The Investigation of Multiple Product Rating Based on Data Mining Approaches

Parnian Zare[1] Mahdi Mehrabi[2]
Department of computer engineering , faculty of engineering, Shiraz branch,
Islamic Azad university, Shiraz, Iran

Abstract

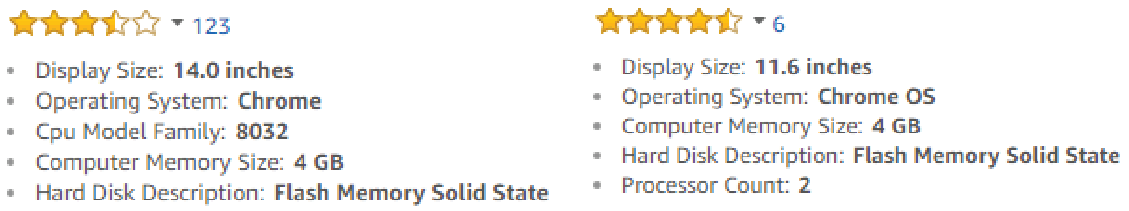
Ratings and product reviews could be considered as one of the main features determining the quality of a product in online store systems, especially in deciding whether to place a product as part of an online store's inventory. Online vendors are attracted by product reviews and ratings in order to study on potential products and related predictions. In this way, different machine learning algorithms such as Support Vector Machine, Bayesian Networks, Random Forests and Logistic Regression are investigated. The performance of each model is evaluated using accuracy, sensitivity and F1 score on the data from amazon online store website, 1996 to 2014. It is noteworthy to mention that the results of this paper can be used as an initial input to long-term product rating predictions.

Keywords: Rating, Machine Learning Algorithm, Text mining, Classification, Resampling
DOI: 10.7176/CEIS/10-5-03
Publication date:June 30th 2019

1- Introduction

Ratings and product reviews are main indicators of product quality in online store systems. Noticeably online vendors pay attention to the ratings in product warehousing retainment process. They also believe long-term product rating predictions would help them introducing certain product on the store website.

By the expansion of electronic commerce such as Amazon and eBay, online purchasing recognized as the most important trading method in the last decade. Although the main advantage of e-purchasing is physical inexistency at the store, this does not allow customers to physically evaluate the product and have to complete purchase process based on their senses. Hence, after price consideration, online customers will be directly pay attention to product rating and reviews in order to make purchasing decision. For each product, the rating information includes two values a) Average product rate b) Number of the voters. As shown in Figure 1-1, if an online product is rated by many users, the customer will ensure that the product information is reliable. On the other hand, if a product is rated only by multiple users, the customer may not feel confident in his purchasing decision.



★★★½☆ ▾ 123

- Display Size: **14.0 inches**
- Operating System: **Chrome**
- Cpu Model Family: **8032**
- Computer Memory Size: **4 GB**
- Hard Disk Description: **Flash Memory Solid State**

★★★★½ ▾ 6

- Display Size: **11.6 inches**
- Operating System: **Chrome OS**
- Computer Memory Size: **4 GB**
- Hard Disk Description: **Flash Memory Solid State**
- Processor Count: **2**

Figure 1-1, Product Rating Sample by Amazon

Increased number of voters will influence average product rating become closer to average population rating in long run. Long-term product predictions are beneficial for both retailers and customers. Various studies showed that user ratings have undeniable effects on customer purchasing decisions, as well business profits. Along with the importance of product ratings for online vendors, product ratings are also known as an important internal part of the world's leading web services such as Amazon, Tryp Odyssey, Epinions, and Yalep where users can express their opinions about a product, company or a business by writing textual reviews. Usually this rating systems contain a text field and star rating evaluation. User's rating system for a product would usually be as follows:

Restaurant: A Restaurants, Shiraz, B. Street

Food Quality ★★★½☆ Service Quality ★★★★☆ Environment ★★★★★

Textual review: I have eaten at restaurant A for many times. Food quality and variety is good. But it's a small place, so you can never go straight there and find a sit available. Another problem would be low speed

[1] MSc of Information Technology Engineering, Shiraz University, Azad Branch, Iran.
[2] Assistant Professor of Computer Software and Information Technology, Shiraz University, Azad Branch, Iran.

serving so you must wait too much.

As can be seen from mentioned rating system, user performs two tasks for multiple rating objectives. One through assigning multiple rating score to different characteristics of a product, and the other by textual reviews needs to be written in the text field. These textual reviews could affect product quality in positive or negative manner. The challenge of low product related information in multiple rating systems makes us to judge the quality of a product on long-term prediction basis.

2- Literature Review

Gano et al. (2009) tried to improve the rating system by examining user experiences. Online comments considered as an important issue for users in purchasing processes. Furthermore, most comments are written in a free text format so computer systems could not easily understand, analyze, and collect them. If the structure and feelings which are provided in the reviews taken in to account, the user experience will be greatly improved. Consequently, they focused on identifying information in free texts and using knowledge to improve the user experience.

Li Hong and colleagues (2010) presented a method to improve numerical ratings. Unigrams and n-grams were the most commonly used. Unigrams could not capture important phrases such as "could have been better", which is essential for prediction models. On the other hand, n-grams considered such expressions, but usually appear to have poor performance in the training set and thus not able to produce powerful predictions. According to the limitations of these two models, a new type of presentation was introduced: root word, set of words which could modify common sentences and negative words. They also provided a limited Ridge regression algorithm for learning outcomes related to the reviews. The experiments showed that the methodology of the Kiev opinion is much better than the earlier advanced techniques for review rating predictions.

Ming and Khademi (2014) presented the text of a review, along with the numerical score. The numerical score was predicted only by reviewing the user's text. Online surveys considered as a valuable source of information for users but due to the large number of these texts, it's almost impossible for users to access the information they seek through all the reviews. To provide a business review, one solution is to assign a rating of 1-5 to the business. This privilege can be personal and equitable to the user's mind. They also predicted a business rank based on the user-generated theory texts which not only provided an overview of the texts high-level ideas, but also abolishes individualism.

Christensen et al. (2017) introduced online communities as an attractive source of ideas which are relevant for new product development and innovation. However, making sense of the ‘big data’ in these communities is a complex analytical task. In the paper they described how to tune the model and which text mining steps to perform. The results conclude that machine learning and text mining could be useful for detecting ideas in online communities.

Miller et al. (2018) mentioned that supervised methods are likely to provide better qualitative results, model selection procedures, and model performance measures. They illustrated that much of the expense of manual corpus labeling comes from common sampling practices such as random sampling that result in sparse coverage across classes, and duplicated effort of the expert who is labeling texts. Furthermore, they outlined several active learning methods for iterative text modeling and article sampling which leads researchers to train high performance text classification models.

Usai et al. (2018) increased awareness of the potential text mining technique to discover knowledge and further promote research collaboration between knowledge management and the information technology communities. Since its emergence, text mining has involved multidisciplinary studies, various database technologies, Web-based collaborative writing, text analysis, machine learning and knowledge discovery.

