Amazon Musical Instrument Reviews

Abhinav Duvvuru   
 Artificial Intelligence and Machine Learning  
Lambton CollegeToronto, Canada  
abhinavduvvuru@gmail.com  
  
Sandeep Reddy Alla  
Artificial Intelligence and Machine LearningLambton CollegeToronto, Canada  
itsmealla05@gmail.com  
  
SivaPriya TatiReddy  
Artificial Intelligence and Machine LearningLambton CollegeToronto, Canada  
sivapriya721@gmail.com   
  
Uma Maheswararao Gutta Artificial Intelligence and Machine LearningLambton CollegeToronto, Canada umamaheshgutta@gmail.com

*Abstract*— In this study, we analyze the reviews given by different customers who got the musical instrumental products. It is hard to go through each review line by line and judge whether the product is good or bad. Hence, we chose NLP technology sentiment analysis to analyze whether incoming messages or reviews are positive, negative, or neutral.

*Keywords used* — nltk, stopwords, spacy, TF-IDF, KNN, SVC

# Introduction

In this digital world, we come across various instrumental products. customers swipe across hundreds of product choices in each category, so in this case, for the customers, it isn't easy to choose the right product. Here the reviews come into the picture. Based on the feedback and reviews given by the customers who purchased the product already will leave a rating after using the product. So, with the help of NLP, we can automatically analyse customer feedback, Reviews, and social responses.

# Data description

The dataset used in this project is the musical instrumental data obtained from Kaggle, which has reviewer ID,User ID,Reviewer Name,Reviewer text,Overall,summary,unix review time,review time.The dataset has total 9 attributes and around 10,000 instances. Below are the list of attributes and their type:

1. reviewerID(categorical) - ID of the reviewer

2. asin(categorical) - ID of the product,

3. reviewerName(categorical) - name of the reviewer

4. helpful(categorical) - helpfulness rating of the review

5. reviewText (categorical)- text of the review

6. overall (numeric) - rating of the product

7. summary (categorical) - summary of the review

8. unixReviewTime (numeric) - time of the review (unix time)

9. reviewTime (categorical) - time of the review (raw)

## Dataset loading

The “Musical\_instruments\_reviews.csv” file can be loaded using Python Pandas and stored as a dataframe. The below figure 1 depicts the data in the dataset.

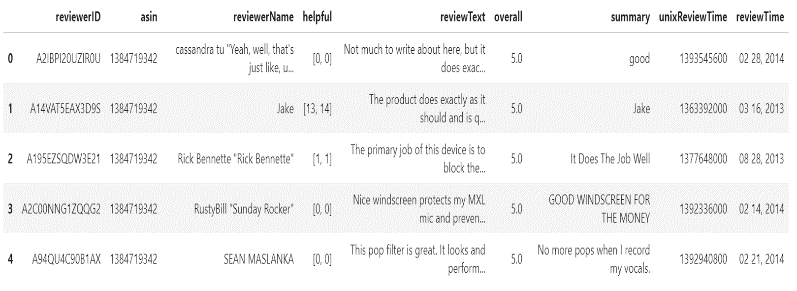


Fig. 1. Overview of dataset

The dataset has total 10261 instances and 9 attributes

# data extraction and cleaniing

The dataset which we use cannot be implemented to the model without proper cleaning and extraction. To apply sentiment Analysis, we do data preprocessing steps.

## Text Preprocessing

We have text data in “review” column of our dataset and explored the text data available in “review” column.

## Handling missing values

The isnull().sum() method in pandas shows the total number of missing values for each attribute.

The NULL values that are present in the “review” column are dropped for further analysis.

## Removing punctuation marks and stopwords

Punctuation marks are removed, and all the text is converted into lowercase letters and then applied the nltk word tokenize method to convert the text into tokens as show in figure 2.

