

**NAME OF THE PROJECT**

Rating Prediction Project

Submitted by:

Dharambir Mahto

**ACKNOWLEDGMENT**

This is to acknowledge that I have done all my work with great sincerity and enthusiasm. I would like to thank madam khushboo garg who has given me the opportunity to work on Rating Prediction Project. While working on the project I experienced lot of problems and hurdles but I am glad that I completed my project on time. In this project I have learned a lot of things and I think that this project will prove to bring a breakthrough in my future career.

**Abstract—**

With the explosion of online user reviews, review rating prediction has become a research focus in natural language processing. Existing review rating prediction methods only use a single model to capture the sentiments of review texts, ignoring users who express the sentiment and products that are evaluated, both of which have great inﬂuences on review rating prediction. In order to solve the issue, we propose a review rating prediction method based on user context and product context by incorporating user information and product information into review texts. Our method ﬁrstly models the user context information of reviews, and then models the product context information of reviews. Finally, a review rating prediction method that is based on user context is proposed. A user-speciﬁc review rating prediction model, which represents the user’s personalized sentiment information, and can be learned from training data of an individual user.

**INTRODUCTION :-**

The rapid development of Web 2.0 and e-commerce has led to a proliferation in the number of online user reviews. Online reviews contain a wealth of sentiment information that is important for many decision-making processes, such as personal consumption decisions, commodity quality monitoring, and social opinion mining. Mining the sentiment and opinions that are contained in online reviews has become an important topic in natural language processing, machine learning, and Web mining. The sentiment classiﬁcation of online reviews is the most fundamental and important work in natural language processing. At present, the review sentiment classiﬁcation (positive and negative) has been widely studied, but the review sentiment classiﬁcation cannot meet the demand for ﬁne-grained review sentiment analysis. For example, how do consumers choose the most appropriate product from several products that are being reviewed? Horrigand’s research shows that consumers are willing to pay an extra 20% to 99% to buy a ﬁve-star rated product instead of a four-star rated one. This shows that the nuances of product ratings can lead to dramatic changes in product sales. In public opinion monitoring, the government not only wants to understand the positive and negative sentiment categories of the reviews, but also wants to further understand the positive and negative sentiment intensity to distinguish the urgency of public opinion events and take different measures. As a result, researchers are paying more and more attention to review rating predictions.

* **Data Gathering:-**

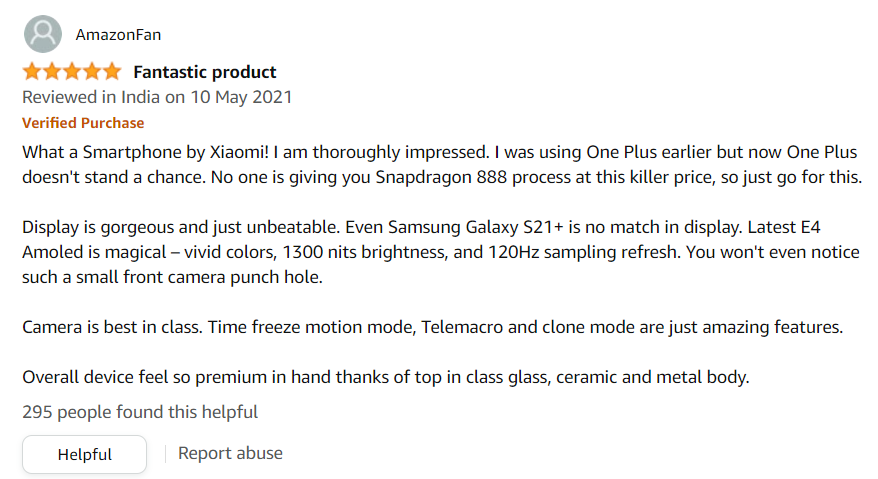
In order to get the required material to achieve this dissertation, files containing consumer reviews for diverse products derived from Amazon.com and flipkart.com were used. The extracted files only represented a subset of the data in which all items had at least 5 reviews (5-core). Duplicates, accounting for less than 1 per cent of reviews, were removed.

* **Methodology:-**

The purpose of this project is to explain the different steps that were carried out in order to develop the machine learning task and therefore being able to predict the ratings based on the text reviews. More precisely, developing a machine learning algorithm requires some well-defined steps such as data gathering, choosing the variables of interest, pre-processing the data, resampling the data, training and finally testing the algorithm.

* **Choice of the variables:-**

1. Dependent variable - As stated in the title of this dissertation, the aim is to predict ratings of Amazon reviews. Therefore, the dependent variable is the numerical rating of a specific review rated by a consumer. The popular Amazon’s rating system works along with a five-star system and allocates a certain number of stars proportionately to the satisfaction of the customers. As illustrated in Figure. Stars corresponds to the lowest rating whereas 5 stars are equivalent to the best rating.



