Volume 65, Issue 5, 2021
Journal of Scientific Research
of
The Banaras Hindu University

Tapping the Sentiment of the Customers through Association Rule Mining for the Products Endorsed by Celebrities

Preeti Nair[1], Manish Kumar Pandey[2] *, Arun Kumar Singh[3]

[*1,3]Department of Commerce, Udai Pratap College, Varanasi.
[2]Banaras Hindu University, Varanasi, India, pandey.manish@live.com

Abstract: **The impact of customers sentiment in the current era is unmatched. Not only celebrities but the customers too are influencers in the revenue growth of a brand. Efficient Marketing analytics in terms of sentiment analysis is required for which association rule mining could be utilized as a prominent model. The current work explored the Predictive Apriori Algorithm to understand the sentiment of the Varanasi customers. The algorithm has performed with an accuracy of 99.34 % which proved that they stop purchasing the products once the celebrity has had negative publicity or is involved in anti-national activities. The work has also explored the correlation between the questions being asked in the survey and found that they strongly agree that the brand is effective if it is endorsed by several celebrities over a while.**

Index Terms: **Predictive Apriori; Sentiment Analysis; Celebrity Brand Endorsement; Association Rule Mining; Marketing Analytics; SMAC**

I. INTRODUCTION

The impact of customers on the revenue of the key industry players are huge and thus the sentiment of the customers is found to be the foundation stone to the success of any product in terms of revenue. Public sentiment is also one of the core factors to decide a brand's popularity. Thus, not only celebrities but also the public has become a great influencer on the success of brands and consequently became a part of the marketing strategy. Now the customers are not blindly following the celebrity's footsteps but also raising their voices against them. This calls for efficient mining of the customer's sentiment. Database mining in the retail industry in the era of Social Mobility Analytics and Cloud (SMAC) (Kumar et al., 2016a, 2016b; MK Pandey, 2017; Pandey et al., 2013; Pandey & Subbiah, 2016, 2018, 2017) & Big Data is found to be a crucial factor in the improvement of marketing strategy (Agrawal et al., 1993; Zaki, 1999). Association rule mining works by finding closed association among the set of data items through correlation study. The identification of these associations helps the companies in making improvised decision making. One such example of association rule mining used in the Fast-Moving Consumer Goods (FMCG) sector is market basket analysis where consumer purchase habit is analyzed and an association is being established between the items purchased. The current work for the first time has utilized the benefits of Association Rule Mining to tap the customer's sentiment for the products endorsed by celebrities. The problem can be represented mathematically as,

Given the database D consisting of questions asked from the customers from Q_1 through Q_N. The problem is to identify n rules $R_1 \ldots R_n \in \{[x \Rightarrow y] | x, y \subseteq \{Q_1, \ldots, Q_N\}; y \neq \emptyset; x \cap y = \emptyset\}$ that results in the maximum predictive accuracy acc ($[x \Rightarrow y]$) (Tobias Scheffer, 2001). The predictive accuracy acc is represented as the probability of the predicting correct rule for new data.

The current work is distributed into six sections. The first section introduces the importance of sentiment analysis and association rule mining. The second section briefly describes the dataset and the methodology used. The third section discusses the metrics incorporated for the performance evaluation. The fourth section describes the result and discussion followed by the conclusion in the fifth section. In the end, references are given.

II. DATASET AND METHODOLOGY

A survey is conducted in Varanasi in 5 different locations namely, Lanka, Chauk, Shivpur, Pandeypur and Sigra. In the survey, the following questions were asked to see the sentiment of the customers during purchase made for the products endorsed by celebrities. The options given to them was whether they strongly disagree, disagree, neutral, agree or strongly agree to these questions. The responses again were mapped to numeric ratings as 1,2,3,4 and 5 for strongly disagree, disagree, neutral, agree or strongly agree respectively. A total of 250 samples were

taken.

Table 1 List of Questions asked during the survey.

Q1	Purchasing products that match the celebrity
Q2	Effectiveness increases if the product is endorsed by different celebrities at different time
Q3	Not Purchasing the products overshadowed by celebrities
Q4	Not Purchasing the products endorsed by celebrities from another region
Q5	Stop purchasing the product of the negative publicity about a celebrity erupts or the celebrity is involved in anti-national activities
Q6	Prefer all products endorsed by my favourite celebrity

The descriptive statistics of the data compiled after the survey is listed in Table 2. The table lists various statistical parameters of the responses received against the questions surveyed.