3- Research Goal

Investigation of multiple product rating based on data mining approaches:
For the objective of rating various attributes of a certain product, text mining approaches and classification methods were used.

3-1- Text Binary Classification

The goal is to achieve user attached classes from corresponded rating and textual reviews provided by Amazon. For this purpose, only two classes will be considered as target outputs which are:

1) Low Rate or Class 0: Rating scores which are less than 3.
2) High Rate or Class 1: Rating scores which are equal or more than 3.

The aim of binary classification algorithm is to assign class0 or class1 in to a new unseen textual review. Figure 3-1 demonstrates binary text classification.

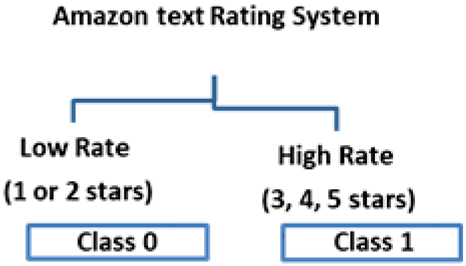


Figure 3-1, Binary Classification

Generally, high rate class would have a positive impact on customer purchasing decisions, while low rate class tend to discourage customers from purchasing products.

3-2- Text Multi-class Classification

The strategy is to extend binary classification mode in order to assign classes in to textual ratings based on the accurate star system. Unlike the binary classification which samples are classified to limited classes, the multi-class classification proposes to classify samples to more than two classes with nominal values. As it can be seen from figure 3-2, each textual rating score is mapped to a specific class using one to one relationship.

x_i ∈ {1 ^ *, 2 ^ *, 3 ^ *, 4 ^ *, 5 ^ *} Formula 3-1

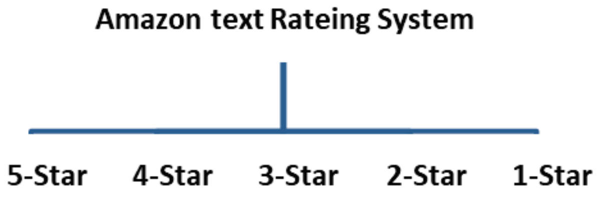


Figure 3-2, Multi-class Classification

In fact, some of the classification algorithms such as Bayesian networks and Logistic regression are naturally designed to obtain multi-class classification objectives. But other algorithms such as Support Vector machines are designed for binary classifications and require further processes to manage multi-class classification such as one versus one strategy.

3-3- Logistic Regression

At first glance, Logistic regression and multi-class classification may look similar, but they are theoretically different. In Logistic regression, class values are numbers between 1 and 5.

Logistic regression also maintains the order. For example, class 4 is better than class 1. In the case of logistic regression, the positioning of a real 5-Star rated product in class 4 increases prediction accuracy.

x_i ∈ {1,2, 3,4,5}. Formula 3-2

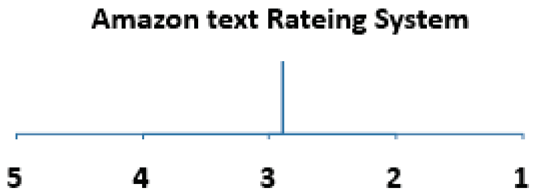


Figure 3-3, Logistic Regression and Product Rate Recognition

3-4- Text Classification Implementation

In order to implement a text classification algorithm, it is necessary to complete a few stages which are wholly mentioned in research methodology.

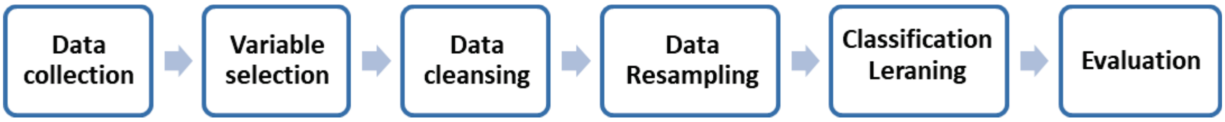
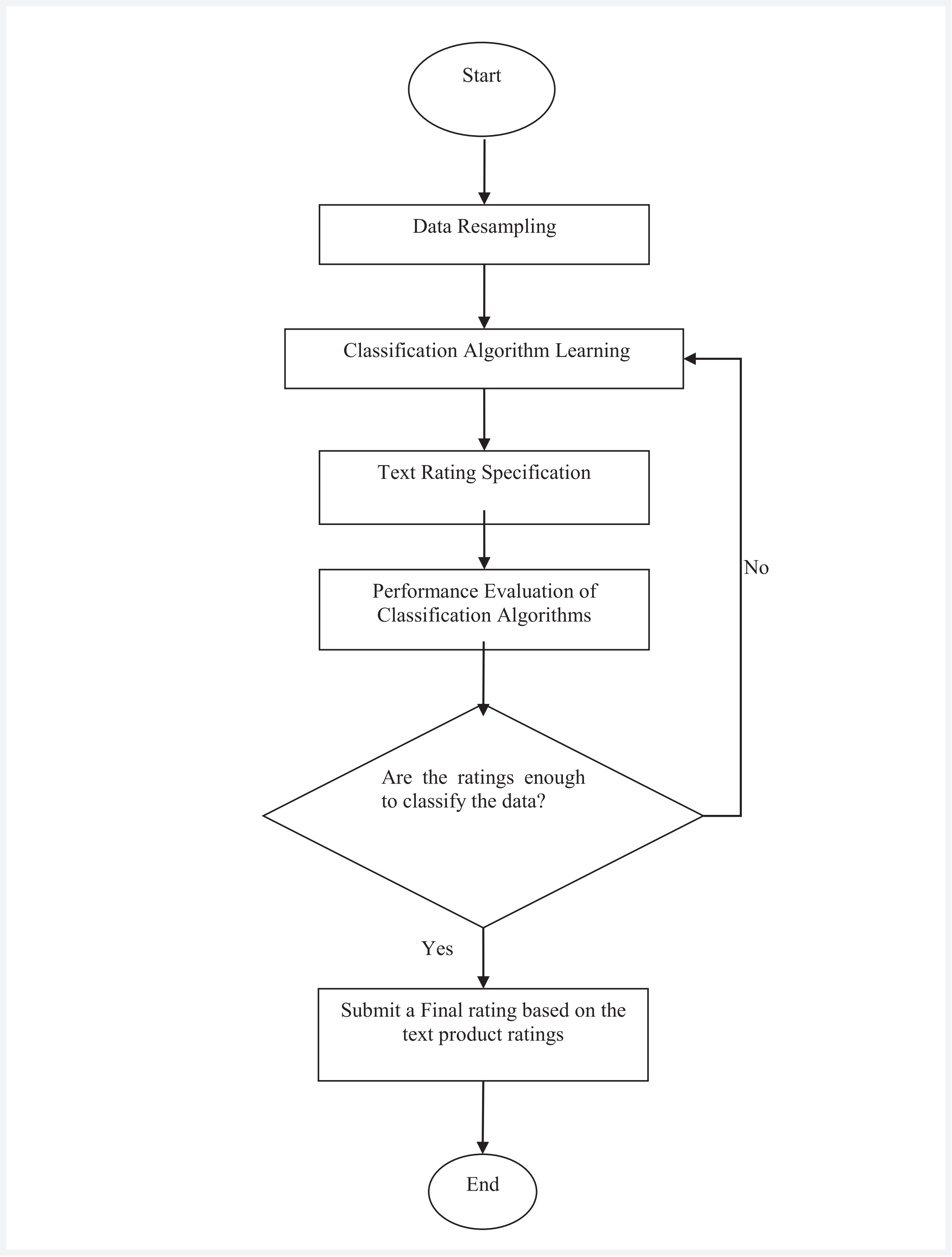