Here is an example of a review belonging to the cell phones and accessories category that has been assigned 5 stars, meaning that the consumer was extremely satisfied with the product.



When a product has been rated by several consumers, an average overall rating can be subsequently computed and displayed in the product detail page, giving a more accurate overview of the product’s quality.

In short, the number of stars provides a summary of the product’s appreciation. This numerical rating system subsequently helps other consumers deciding which products to buy or not. Indeed, when looking at the overall rating, the consumer can directly identify positive reviews from negative reviews.

1. Independent variables:-Independent variables are defined as variables having an impact on the variation of the dependent variable. In the framework of this dissertation, we decided in a first phase only to concentrate on the text review to extrapolate the value of the rating. Therefore, the main independent variable is the text review. As the text review justifies a user’s rating, we expect those two variables to be highly positively correlated. In addition, we also considered the review summary. This is because many users prefer to read the summary (rather than the entire review). Thus, the review summary could also be a relevant predictor of star-rating. We will test to what extend the performance of the classifiers varies when taking into account the text review as well as its review summary.

* **Data Cleaning:-**

This step, also known as pre-processing, is aimed at cleaning the data. Indeed, text reviews contain unnecessary and redundant characters and have to be normalized. More precisely, the objective of the data cleaning consists in:

a) Removing punctuation

b) Removing numbers

c) Converting text to lower case (no capital letters)

d) Removing extra whitespace

e) Removing stop-words (extremely common words which do not provide any analytic information and tend to be of little value i.e. a, and, are etc.)

* Hardware and Software Requirements and Tools Used:-

1. **Hardware Required:-**

Processor – intel core i3 used to process the software faster but intel i5 or i7 is recommended.

RAM – 4 GB DDR :- used to run software run faster.

1TB Hard Disk(HDD) :- used for the storage but its always recommended to use SSD.

**B)Software Required:-**

1)Anaconda :- anaconda is a software which contains lots of libraries and software in itself like jupyter notebook, spyder, pycharm etc.

2)jupyter notebook /spyder:- Jupyter Notebook is a software to write codes efficiently and effectively and it is user friendly.

**C)Tools:-**

1)Matplot library :- Matplotlib is a cross-platform, data

visualization and graphical plotting library for Python

and its numerical extension NumPy.

2)Pandas library:- pandas library is used for data

manipulation and analysis.

3)Numpy library:- NumPy is a Python library used for working with arrays. It also has functions for working in domain of linear algebra, fourier transform, and matrices.

4)Sklearn library:- Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python.

5)Seaborn library:- Seaborn is an open-source Python library built on top of matplotlib. It is used for data visualization and exploratory data analysis. Seaborn works easily with dataframes and the Pandas library. The graphs created can also be customized easily.

* **Training a classifier:-**

In the training phase, the classifier assigns a class to each instance that has been annotated beforehand. This steps enables the classifier to learn from the instances to be later able to accurately classify some new instances. We can see the different classes that can be assigned depending on the chosen approach. In order to implement this training step, the Python programming language and its Scikit-learn libraries were used. Firstly, the training corpus (reviews and corresponding classes) developed in the previous steps has been treated as a bag-of-words (BoW) and turned into numerical features vectors using Count Vectorizer method. Indeed, the latter allows to tokenize text in order to obtain a dictionary of features and a term-document matrix. In order to attribute each feature a more relevant index than just its occurrence, the tdIdf Transformer method has been used in order to obtain its tf-idf. As explained above, this tf-idf allows to capture the relevance of terms (tf) while taking the importance (discriminative power) into account by assigning them different weights (idf). Finally, the different classifiers have been implemented in order to figure out which one of them was the most appropriate for this text classification task.

* **Evaluation:-**

This step enables to measure the performance and test the effectiveness of the trained classifiers. In other words, we can see whether the classifier learned some general principles and is able to predict an accurate outcome on new unseen instances. In order to perform the evaluation of the different classifiers, a pipeline has been built in order to make the previous vectorization-transformation-classification steps easier to work with. The classifiers are evaluated using a 10-fold cross-validation. Moreover, a classification report displays several basic evaluation performance metrics such as precision, recall and f1-score. All the classifiers are all evaluated through these accuracy metrics so that we are able to compare them afterwards. Finally, a confusion matrix is also computed to get an overview of the actual values vs. predicted values for each classifier.

* **Conclusion:-**

In this paper, we present a new RRP method that is based on the user context. More speciﬁcally, in order to solve problem of existing RRP methods based on review content, considering the user personalized information and product context information, which are useful to predict rating of review, we propose a RRP method based user context and product context. Experimental results on the datasets show that our proposed methods have more performance than the state-of-the-art baselines in RRP.