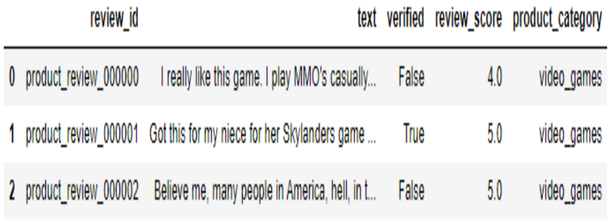
# Text classification - Amazon Reviews

Yogi Bhoot
*COMP-1804 - Applied Machine Learning*
001189309

***Abstract*—Over the last few decades, there are many e-commerce websites or markets that have been growing in popularity. In a growing company competition, it takes feedback from customers to figure out what to do to get their company up and making a good position in the market. As a result, millions of evolutions are generated every day by customers and making it difficult for genuine buyers to decide whether or not to purchase an item. This vast amount of reviews is difficult and hard to analyze to make a decision for sellers and merchants who sell their products on the online market and website. In the study, I use two different types of supervised machine learning models, Support vector classification and Logistic regression on product category and product rating from amazon. The SVC model performed on product rating VS reviews and the result showed a great accuracy rate. while the logistic model used for product category and accuracy rate is good.**

## I. Introduction and related work

- Nowadays, people prefer to buy online products from amazon. Also, it is one of the largest online buying and selling markets in the world. It not only sells one product but also sells many other products to related different categories.Also, Amazon provides feedback and rating service for people's opinions. It expands all over the world, so people give their reviews and ratings from the different regions for products. There is a vast amount of data stored in the amazon data store. Also, there are positive and negative opinions about the same product, so it makes it difficult for customers to buy products. Sentiment classification is used for determine the overall meaning and purpose to use of a text which can be negative, positive and neutral. It is challenging in making decision for merchants for increasing selling their products. In addition, The real challenges are similarity of word, handling large word in documents and organisation of ext in documents.
- With the rapid increase in the number of applications and upgrading technology, its data has become one of the most important information sources in the world. Text classification plays the most important role in the arrangement of text documents. Text classification is a multi-labelled problem.It is used for documents where more than one class problem. [1]
- Similar work has been already done in [2] and [3]. In number 2, they have already separate train and test data set for classification and did small cleaning processing including converting into lowercase and lemmatization for WordCloud then apply the logistic model and get more than 0. On the other hand, reference number 3, they have done deep cleaning processing including tokenization, stopword, Capitalisation, noise removal, slang and lemmatization etc. Also, they have applied around 15 models for classification and give an evaluation of all models.
- Our amazon review dataset is provided by the professor. It has 5 columns and 32918 rows. All columns have character values except the name of the review score column. The verified and product category has two different values false and true, video games and music instrument value respectively while, the review score has 5 6 different values -1.0, 1.0, 2.0, 3.0, 4.0 and 5.0 that is provided by customer satisfaction. Moreover, there are many Nan and null values in the dataset. Before applying the model, I need to clean some noise and not available value because the model runs on structure data which has an error-free dataset. So, I did pre-processing including dropping the null value, count review according to the category, removing punctuation, removing stop words and speaking voice, lemmatization, and applying 0 and 1 values for positive and negative values for word-cloud and vectorizer using term frequency-inverse document frequency. Then, apply both models according to their x train, x test, y train and y test data. Support vector classification used x value as review text and y as review score, while logistic regression used x value as a review text and y value as a product category.



| | review_id | text | verified | review_score | product_category |
|---|---|---|---|---|---|
| 0 | product_review_000000 | I really like this game. I play MMO's casually... | False | 4.0 | video_games |
| 1 | product_review_000001 | Got this for my niece for her Skylanders game ... | True | 5.0 | video_games |
| 2 | product_review_000002 | Believe me, many people in America, hell, in t... | False | 5.0 | video_games |

Fig. 1. DataSet

- SVC and NuSVC are slightly different sets of variables. Also, They are running on different types of mathematical formulas. LinearSVC is a faster implementation compared to other support vector classification when we use a linear kernel. The linear kernel is not used as a parameter in LinearSVC.There are three classifiers SVC, NuSVC, and LinearSVC take two parameters and both are arrays. One parameter is used for shape (n sample, n feature) for holding training data and the other is the class label. [4]