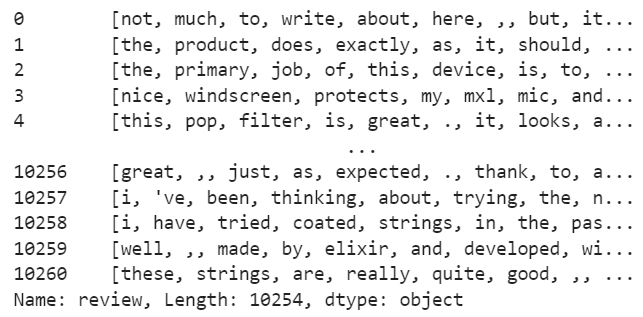


Fig. 2. Tokenized text

The tokenized words are used to apply the nltk stopwords to remove the stopwords from the text.

# Exploratory data analysis and visualization

We applied different visualization libraries like matplotlib and seaborn to analyze the data by plotting graphs and charts.

We are looking into the "Overall" column to check on how the ratings of the products are distributed.

From the figure 3 it is clear that majority of the reviewers have given 5 star rating followed by 4 star and 3 star.

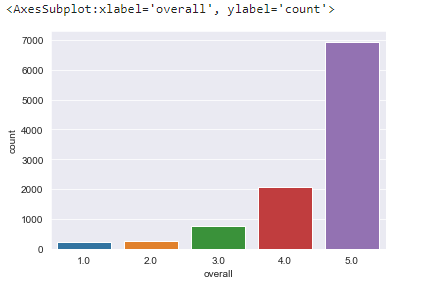
**

Fig. 3. Count of ratings given by reviewers

## Labelling the rating attribute

The overall attribute is renamed as “rating” attribute and ratings are labelled into 3 labels as following:

* Ratings 5 and 4 are considered as positive reviews and labelled as 2.
* Rating 3 is considered as neutral and labelled as 1
* Rating 0 and 1 are considered as negative reviews and labelled as 0

Figure 4 depicts the labelled ratings

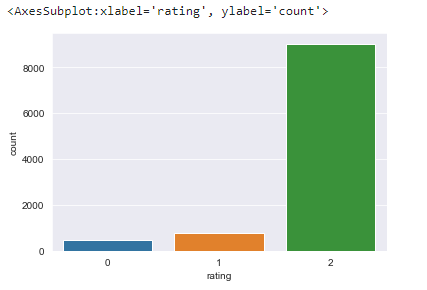
**

Fig. 4. Labelled ratings

## Visualizing text using Spacy

Parts of speech of the text can be visualized using the Spacy library as shown in figure 5.

*Text

Description automatically generated with low confidence*

Fig. 5. Parts of speech of the text

The named entity recognition of the text can also be found using the Spacy library as shown in figure 6.



Fig. 6. Named entity recognition of a single review text

# model building

After successfully processing text data. The reviews need to be converted into the sparse matrix before feeding to the classification model.

The data is split into the test and train data with test data as 30%.

We have applied various classification algorithms to predict the reviews. K Nearest Neighbors, Decision Tree Classifier, Support Vector classifier, Random Forest Classifier, Logistic Regression and Linear Support Vector Classifier were applied and predicted the model performance and evaluated using confusion matrix and classification report.

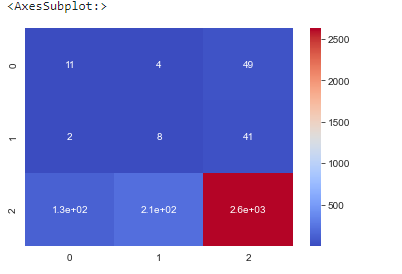


Fig. 7. Heatmap of confusion matrix of KNN classifier

The accuracy score of our data using K nearest neighbors classifier is 0.86

Applied linear support vector machine classifier and evaluated the model performance and plotted heat map as shown in below figure 8.

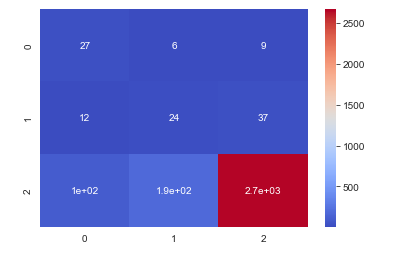


Fig. 8. Heatmap of confusion matrix of Linear support vector machine classifier

Got an accuracy score as 0.88 with linear support vector classifier. The classification report (figure 9) shows that the precision, recall and f1-score are better when compared to other classification algorithms.