Table 2 Descriptive Statistics of the data set

	Q1	*Q2*	*Q3*	*Q4*	*Q5*	*Q6*
Mean	3.636	4.072	4.036	2.336	3.984	3.956
Standard Error	0.067767142	0.050809	0.045817	0.049051	0.056389	0.056186
Standard Deviation	1.071492597	0.803366	0.724433	0.775571	0.891585	0.888381
Sample Variance	1.148096386	0.645398	0.524803	0.60151	0.794924	0.789221
Kurtosis	-0.918953813	1.340965	-0.5074	-0.44776	0.304811	0.391142
Skewness	-0.318453089	-0.92765	-0.24625	-0.29903	-0.75685	-0.77974
Sum	909	1018	1009	584	996	989
Confidence Level (95.0%)	0.133469885	0.100071	0.090239	0.096609	0.11061	0.110661

Association Rule Mining is further applied to the data for finding the association among the customer's sentiment while purchasing a product endorsed by celebrities. A candidate set is created with large itemsets and based on various heuristics; a subset is selected with large itemsets. The process is iteratively repeated using the large itemsets as an input to generate the candidate set. During the evaluation of the association rules, every rule which may differ on confidence or support was compared (Agrawal, 2013). The current work incorporated the Predictive Apriori algorithm to study the tradeoff between confidence and support (Tobias Scheffer, 2001). This is utilized to improve the predictions of the association rules. The pseudo code of the methodology (T Scheffer, 2005) is given in Table 3.

Table 3 Pseudo code of the methodology

1	Input: n represents the required number of association rules in a database D with questions q_1 through q_N	
2	Let α =1, for i=1…k Do: Create a number of association rule [$x \Rightarrow y$] with i items randomly and measure respective confidences along with their distribution (say $\pi_i(c)$).	
3	$\forall\, c, \pi\,(c) = \frac{\sum_{i=1}^{k} \pi_i(c)\binom{k}{i}(2^i-1)}{\sum_{i=1}^{k}\binom{k}{i}(2^i-1)}$ (1)	
4	Assume X_0= {∅}; then X_1={{q_1} …, {q_k}} is all questions set with single element.	
5	For i=1…(k-1) While (i=1 or ($X_{i-1} \neq \emptyset$)	
	(a)	If i>1 Then Evaluate the candidate sets of i length using $X_i = \{x \cup x` \mid x, x` \in X_{i-1}, \mid x \cup x` \mid = i\}$. The Evaluation would be optimized by choosing $x$ and $x` \in X_{i-1}$ and eliminating multiple occurrences of items in X_i.
	(b)	π would be used to eliminate items with lesser value from X_i.
	(c)	$\forall\, x \in X_i$, search efficiently the best rules with x
	(d)	If the best is changed, value of π would be increased to the smallest number such that E (c|1, π)> E (c (best [n]) | $\hat{c}$ (best[n]), s(best[n])). In case the value of π > total questions, Then Exit.
	(e)	If π is increased in previous steps, Then eliminate all item sets from X_i with support below π.
6	Output best [1] …best[n] would provide the list of the n association rules.	

III. Performance Evaluation metrics

Accuracy & Correlation matrix is used to evaluate the performance of the Predictive Apriori Algorithm. Accuracy is defined as the probability of predicting the rules correctly among all the available rules for the new data. The correlation matrix denotes the significance of the relationship in terms of the probability levels. The value of 0 implies that no relationship exists. The value ≤0.5 implies that the correlation is statistically significant while the value >.0.5 implies that the correlation is not statistically significant.

IV. Result and discussion

The results obtained after the correlation study of the questions asked were illustrated in Figure 1. The Figure

demonstrates that there is a significant relationship between Q1 and Q2 with a value of 0.21 which proves that the customers in Varanasi bought those products that match the celebrity and they believe that the effectiveness of the brand increased whenever the same brand is endorsed by many celebrities over the time. A strong relationship is also observed between Q2 and Q5 with a value of 0.22 that proves that though the customers are sure that the brand is effective if it is endorsed by several celebrities, they stop using the brands if some negative publicity is being made for the celebrity or the celebrity itself is involved in anti-national activities.

	Q1	Q2	Q3	Q4	Q5	Q6
Q1	1					
Q2	0.21	1				
Q3	0.09	0.10	1			
Q4	0.02	-0.02	0.02	1		
Q5	0.05	0.22	0.09	0.02	1	
Q6	0.04	0.13	0.06	-0.05	0.12	1

Figure 1 Correlation Matrix of the questions surveyed.