- Logistic regression is a supervised machine learning model that uses binary values for classification problems. Logistic regression is more efficient compared to linear regression. Also, there is no need for a linear relationship between input and output.
- Both models are using the same review text after completing the clean process. Both use two input parameters. In the SVC model, use the input variable that one is score value in integer and other is Review text in vector. The same as input in the logistic model, but one parameter is the product category which has character value and the other is the Review text in vector. I choose score integer in SVC because it performs on numerical value, while category does not work in this model. So, I decided to use categories in a logistic model.

## II. Ethical discussion

there are many ethical challenges for the amazon review dataset. Because there are many online markets in the real world which compete with each other for making a good place in the real world. if some case dataset is a leak, it will less, or use other company for targeting their customer for selling product. Also, It will be misused and sold to other companies with bad intentions. Moreover, the dataset is not worked properly because of error and null values with mixing other data types in data. For these reasons, the model can't predict the right value for using test data.

## III. Dataset preparation

The given dataset is a combination of integer and character values. In the dataset, I have only one integer value which is the review score in the CSV file. Then I checked Not Available and not a number value in the dataset which contains 5 columns and 32918 rows. All columns have character values except the name of the review score column. there are 906 and 20 null values in the product category and review text respectively. According to a given task, I need just a review score, review text and product category for performing to selected machine learning models. So, I have created a new data frame for which data I need including review text, rating and product category. Now move on to the data cleaning part. Firstly, I removed all null values that we detected before we mentioned the null value part and again performed the null value checking code to verify that is really deleted or not. Secondly, check the total number of reviews for only our two product categories which have video games and music instruments. Then I performed text length and stored it in the new variable called review length. In addition, the describe function gives all information about the maximum, minimum, average, quartile and total number length. I used this function for review length so we can choose a reducing method. After that, I removed the punctuation sign that is not convertible in the binary value. Also, remove the stop word which has no meaning for text clarification. After these changes, I show the bar graph rating Vs review text and product category Vs review text. Then, remove all speech word values. next, I used the lemmatization method to reduce the same meaning and store one word in the review. Then, I use true means 1 value and false values as 0. 1 value stored more than 3 rating reviews text and 0 value contains lower than 2 rating reviews And both values stored in the same column. Next, I will display a word cloud image. so, I use the if condition for1 value stored in one new data frame. In the same way, I stored negative words in a new data frame. These two data frames are applied to the word cloud function for creating images.

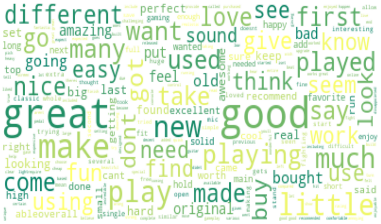


Fig. 2. Positive review cloud

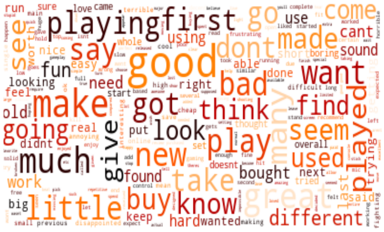


Fig. 3. Nagetive review cloud

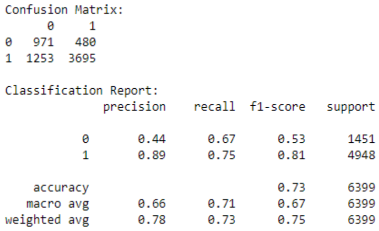
Then, I used the term frequency-inverse document frequency method for creating a vector that we convert into tokenizing. Lastly, this vector and another parameter we applied to the machine learning model that we used for prediction.