Figure 3-4 Text Classification Implementation

4- Research Flowchart


Start
Data Resampling
Classification Algorithm Learning
Text Rating Specification
Performance Evaluation of Classification Algorithms
Are the ratings enough to classify the data?
No
Yes
Submit a Final rating based on the text product ratings
End

5- Research Methodology

5-1- Data Collection

For the proposed objectives of this paper different product datasets containing customer ratings were used. These extracted files which are derived from Amazon only represent a subset of data, that all products have 5 ratings or each user has rated at least 5 times. Duplicate ratings which are less than 1% of the total were eliminated. Each rating includes following labels:

1) Rating ID
2) Product ID
3) Voter Name
4) Textual Reviews
5) Rating Efficiency Score to other users
6) Rating Score (Star system)
7) Rating Result and Summary (Text format)
8) Rating time (Unix time)
9) Rating time (Day/Month/year format)

Rating labels example: {ACU3LCRX4A8RV, 3998899561, Zonaldo Reefey, [2,3], works well. Unlike the previous models, it does not have temperature problems. Without any doubt this is the best charger you can buy with such a reasonable price. That's why I bought one for myself and one for my wife, 5, really great quality, 1377388800, 08/25/2013}

Among above labels, we've only work with text reviews plus rating results and summaries. The rest of the existing elements considered unrelated in a context-sensitive framework. For extracting this information JSON data format, R programming language and Excel have been used. Furthermore, in this paper experienced and searched products are considered as two main categories of data collection. Quality evaluation of experienced products are difficult as huge number of users should buy the product in order to assess the quality. The specific sample utilized from this category is computer games. On the other hand, quality evaluation of searched products can take place through considering their key features on the internet. In this way, purchasing the product is not a necessary task. The specific sample utilized from this category is Mobile and Accessories.

5-2- Variable Selection

Dependent variables: Amazon rating system which is based on a 5-Star evaluation, used as a dependent variable. Dependent variables could take different values due to specific classification mode. For example:

1) Dependent variables for text binary classification algorithm would be low rate and high rate.
2) Dependent variables for text multi-class classification would be 1-Star, 2-Star, 3-Star, 4-Star and 5-Star.
3) Dependent variables for Logistic regression would be 1, 2, 3, 4, 5.

Independent variables: textual reviews, rating results plus summaries and the combination of them are used as independent variables. The goal is to find out which of these three variables provide better performance for the complex structure of the text.

5-3- Data Cleansing

Naturally, textual reviews contain duplicate and non-essential words therefore data preprocessing is used for the cleaning objectives. The purpose of data preprocessing mainly highlighted below:

1) Punctuation deletion
2) Number deletion
3) Extra space deletion
4) Stop word deletion
5) Lowercase conversion

Data preprocessing simplify the data and gain more accuracy in classification task. Data cleaning process is done using R programming language and its related packages. This stage provides cleaned training dataset which includes rating score and cleaned textual reviews.

5-4- Data Resampling

Data resampling checks data distribution before classification process. The aim is to figure out whether the data needs to be re-sampled or not.

Computer games dataset includes ratings in 14% low rate and 86% high rate. It also presents 7% ratings in 1-Star, 7% ratings in 2-Star, 14% ratings in 3-Star, 26% ratings in 4-Star and 46% ratings in 5-Star.

Mobile and accessories dataset include ratings in 13% low rate and 87% high rate. It also presents 7% ratings in 1-Star, 6% ratings in 2-Star, 11% ratings in 3-Star, 12% ratings in 4-Star and 55% ratings in 5-Star.

It is obvious that both datasets have unbalanced distribution therefore, the data must be resampled in order to prevent bias and false results. In this paper sample overfitting and underfitting methods are used to eliminate data distribution problems.

5-5- Classification Algorithm Learning

In the learning stage, classification algorithms assign a class to each labeled sample. This stage enables classification algorithms to learn from the samples and then correctly classify the new ones. Different classification algorithms are developed in order to determine the best according to performance measurements.
To implement this learning stage, Python programming language and its related libraries have been used. All implemented classification algorithms have been learned with both datasets utilizing sample overfitting and underfitting method.

5-6- Evaluation

Evaluation enables us to measure performance and effectiveness of trained classification algorithms. In other words, we want to see if the classification algorithm has been able to correctly classify new and unseen instances. Performance evaluation of a classification algorithm includes tasks below:

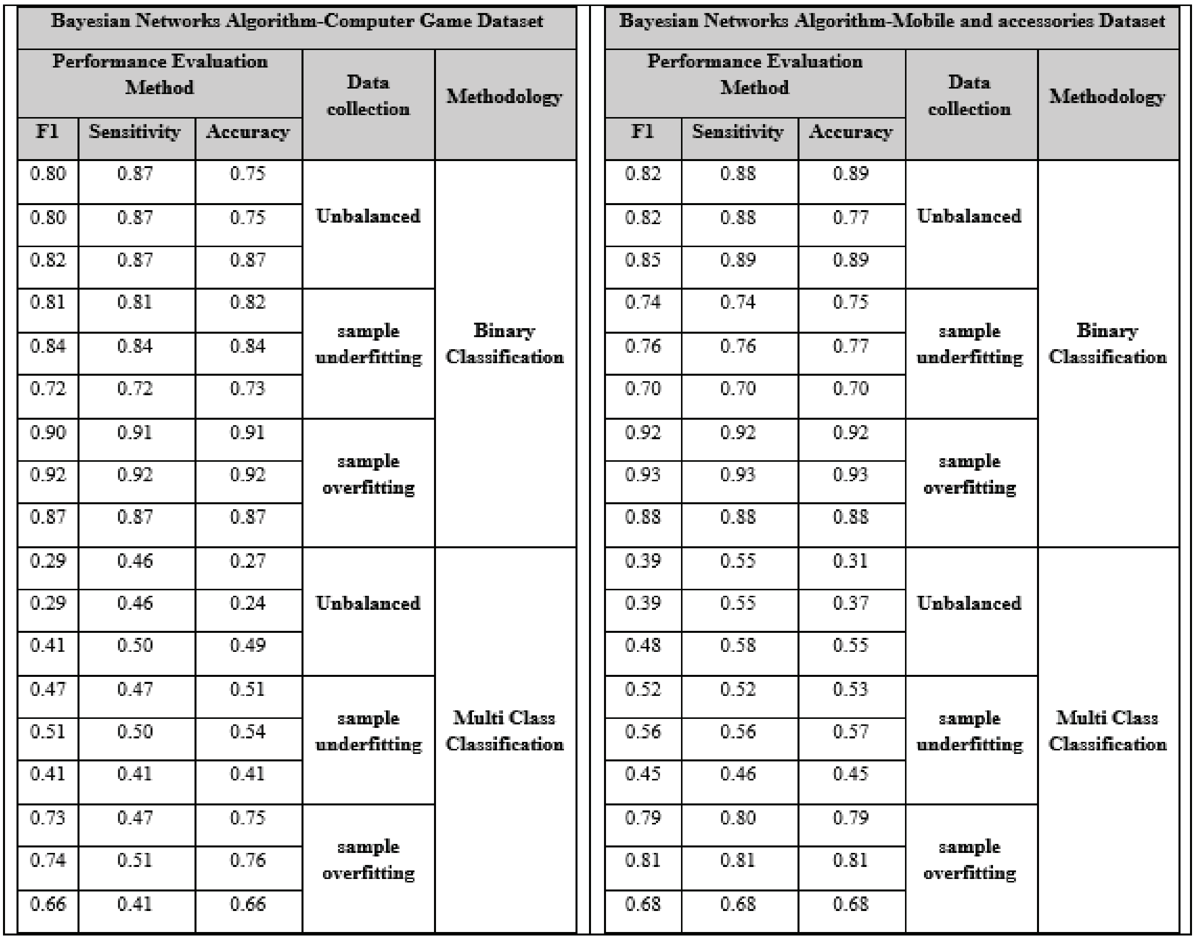
1) parallel environment creation to simplify the conversion, transformation, and classification of earlier stages.
2) 10-cross-validation implementation.
3) accuracy, sensitivity and F1 score measurements.

6- Numerical Results Obtained from Classification Algorithms

Performance related measurements due to different classification algorithms are presented in the

1) Table 6-1, Bayesian Networks Algorithm
2) Table 6-2, Support Vector Machines Algorithm
3) Table 6-3, Random Forests Algorithm
4) Table 6-4, Logistic Regression Algorithm

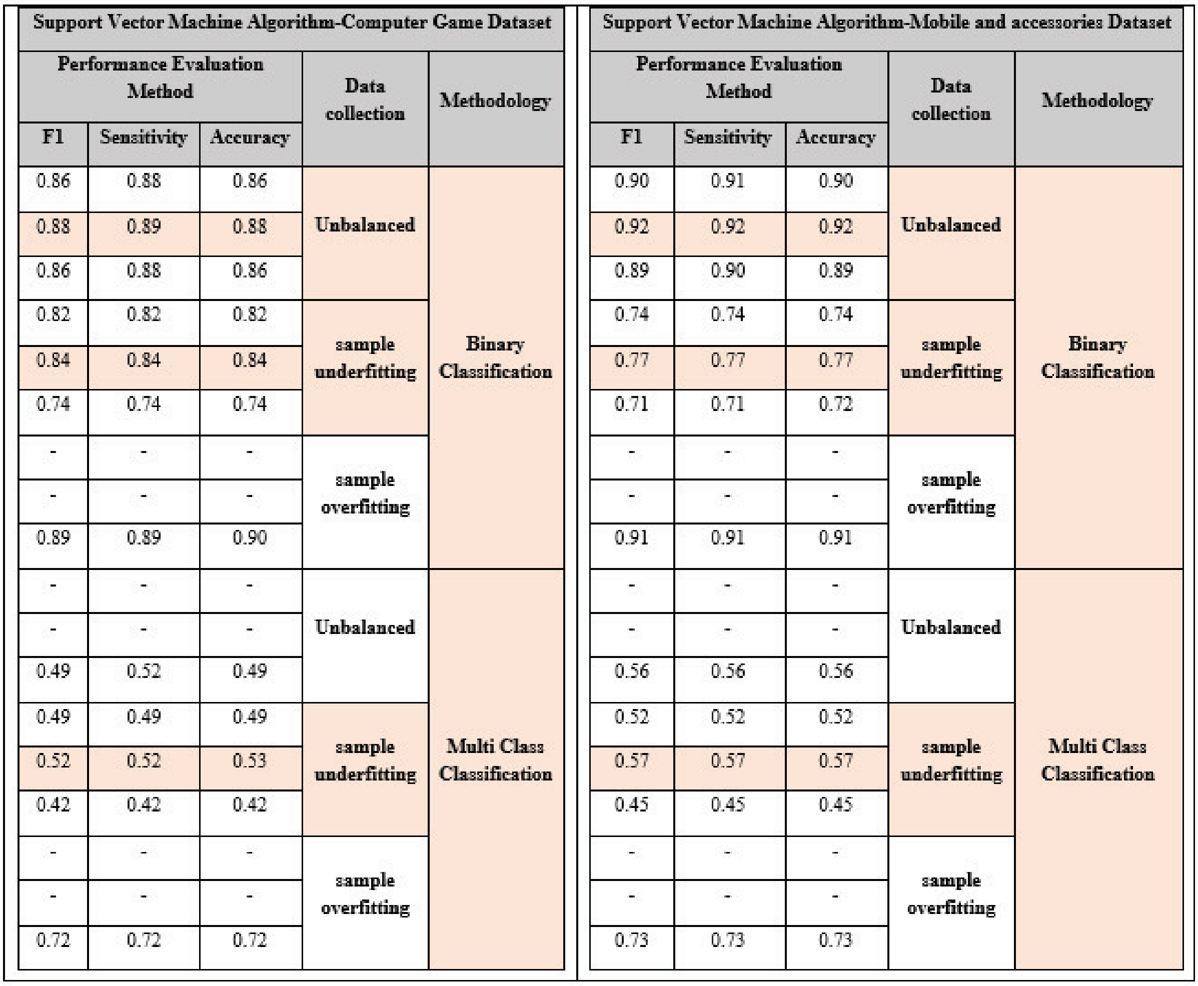
The measurements have been done on both computer games and mobile datasets utilizing sample overfitting and sample underfitting method.



