating of 5 outnumber the rest**. It will cause a problem in supervised learning models since model performance depends on the quality of data input. If we do not solve the imbalance problem, the model will be designed for detecting rating 5 comments and in a non-survey environment, it will be no difference with merely 'guessing'.

To balance the samples, **up-sampling** and **down-sampling** are used to augment the data.

4.1.1 Up-sampling

In up-sampling, the task is to increase the sample size of rating 1~4. To fulfill this need, **back translation**, which is a regular NLP text augmentation method, is applied. The result of the back translation is as follows.

Original Sentence:

1. The improvement over the old formulation is noticeable.

Sentence After Back Translation:

1. The improvement is obvious compared with the old formula.
2. This is a significant improvement over the previous recipes.

3. Improvements in old formulations are remarkable.

We can see that after back translation, keywords in the sentence are randomly replaced with the **synonym** and the structure **of the sentence** will also be changed randomly.

4.1.2 Down-sampling

Meanwhile**,** comments with a rating of 5 are **randomly deleted** to make the dataset more balanced regarding the distribution of ratings 1~5.

After data augmentation, the distribution of the dataset is shown in Figure 2.

4.2 Topic Modeling

Based on the pre-processing data, we applied the topic model (via Latent Dirichlet Allocation) to analyze the main topics of reviews in the dataset.

First, we normalize and vectorize the reviews by Bag-of-Words vectorizer. Then, we apply the LatentDirichletAllocation function to analyze the topics. Here, we keep doc_topic_prior(parameter of Dirichlet distribution for topics) and topic_word_prior(parameter of Dirichlet distribution for words in a topic) the same(0.25). We change n_components, which represents the number of topics in the model, from 2 to 4. We want to find the best topic model based on the perplexity score. The lower perplexity score is better. Finally, we use the best topic model to visualize our result.

In the visualization part, we assign parameter lambda to 0.8, which is better than 1 to show topic-relevant terms[1].

4.3 Supervised ML Models

Based on the augmented dataset, we are able to apply supervised machine learning models to train a customized model for this specific need.

As for the tokenizer, several different methods of tokenizing are used including word-bags, TF-IDF with Uni Gram, TF-IDF with Uni+Bi Gram, and TF-IDF with Bi Gram. All of the tokenizers consider the most frequent 1,000 words in the training set after removing stop words.

As for the model, multiple popular algorithms used for high-dimension data are used including SVC-support vector classifier, Random Forest, Decision Tree, ANN-artificial neural network.

By combining different **tokenizers** and **models**, we hope to find the best combination of tokenizer and algorithm to fit this sentiment analysis task using k-fold validation.

Pro:

1. A customized tokenizer and model might fit the task better compared to a pre-trained tokenizer.
2. Complicated non-linear models will fit the data better.

Con:

1. The performance of the data depends highly on the quality of the data.
2. Overfitting may exist. Thus when training the model, we might need to introduce regularization.

5 Result and Discussion

5.1 Topic modeling

Based on the perplexity score of each model, we find the model with topics has the lowest perplexity score (294). Based on this model, we visualize the topic as shown in Figure 4. It shows three topics and their top 10 most relevant terms.

We can see that topic 1 is mostly about the pedal, cable, and so on. They are about the most frequently used components. We think these users who give reviews related to topic

1 are very familiar with instruments and they can be called professional instruments users. They are our potential customers.

Topic 2 is more about the mic, strap, and other kinds of most frequently used components. We think these users are only familiar with the most common knowledge in instruments so they can be defined as entry-level music lovers.

Topic 3 is more about guitar, string, and other terms related to guitar. We think these users are guitar players and they are our potential guitar buyers. We can also know among those reviews, the guitar is the most popular instrument that many users talk about.

5.2 Performance of Supervised Models

The accuracy of different combination tokenizers and models on the test set is shown in Figure 3. According to the accuracy matrix, we found that:

1. Compared to word-bags, TF-IDF is a better way to tokenize in a comment environment.
2. Poly SVC does well in both high and low dimension tokens, reaching an **accuracy of over 90%**.
3. ANN can do better on high dimension token Uni+Bi Gram TF-IDF but will overfit on simple word-bags or Uni Gram TF-IDF.

Based on the model, managers do not need to label the rating manually in a non-survey environment like comments under Youtube videos regarding their product. Instead, the company can generate more text comment samples with ratings labeled using the model.

6 Conclusion

According to our result, the topic modeling clustered the customer into 3 different segments demonstrating a distinguished preference difference. Therefore, precision marketing based on classification preferences can be very helpful.

On the other hand, the score evaluation system has shown high accuracy in the test data set, so the performance of the scoring system actually used for score estimation is also guaranteed.

6.1 Current situation analysis

Based on topic modeling, the existing customers could be segmented into 3 groups: entry-level music lovers, professional instruments users, and guitar players. Therefore we propose some practical strategies for the Amazon platform to boost sales for the instrument category and to reduce costs for marketing activity.

1) Product selection list: create a detailed collection of instrument products based on each group's demand. For example, for entry-level music lovers, we could provide a commonly used list, such as strap and mic, to allow them quickly browse through and make purchase decisions. For professional musicians, we need to include more specific attributes and parameters in the list to help them make comparisons more efficiently. For guitar players, we provide a recommendation list for high-rated guitar accessories and encourage bundle sales.

