Ratings Prediction Project

Submitted by

Akhil Vangeti

Acknowledgement

I sincerely thank Mr. Sajid Choudhary for providing me valuable information and guidance for completion of this project. His assistance has helped me a lot during the course of entire project.

I also thank datatrained institute for helping me gain hands on experience about machine learning.

I also thank the kaggle community for providing valuable insights intlo Natural Language Processing techniques.

Introduction

The technology in this era helps the people to do various task at their fingertips. The times such as covid outbreak has made it quite impossible for them to visit any retail store to purchase any product. This indeed has resulted in an enormous growth for the e-retailers.

The e-retailers provide convenience, wide variety and price discounts for attracting more and more customers but the problem which arise in this scenario is the quality which cannot be checked by the customer before making a purchase. The only option to solve this problem is to see the reviews of the product.

The ratings of the products provided by users will help the new customers make a purchase decision. There are many e-retailers which have reviews as well as ratings on a scale on 1-5, 1 being the worst and 5 being the good. In this project we will help an organization which has reviews but no ratings.

This indeed is planned by using natural language processing along with the machine learning algorithms to build a model which analyses the review text and gives the rating.

Analytics and Modelling

The steps followed in this study are as follows:

1) Data collection
2) Data Study and Analysis
3) Data preprocessing
4) Modelling

Data Collection:

The data for this study is collected from the online e-retailer Amazon.in.

The reviews and ratings are collected for the products such as laptop, phone, printer, smart watch, router, speakers etc;. The target for collection of data is 20000 data points with balanced ratings of 1 to 5.

The Selenium library is used for the scraping process and addressing nulls is also taken into consideration while scraping. After the scraping process is complete, we had around 23000 data points i.e; review summary, review content and ratings. The data then is saved into a csv file. This data is used in the next sections.

The ipynb file for the above is named as RatingData and can be found using this link: https://github.com/vangetiakhil/FRT2

Data Study and Analysis:

The data used in this study consists of text data, categorical data.

The format of data is string. There are 4 variables in this dataset and are as below:

a) Product type
b) Title
c) Review
d) Rating

The dataset contains 23044 rows.

The first step taken in the analysis part is to check for the missing values and it is found that there is one missing value in the dataset. Since all the data values are in text, the nulls are filled by a word Empty.

To proceed further with the data analysis part, the text has to be analyzed. Since, the text contains punctuations, numbers, emojis there is a need to remove these before analysis.

The WordCloud has been plotted for the title columns and reviews column after cleaning the text in it and found that the most used words are:

a) Product
b) Good
c) Bad
d) Worth
e) Value
f) Price and so on…

Since, the title column is a short summary of the review columns, these two can be merged.

Data preprocessing:

The Ratings column is our target variable and the data in it is in string format. This column is converted to integer type categorical variables.

The product type is a categorical variable and needs to be encoded to be fed into the model. This is done by using the LabelEncoder from sklearn library.

The title and reviews columns have been merged and the text cleaning has been done by removing the stopwords, punctuations, digits, white spaces, emojis and then the data has been split into list of lists by splitting the sentences into words in each rows.

Stemming has been done with the help of snowball stemmer and the words are not satisfactory. Hence, Stemming is ignored for this project. To model the available data, everything must be in numbers. Since, the data split is not in integers, vectorization needs to take place to convert them into numbers.

For the vectorization part, Word2Vec method has been selected. The number of features has been selected as 100 which indeed is the vector dimensionality. The average of the vectors for the words in the text will be used as the input for model to substitute the text.

Modelling:

The target variable is categorical and have five classes in number. Therefore, this is a multi-class classification problem.

For this problem, we have used three techniques listed below:

a) Logistic Regression with OneVsRest method
b) SVC with OneVsOne method
c) Decision Tree Classifier
d) Random Forest Classifier
e) KNeighbors Classifier

The above mentioned modeling techniques are applied and their accuracy scores ranges from 40 – 60%. To check which model performs better, cross validation has been used.

After the cross validation scores have been checked, it is observed that the Logistic regression and SVC are performing better with minimal difference in accuracy scores from cross validation.

GridsearchCV is used to find the best parameters for both the models. Though, there isn't much increase in the accuracy scores even after they are modelled with best parameters.

To select a model between these two, classification report and confusion report metrics have been used. After comparing both the models based on the terms of recall and precision, it is noticed that SVC is performing better. Hence, SVC is our finalized model.

Conclusion

The data has been collected by scraping the e-retailer Amazon.in. The data has been cleaned, analyzed, encoded and models based on different techniques have been built.

The models are predicting with an accuracy score near to 50%. This can be further improved by expanding the reviews collection from many other sources.

Note:- The full details of the project (ipynb file) can be found in this link: https://github.com/vangetiakhil/FRT2