Bayesian Networks Algorithm-Computer Game Dataset				
Performance Evaluation Method			Data collection	Methodology
F1	Sensitivity	Accuracy		
0.80	0.87	0.75	Unbalanced	Binary Classification
0.80	0.87	0.75		
0.82	0.87	0.87		
0.81	0.81	0.82	sample underfitting	
0.84	0.84	0.84		
0.72	0.72	0.73		
0.90	0.91	0.91	sample overfitting	
0.92	0.92	0.92		
0.87	0.87	0.87		
0.29	0.46	0.27	Unbalanced	Multi Class Classification
0.29	0.46	0.24		
0.41	0.50	0.49		
0.47	0.47	0.51	sample underfitting	
0.51	0.50	0.54		
0.41	0.41	0.41		
0.73	0.47	0.75	sample overfitting	
0.74	0.51	0.76		
0.66	0.41	0.66		

Bayesian Networks Algorithm-Mobile and accessories Dataset				
Performance Evaluation Method			Data collection	Methodology
F1	Sensitivity	Accuracy		
0.82	0.88	0.89	Unbalanced	Binary Classification
0.82	0.88	0.77		
0.85	0.89	0.89		
0.74	0.74	0.75	sample underfitting	
0.76	0.76	0.77		
0.70	0.70	0.70		
0.92	0.92	0.92	sample overfitting	
0.93	0.93	0.93		
0.88	0.88	0.88		
0.39	0.55	0.31	Unbalanced	Multi Class Classification
0.39	0.55	0.37		
0.48	0.58	0.55		
0.52	0.52	0.53	sample underfitting	
0.56	0.56	0.57		
0.45	0.46	0.45		
0.79	0.80	0.79	sample overfitting	
0.81	0.81	0.81		
0.68	0.68	0.68		

Table 6-1, Bayesian Networks Algorithm Results



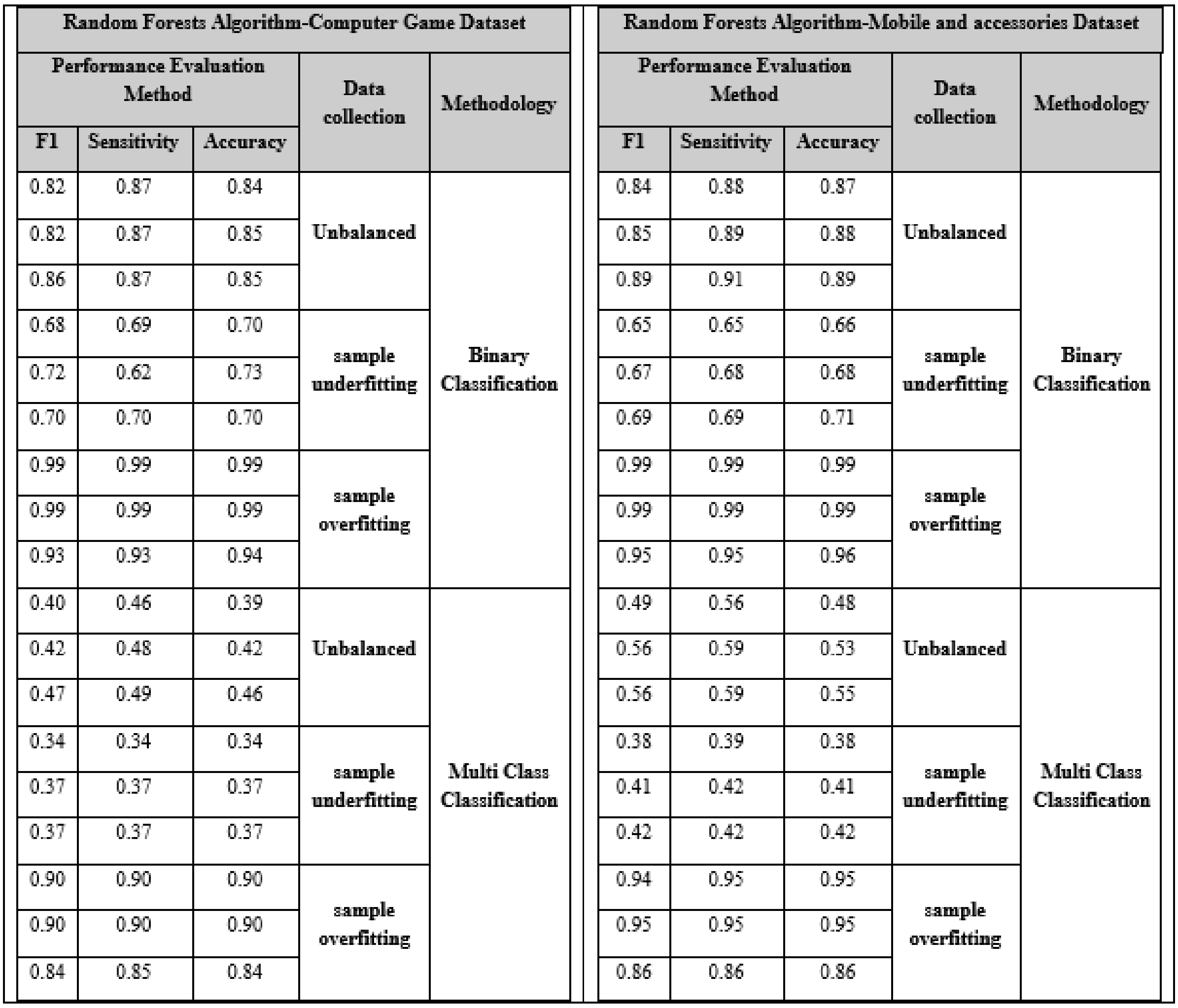
Support Vector Machine Algorithm-Computer Game Dataset				
Performance Evaluation Method			Data collection	Methodology
F1	Sensitivity	Accuracy		
0.86	0.88	0.86	Unbalanced	Binary Classification
0.88	0.89	0.88		
0.86	0.88	0.86		
0.82	0.82	0.82	sample underfitting	
0.84	0.84	0.84		
0.74	0.74	0.74		
-	-	-	sample overfitting	
-	-	-		
0.89	0.89	0.90		
-	-	-	Unbalanced	Multi Class Classification
-	-	-		
0.49	0.52	0.49		
0.49	0.49	0.49	sample underfitting	
0.52	0.52	0.53		
0.42	0.42	0.42		
-	-	-	sample overfitting	
-	-	-		
0.72	0.72	0.72		

Support Vector Machine Algorithm-Mobile and accessories Dataset				
Performance Evaluation Method			Data collection	Methodology
F1	Sensitivity	Accuracy		
0.90	0.91	0.90	Unbalanced	Binary Classification
0.92	0.92	0.92		
0.89	0.90	0.89		
0.74	0.74	0.74	sample underfitting	
0.77	0.77	0.77		
0.71	0.71	0.72		
-	-	-	sample overfitting	
-	-	-		
0.91	0.91	0.91		
-	-	-	Unbalanced	Multi Class Classification
-	-	-		
0.56	0.56	0.56		
0.52	0.52	0.52	sample underfitting	
0.57	0.57	0.57		
0.45	0.45	0.45		
-	-	-	sample overfitting	
-	-	-		
0.73	0.73	0.73		

Table 6-2 Support Vector Machine Algorithm Results

Explanations:

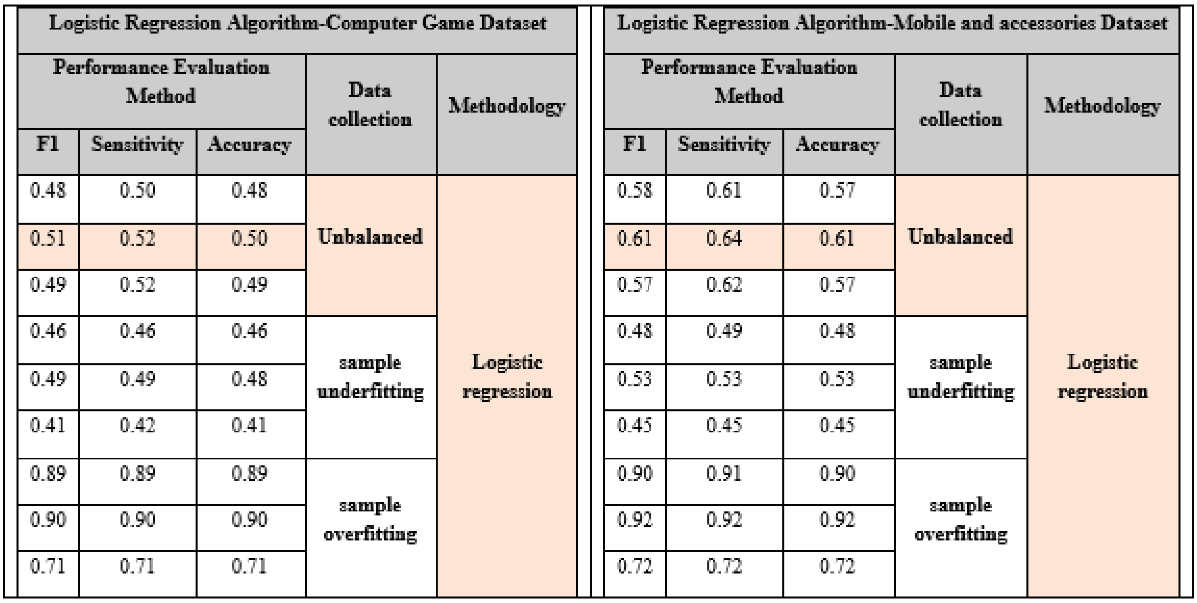
1) The result of support vector machine algorithm using sample underfitting method could not be displayed due to low speed execution. Therefore, these results are marked with "-".
2) Colored sections represent the best performance among all mentioned classification algorithms.



Random Forests Algorithm-Computer Game Dataset				
Performance Evaluation Method			Data collection	Methodology
F1	Sensitivity	Accuracy		
0.82	0.87	0.84	Unbalanced	Binary Classification
0.82	0.87	0.85		
0.86	0.87	0.85		
0.68	0.69	0.70	sample underfitting	
0.72	0.62	0.73		
0.70	0.70	0.70		
0.99	0.99	0.99	sample overfitting	
0.99	0.99	0.99		
0.93	0.93	0.94		
0.40	0.46	0.39	Unbalanced	Multi Class Classification
0.42	0.48	0.42		
0.47	0.49	0.46		
0.34	0.34	0.34	sample underfitting	
0.37	0.37	0.37		
0.37	0.37	0.37		
0.90	0.90	0.90	sample overfitting	
0.90	0.90	0.90		
0.84	0.85	0.84		

Random Forests Algorithm-Mobile and accessories Dataset				
Performance Evaluation Method			Data collection	Methodology
F1	Sensitivity	Accuracy		
0.84	0.88	0.87	Unbalanced	Binary Classification
0.85	0.89	0.88		
0.89	0.91	0.89		
0.65	0.65	0.66	sample underfitting	
0.67	0.68	0.68		
0.69	0.69	0.71		
0.99	0.99	0.99	sample overfitting	
0.99	0.99	0.99		
0.95	0.95	0.96		
0.49	0.56	0.48	Unbalanced	Multi Class Classification
0.56	0.59	0.53		
0.56	0.59	0.55		
0.38	0.39	0.38	sample underfitting	
0.41	0.42	0.41		
0.42	0.42	0.42		
0.94	0.95	0.95	sample overfitting	
0.95	0.95	0.95		
0.86	0.86	0.86		

Table 6-3 Random Forests Algorithm Results



Logistic Regression Algorithm-Computer Game Dataset				
Performance Evaluation Method			Data collection	Methodology
F1	Sensitivity	Accuracy		
0.48	0.50	0.48	Unbalanced	Logistic regression
0.51	0.52	0.50		
0.49	0.52	0.49		
0.46	0.46	0.46	sample underfitting	
0.49	0.49	0.48		
0.41	0.42	0.41		
0.89	0.89	0.89	sample overfitting	
0.90	0.90	0.90		
0.71	0.71	0.71		

Logistic Regression Algorithm-Mobile and accessories Dataset				
Performance Evaluation Method			Data collection	Methodology
F1	Sensitivity	Accuracy		
0.58	0.61	0.57	Unbalanced	Logistic regression
0.61	0.64	0.61		
0.57	0.62	0.57		
0.48	0.49	0.48	sample underfitting	
0.53	0.53	0.53		
0.45	0.45	0.45		
0.90	0.91	0.90	sample overfitting	
0.92	0.92	0.92		
0.72	0.72	0.72		

Table 6-4 Logistic Regression Algorithm Results

Explanation:
Colored sections represent the best performance in predicting precise score using logistic regression.

7- Numerical Analysis of Classification Algorithms Performance

7-1- Analyze the Results of Unbalanced Data

As it can be seen from numerical results of computer games dataset:

1) Binary classification methodology with an unbalanced dataset provides fairly good results in different classification algorithms. F1 performance evaluation score take different values between 0.8 and 0.88. Specifically support vector machine algorithm provides the best performance.
2) Logistic regression algorithm with an unbalanced dataset would predict a precise rating score. F1 value is equal to 0.51 for its best performance.

As it can be seen from numerical results of mobile and accessories dataset:

1) Binary classification methodology with an unbalanced dataset provides fairly good results in different classification algorithms. F1 performance evaluation score take different values between 0.82 and 0.92. Specifically support vector machine algorithm provides the best performance. The results are mostly similar to computer games dataset.
2) Logistic regression algorithm with an unbalanced dataset would predict a precise rating score. F1 value is equal to 0.61 for its best performance.

7-2- Analyze the Results of Resampled Data

Utilizing sample overfitting method provides better results than sample underfitting in both binary and multi class classification. Multi-class classification gives better results in predicting precise rating score using data resampling. Although the results of sample overfitting are optimistic, as F1 performance evaluation score is identical to 0.90, it has a notable weakness. Sample overfitting method engages bigger datasets and consequently execution speed would increase. According to mentioned issue, in this paper we used the results of sample underfitting for further approaches.