2) Advertising and promotion: we could be more accurate and specific in sending advertising messages or holding promotion activities. For example, for guitar players, we could allow them to set up auto-delivery or send coupons for accessories that need to be replaced after a certain period.

6.2 Rating Score System

One of the most practical uses of our rating score system is comment evaluation. Besides the traditional post-purchase rating system, the era of new retail introduces more diverse channels such as comments on youtube, Facebook, or live commerce,

which do not have a score option to get feedback in a statistical way. The importance of our model could be realized in 2 ways:

1) Quantify non-rating reviews on social media platforms. We could use our score system to transfer text messages into specific rating numbers and then calculate overall performance in each social channel.

2) Optimizing mechanism of review display. For example, we prioritize positive and high-rated comments automatically in a live show, which provide more valuable information for potential customers and therefore increase the conversion rate.

Appendix

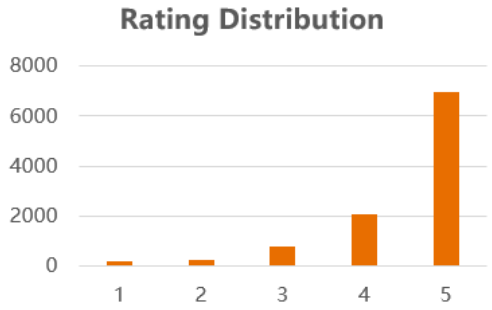

Figure 1. Sample distribution before augment

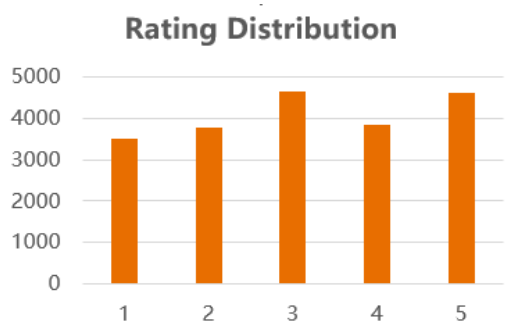

Figure 2. Sample distribution after augment

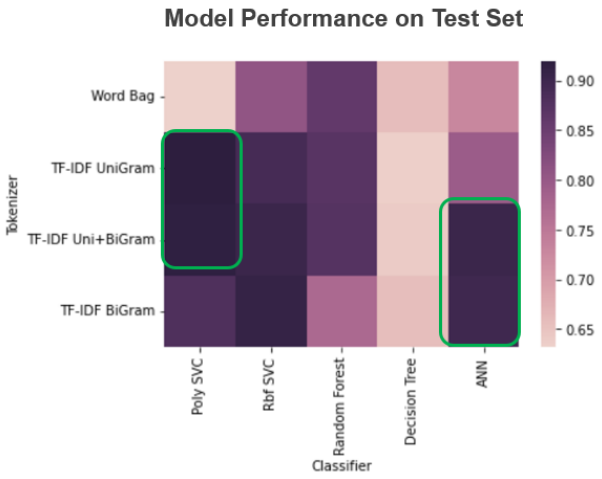

Figure 3. Model performance on test set

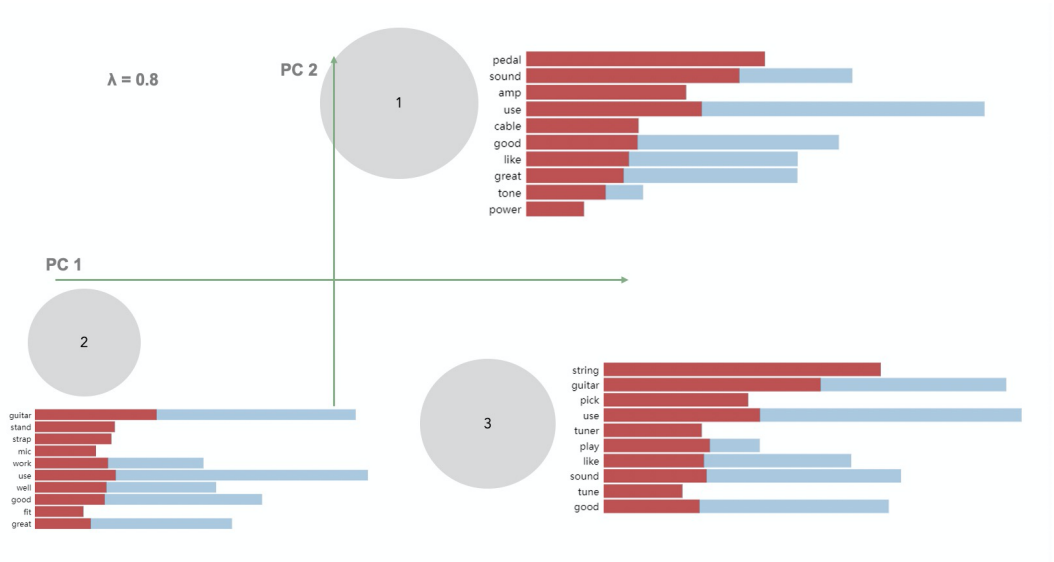

Figure 4. Three topics and their top 10 most relevant terms

Reference

1.Sievert, C., & Shirley, K. (2014, June). LDAvis: A method for visualizing and interpreting topics. In Proceedings of the workshop on interactive language learning, visualization, and interfaces (pp. 63-70).