Figure 2 demonstrates the responses to the research questions in their quartiles. It is clear from the figure that most of the customers from Varanasi strongly agrees that they will stop purchasing the products if the celebrities endorsing them are involved in anti-national activities. They also strongly believe that effectiveness increases if multiple celebrities have endorsed the products over time. They also strongly agree that the products overshadowed by the celebrities would not be purchased. They also strongly believe that the products would not be purchased if they are not aware of the celebrities or the celebrity is from some other region.

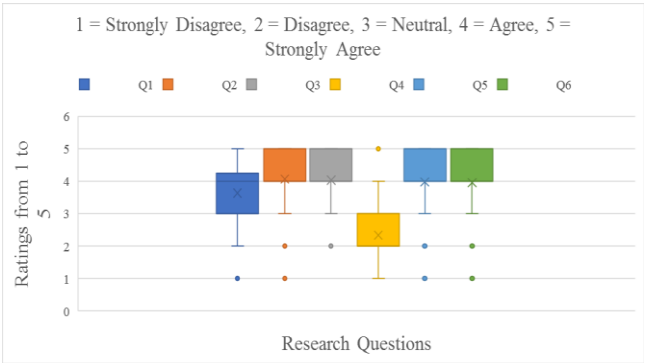

Figure 2 Box Whiskers plot for depicting the responses to their quartiles.

The rules generated and the accuracy of this generation using the Predictive Apriori algorithm is given in Table 4. As described by (Tobias Scheffer, 2001), the algorithm estimates the prior π(c) by extracting several hypotheses randomly to compute their confidence. A fixed number of rules (here 100) were generated using a conditional check on the length of the rules. Mathematically, the probability (Tobias Scheffer, 2001) that i items would be there in the rule from all association rules over n items is given as

$$P[i] = \frac{\binom{n}{i}(2^i - 1)}{\sum_{j=1}^{n}\binom{n}{j}(2^j - 1)} \quad (2)$$

The best rule obtained has an accuracy of 99.34 % and this states that if the customers of Varanasi strongly agreed that they purchase the products that strongly matches the celebrity, the effectiveness increases if more celebrities endorsed the same product over time and that they prefer all the brands endorsed by their favourite celebrities, then they stop purchasing the products once the celebrity has had negative publicity or is involved in anti-national activities.

Table 4 Rules generated and the accuracy of their generation.

S No	Rules	Accuracy(acc)
1	Q1=5 Q2=5 Q6=5 13 ==> Q5=5 13	0.99342
2	Q1=4 Q2=3 12 ==> Q3=4 12	0.99297
3	Q1=3 Q4=2 Q5=4 8 ==> Q3=4 8	0.98776
4	Q2=5 Q3=5 Q4=3 Q6=5 8 ==> Q5=5 8	0.98776
5	Q1=4 Q3=5 Q5=5 7 ==> Q2=5 7	0.98424
6	Q2=4 Q5=3 Q6=3 7 ==> Q3=4 7	0.98424
7	Q3=4 Q5=3 Q6=3 7 ==> Q2=4 7	0.98424
8	Q1=3 Q3=4 Q4=1 6 ==> Q2=4 6	0.97844
9	Q1=3 Q4=1 Q6=4 6 ==> Q2=4 6	0.97844
10	Q1=4 Q3=3 Q4=3 6 ==> Q6=4 6	0.97844
11	Q1=5 Q2=5 Q5=4 Q6=4 6 ==> Q4=3 6	0.97844
12	Q1=3 Q5=4 Q6=4 18 ==> Q2=4 17	0.97607
13	Q1=2 Q4=1 5 ==> Q6=4 5	0.96863
14	Q1=3 Q3=3 Q4=3 5 ==> Q2=4 5	0.96863
15	Q1=3 Q3=3 Q6=4 5 ==> Q2=4 5	0.96863
16	Q1=2 Q3=4 Q4=3 Q5=4 5 ==> Q2=4 5	0.96863
17	Q1=5 Q3=4 Q4=3 Q6=4 5 ==> Q2=5 5	0.96863
18	Q3=4 Q4=1 Q5=4 Q6=4 5 ==> Q2=4 5	0.96863
19	Q1=2 Q3=4 Q6=3 4 ==> Q2=4	0.95163