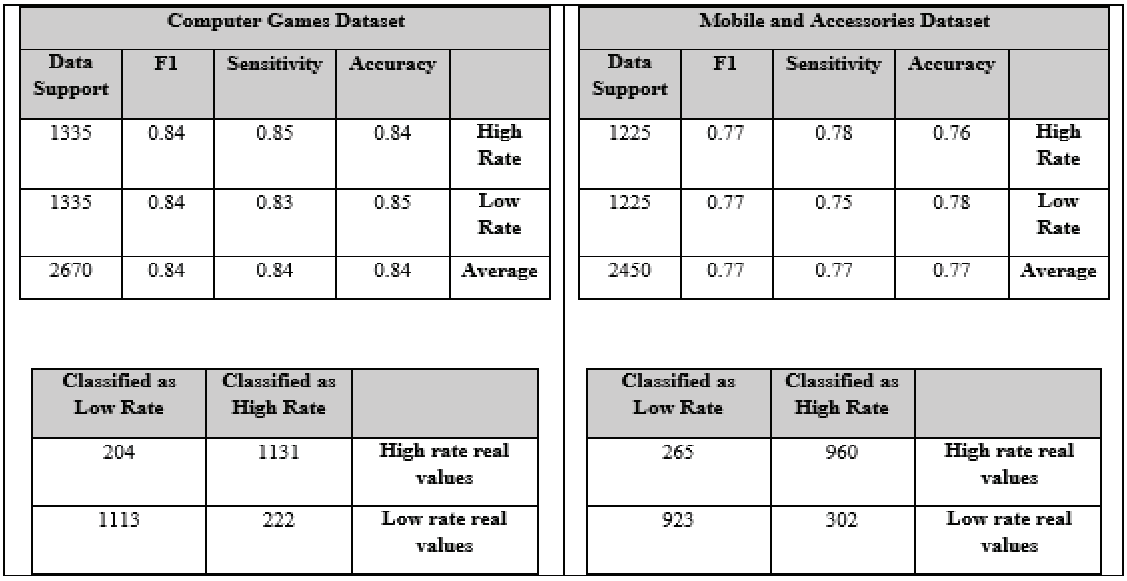
As it can be seen from numerical results of computer games and mobile and accessories datasets support vector machine and Bayesian network outcomes are close to each other in terms of accuracy. Although the support vector machine algorithm works a bit better than Bayesian networks, it has lower execution speed problem. We have to mention that classification algorithms provide better outcomes using rating results and summaries.

8- Examine the Performance of Support Vector Machine algorithm

In the following, we decided to investigate efficiency of support vector machine algorithm in terms of binary and multi class classifications. This means that the support vector machine algorithm used sample underfitting method plus rating results and summaries. The support vector machine algorithm obtained F1 performance evaluation score of 0.84 for binary classification and 0.52 for multi class classification.

8-1- Support Vector Machine Algorithm - Binary Classification

As it can be seen from table 8-1 the results of low rate classification have been greatly improved. It is also obvious that low rate classification is done more accurate than high rate classification.



Computer Games Dataset				
Data Support	F1	Sensitivity	Accuracy	
1335	0.84	0.85	0.84	High Rate
1335	0.84	0.83	0.85	Low Rate
2670	0.84	0.84	0.84	Average

Mobile and Accessories Dataset				
Data Support	F1	Sensitivity	Accuracy	
1225	0.77	0.78	0.76	High Rate
1225	0.77	0.75	0.78	Low Rate
2450	0.77	0.77	0.77	Average

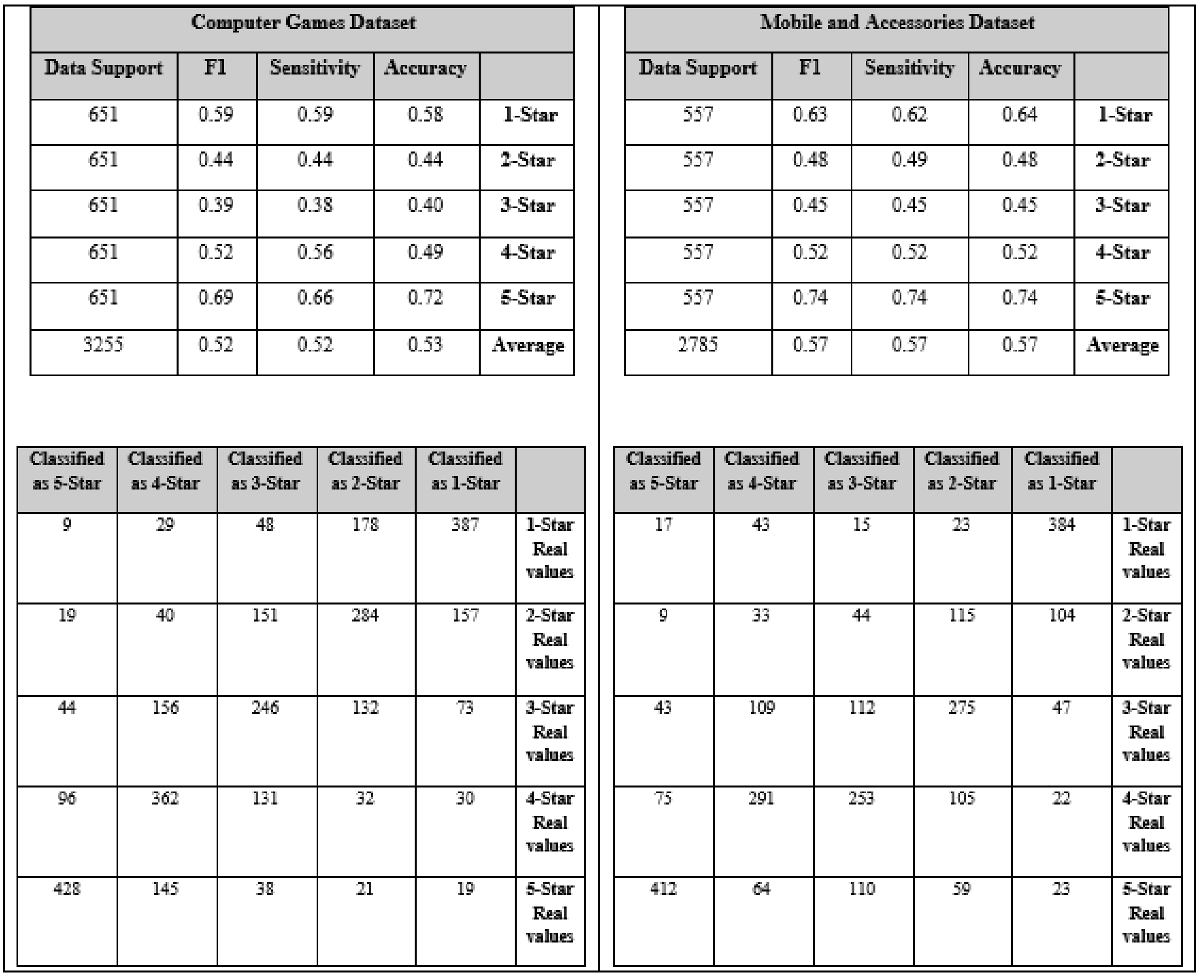
Classified as Low Rate	Classified as High Rate	
204	1131	High rate real values
1113	222	Low rate real values

Classified as Low Rate	Classified as High Rate	
265	960	High rate real values
923	302	Low rate real values

Table 8-1, Support Vector Machin Function in binary classification

8-2- Support Vector Machine Algorithm – Multi class Classification

As it can be seen from table 8-2 classifying test data to 1-Star and 5-Star classes is count as a simple task for the algorithm. It is clear that the classification algorithm has weakness in correctly classifying ratings to 3-Star class. Furthermore, support vector machine algorithm may mistakably consider 5-Star ratings as the worst-case scenario with 4-Star rating.