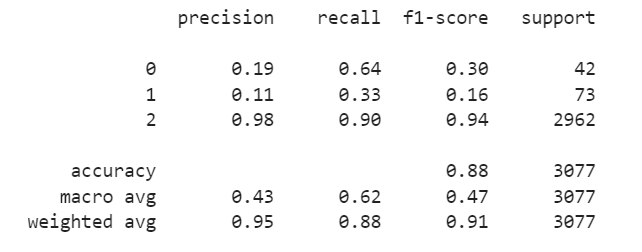


Fig. 9. Classification report of Linear Support Vector Machine Classifier

we further applied Random Forest classifier, Decision Tree classifier, Logistic Regression and got an accuracy score of 0.88

Figure 10 shows the accuracy score of various classification algorithms with x-axis as algorithms and y-axis as the accuracy score.

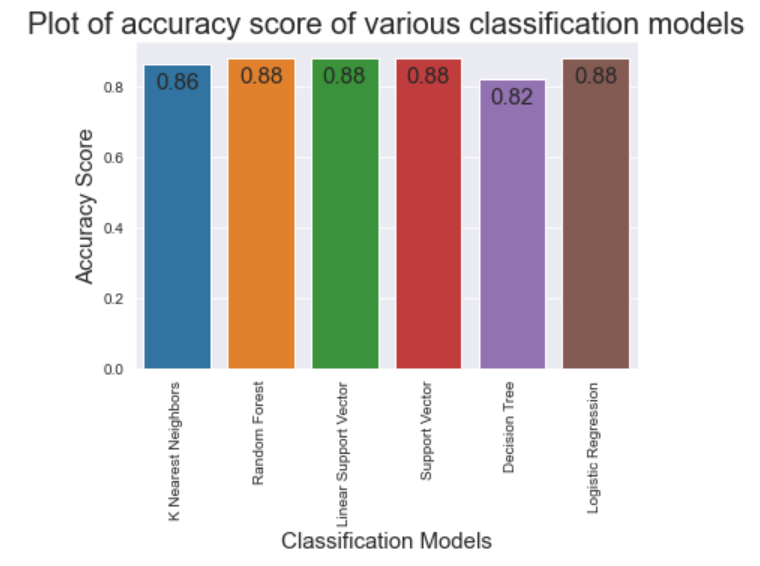


Fig. 10. Bar graph of accuracy scores of classification algorithms

# Conclusion

##### We can see from the Text Classification model performance, that all the models that we applied have performed with an accuracy of 80-88 percentage. All the models are performing extremely well in predicting the positive reviews of our data.

##### But most of the models have not predicted well with the neutral and negative reviews as our model have them in very less numbers.

##### Only Linear Support Vector Classifier performed well in predicting the positive, negative as well as neutral reviews. The accuracy of the Linear support vector model is 88% and we can see better results of recall and precision for negative as well as neutral reviews along with positive reviews.

##### So, on a positive note, we are able to achieve an accuracy of 88% with Linear Support vector classifier which is pretty good in predicting the instrument reviews.

##### References

1. “Amazon Music Instrument Dataset”. Kaggle [online dataset], November 2021. Available: Kaggle, <https://www.kaggle.com/eswarchandt/amazon-music-reviews> [Accessed: November 11, 2021]
2. “Sunil Ray”, “Understanding Support Vector Machine(SVM) algorithm”. Analyticsvidya [online document], September 2017. Available Analyticsvidya, <https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/> [Accessed: December 12, 2021]
3. “Linguistic Features”. Spacy [online document]. Available: Spacy, <https://spacy.io/usage/linguistic-features> [Accessed: December 12, 2021]
4. “Scott Robinson”, “K-Nearest Neighbors Algorithm in Python and Scikit-Learn”. Stackabuse [online document]. Available: Stackabuse, <https://stackabuse.com/k-nearest-neighbors-algorithm-in-python-and-scikit-learn/> [Accessed: December 12, 2021]