## IV. Methods

In our study, we performed two machine learning models one is to support vector classification and another is the Logistic regression model. Support vector classification takes two parameters which have our x train data and y train data.The SVC model is used to predict our review according to review rating. Also, I defined it using a linear kernel, balanced class weight, the probability is true and given a specific random state is 111 for good performance. After that, we applied our data set to fit the model. Here, it takes time to fit our data. Then, I perform a prediction on our fitting model. Lastly, I check out the prediction confusion matrix and classification report. I create a function for confusion and a classification method that gives both outputs. The logistic regression method takes two inputs which have our x train and y train data the same we used before for our support vector classification model. Here, we use x train data as a review text and y train. In this model, we use x test data as a product category. I don't use the product category in the SVC model because it does not convert into the binary value. so,I use logistic regression for the category. After, I fit our data into the model and predict the value. Then, I call the confusion and classification report function to get the metric and classification value. Also, I display ROC curves for both models.

## V. Experiments and Evaluation

This section presents the result and evaluation of the study. The percentage of testing data is calculated by the accuracy value which is classified by how fit and accurate our data is. Evaluation is predicted by our external data that is predicted by our model and shows the result how it is. The result of the support vector classification model shows a 73 percent accuracy rate and a positive review rate is higher than to negative review rate. Also, I take 10 reviews by myself to evaluate the model and the given result is correct.



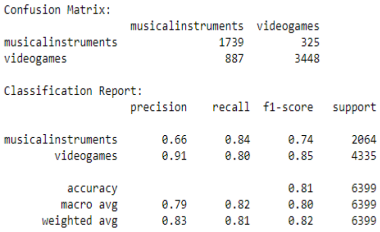
```
Confusion Matrix:
      0     1
0   971   480
1  1253  3695

Classification Report:
              precision    recall  f1-score   support

           0       0.44      0.67      0.53      1451
           1       0.89      0.75      0.81      4948

    accuracy                           0.73      6399
   macro avg       0.66      0.71      0.67      6399
weighted avg       0.78      0.73      0.75      6399
```

Fig. 4. SVC Result



```
Confusion Matrix:
                    musicalinstruments  videogames
musicalinstruments                1739         325
videogames                         887        3448

Classification Report:
                    precision    recall  f1-score   support

musicalinstruments       0.66      0.84      0.74      2064
        videogames       0.91      0.80      0.85      4335

          accuracy                           0.81      6399
         macro avg       0.79      0.82      0.80      6399
      weighted avg       0.83      0.81      0.82      6399
```

Fig. 5. Logistic Result

On the other hand, The result of the logistic regression model shows a 81 percent accuracy rate and a video games review rate is higher than the music instrument review rate. Also, I take 10 reviews by myself to evaluate the model and the given result is correct.

## VI. Discussion and future work

In this section, the main purpose of this study is to see selected machine learning models support vector classification and logistic regression to show accuracy in predicting the training data. Here, the given amazon video games and music instrument category reviews with rating score dataset. SVC model gives more accuracy for rating and the logistic model gives good accuracy for the product category.there are a few limitations of these two machine learning models. In future, we use a summary of the text review that gives more accuracy compared to the tokenize method. In reduce method, we can't get the same meaning word and also get the right word for doing this method. In future, these models are upgraded and use summarize word for better predicting data.

## VII. Conclusions

In the end, this study shows two different machine learning models that support vector classification and logistic regression models on the amazon video games and the music instrument and review score reviews. The study results showed that SVC accuracy is 73 percent for predicting reviews from rating scores, while logistic model accuracy is 81 percent for predicting reviews from product categories.

## References

[1] Chrystal, Jincy  Joseph, Stephy. (2015). Text Mining and Classification of Product Reviews Using Structured Support Vector Machine. Computer Science  Information Technology. 5. 21-31. 10.5121/csit.2015.50803.

[2] Vanikul, Rajath Akshay. "Sentiment-Classification-For-Product-Reviews." GitHub, 30 Oct. 2021, github.com/RajathAkshay/Sentiment-Classification-for-Product-Reviews. Accessed 26 Apr. 2022.

[3] kowsari. "Text-Classification/Code at Master · Kk7nc/TextClassification." GitHub, 8 Mar. 2020, github.com/kk7nc/Text-Classification/tree/master/code. Accessed 26 Apr. 2022.

[4] B. Scholkopf, C. J. Burges, A. J. Smola ¨ et al., Advances in kernel methods: support vector learning. MIT press, 1999.