Computer Games Dataset				
Data Support	F1	Sensitivity	Accuracy	
651	0.59	0.59	0.58	1-Star
651	0.44	0.44	0.44	2-Star
651	0.39	0.38	0.40	3-Star
651	0.52	0.56	0.49	4-Star
651	0.69	0.66	0.72	5-Star
3255	0.52	0.52	0.53	Average

Mobile and Accessories Dataset				
Data Support	F1	Sensitivity	Accuracy	
557	0.63	0.62	0.64	1-Star
557	0.48	0.49	0.48	2-Star
557	0.45	0.45	0.45	3-Star
557	0.52	0.52	0.52	4-Star
557	0.74	0.74	0.74	5-Star
2785	0.57	0.57	0.57	Average

Classified as 5-Star	Classified as 4-Star	Classified as 3-Star	Classified as 2-Star	Classified as 1-Star	
9	29	48	178	387	1-Star Real values
19	40	151	284	157	2-Star Real values
44	156	246	132	73	3-Star Real values
96	362	131	32	30	4-Star Real values
428	145	38	21	19	5-Star Real values

Classified as 5-Star	Classified as 4-Star	Classified as 3-Star	Classified as 2-Star	Classified as 1-Star	
17	43	15	23	384	1-Star Real values
9	33	44	115	104	2-Star Real values
43	109	112	275	47	3-Star Real values
75	291	253	105	22	4-Star Real values
412	64	110	59	23	5-Star Real values

Table 8-1, Support Vector Machin Function in Multi class classification

9- Conclusion and Future Work

In this paper, we consider different models to obtain textual score. Text classification algorithms automatically assign a text document into a fixed set of classes. The goal of binary classification methodology is to classify data in to high rate or low rate classes on the other hand multiclass classification and logistic regression objectives are to find precise category or rating score. These methodologies were tested on two different datasets which are experienced and searched products. It is noteworthy to mention that implementing a text classification algorithm in unbalanced dataset is not an easy task. Indeed, resampling techniques are needed in order to balance the datasets. According to the results, successful classification algorithms are Support vector machine and Bayesian networks due to performance evaluation.

For future work analyzing further datasets form Amazon, utilizing resampling methods other than sample overfitting or underfitting, considering K famous words in the context, improving bag of words method, one-gram – two-gram – three-gram application and sentimental analysis are highly recommended.

10- References

Bannach-Brown, A., Przybyła, P., Thomas, J., Rice, A. S., Ananiadou, S., Liao, J., & Macleod, M. R. (2018). The use of text-mining and machine learning algorithms in systematic reviews: reducing workload in preclinical biomedical sciences and reducing human screening error. bioRxiv, 255760.

Ming Fan and Maryam Khademi. Predicting a business star in yelp from its reviews text alone arXiv preprint arXiv:1401.0864, 2014.

McAuley, J., Pandey, R., & Leskovec, J. (2015, August). Inferring networks of substitutable and complementary products. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). ACM.

Usai, A., Pironti, M., Mital, M., & Aouina Mejri, C. (2018). Knowledge discovery out of text data: a systematic review via text mining. Journal of Knowledge Management.

McAuley, J., Targett, C., Shi, Q., & Van Den Hengel, A. (2015, August). Image-based recommendations on styles and substitutes. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 43-52). ACM.

Dover, Y., Goldenberg, J., & Shapira, D. (2012). Network traces on penetration: Uncovering degree distribution from adoption data. Marketing Science, 31(4), 689-712.

Singh, G., Thomas, J., & Shawe-Taylor, J. (2018). Improving Active Learning in Systematic Reviews. arXiv preprint arXiv:1801.09496.

Tong, S., & Koller, D. (2001). Support vector machine active learning with applications to text classification. Journal of machine learning research, 2(Nov), 45-66.

Christensen, K., Nørskov, S., Frederiksen, L., & Scholderer, J. (2017). In search of new product ideas: Identifying ideas in online communities by machine learning and text mining. Creativity and Innovation Management, 26(1), 17-30.

Xie, Y., & Jiang, H. (2017). Stock Market Forecasting Based on Text Mining Technology: A Support Vector Machine Method. JCP, 12(6), 500-510.

Wright, M. N., & Ziegler, A. (2015). Ranger: a fast implementation of random forests for high dimensiona1 data in C++ and R. arXiv preprint arXiv:1508.04409.

Thearling, K. (2017). An introduction to data mining.

Mohammad, A. H., Alwada'n, T., & Al-Momani, O. (2018). Arabic text categorization using support vector machine, Naïve Bayes and neural network. GSTF Journal on Computing (JoC), 5(1).

Tilve, A. K. S., & Jain, S. N. (2017). Text Classification using Naïve Bayes, VSM and Pos Tagger. International Journal of Ethics in Engineering & Management Education (ISSN: 2348-4748, Volume 4, Issue 1.

Sperandei, S. (2014). Understanding logistic regression analysis. Biochemia medica: Biochemia medica, 24(1), 12-18.

Christensen, K., Liland, K. H., Kvaal, K., Risvik, E., Biancolillo, A., Scholderer, J., & Næs, T. (2017). Mining online community data: The nature of ideas in online communities. Food Quality and Preference, 62, 246-256.

Kranjc, J., Orač, R., Podpečan, V., Lavrač, N., & Robnik-Šikonja, M. (2017). ClowdFlows: Online workflows for distributed big data mining. Future Generation Computer Systems, 68, 38-58.

Miller, B., Linder, F., & Mebane Jr, W. R. (2018). Active Learning Approaches for Labeling Text.

Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., & Den Hartog, D. N. (2017). Text classification for organizational researchers: A tutorial. Organizational research methods, 1094428117719322.

Lever, J., Krzywinski, M., & Altman, N. (2016). Points of significance: classification evaluation.

Diab, D. M., & El Hindi, K. M. (2017). Using differential evolution for fine tuning naïve Bayesian classifiers and its application for text classification. Applied Soft Computing, 54, 183-199.

Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., & Lichtendahl Jr, K. C. (2017). Data mining for business analytics: concepts, techniques, and applications in R. John Wiley & Sons.

López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information Sciences, 250, 113-141.