	4	
20	Q3=4 Q5=5 Q6=3 4 ==> Q2=4 4	0.95163
21	Q1=1 3 ==> Q2=4 Q3=4 3	0.92155
22	Q3=2 3 ==> Q2=5 3	0.92155
23	Q4=4 3 ==> Q2=4 3	0.92155
24	Q1=2 Q6=2 3 ==> Q5=5 3	0.92155
25	Q3=3 Q5=2 3 ==> Q6=4 3	0.92155
26	Q1=3 Q3=5 Q4=2 3 ==> Q5=5 3	0.92155
27	Q1=3 Q5=5 Q6=3 3 ==> Q2=4 3	0.92155
28	Q1=4 Q2=3 Q5=3 3 ==> Q3=4 Q4=3 3	0.92155
29	Q1=3 Q2=4 Q3=5 Q6=5 3 ==> Q4=3 3	0.92155
30	Q1=3 Q4=2 Q5=4 Q6=4 3 ==> Q2=4 Q3=4 3	0.92155
31	Q1=4 Q2=4 Q3=3 Q4=2 3 ==> Q6=4 3	0.92155
32	Q1=5 Q2=5 Q5=5 Q6=4 3 ==> Q3=5 3	0.92155
33	Q2=4 Q4=3 Q5=3 12 ==> Q3=4 11	0.90934
34	Q2=5 Q4=3 Q6=5 12 ==> Q5=5 11	0.90934
35	Q2=1 2 ==> Q1=4 Q4=2 2	0.86736
36	Q2=1 2 ==> Q1=4 Q5=3 2	0.86736
37	Q5=1 2 ==> Q3=4 2	0.86736
38	Q6=1 2 ==> Q3=3 Q4=2 2	0.86736
39	Q1=1 Q4=3 2 ==> Q5=4 2	0.86736
40	Q1=1 Q5=4 2 ==> Q4=3 2	0.86736
41	Q1=1 Q6=4 2 ==> Q4=3 2	0.86736
42	Q1=4 Q2=2 2 ==> Q3=4 2	0.86736
43	Q1=4 Q3=2 2 ==> Q6=4 2	0.86736
44	Q1=4 Q4=4 2 ==> Q2=4 Q5=4 2	0.86736
45	Q1=5 Q2=2 2 ==> Q4=2 2	0.86736
46	Q2=2 Q4=3 2 ==> Q1=3 2	0.86736
47	Q2=2 Q5=3 2 ==> Q1=3 2	0.86736
48	Q2=3 Q4=1 2 ==> Q6=4 2	0.86736
49	Q3=2 Q4=3 2 ==> Q2=5 Q5=4 2	0.86736
50	Q3=2 Q5=4 2 ==> Q2=5 Q4=3 2	0.86736
51	Q3=2 Q6=4 2 ==> Q1=4 2	0.86736
52	Q1=3 Q2=5 Q4=3 2 ==> Q3=5 2	0.86736
53	Q1=3 Q3=4 Q6=2 2 ==> Q5=4 2	0.86736
54	Q1=4 Q3=3 Q5=2 2 ==> Q4=3 2	0.86736
55	Q1=4 Q3=3 Q5=3 2 ==> Q6=4 2	0.86736
56	Q1=4 Q3=5 Q6=2 2 ==> Q2=5 2	0.86736
57	Q1=5 Q3=3 Q4=1 2 ==> Q2=4 2	0.86736
58	Q1=5 Q4=1 Q5=2 2 ==> Q6=4 2	0.86736
59	Q1=4 Q2=3 Q5=4 Q6=4 2 ==> Q3=4 Q4=3 2	0.86736
60	Q1=5 Q2=5 Q4=1 Q5=3 2 ==> Q6=4 2	0.86736
61	Q1=5 Q2=5 Q4=1 Q5=5 2 ==> Q3=5 2	0.86736
62	Q2=4 Q3=4 Q5=2 Q6=4 2 ==> Q4=1 2	0.86736
63	Q2=5 Q3=5 Q5=3 Q6=3 2 ==> Q4=3 2	0.86736
64	Q1=2 Q3=4 Q5=4 10 ==> Q2=4 9	0.84985
65	Q1=5 Q3=5 Q6=5 10 ==> Q5=5 9	0.84985
66	Q4=1 Q5=4 Q6=4 10 ==> Q2=4 9	0.84985
67	Q1=5 Q4=2 Q6=5 9 ==> Q5=5 8	0.80978
68	Q2=5 Q3=5 Q6=5 17 ==> Q5=5 15	0.78897
69	Q2=4 Q5=4 Q6=3 8 ==> Q4=3 7	0.7641
70	Q1=3 Q6=4 25 ==> Q2=4 21	0.75179
71	Q3=4 Q5=4 Q6=4 41 ==> Q2=4 32	0.72942
72	Q1=3 Q2=4 Q3=4 Q6=4 14 ==> Q5=4 12	0.72809
73	Q2=4 Q3=4 Q4=3 Q5=4 22 ==> Q6=4 18	0.72169
74	Q1=3 Q5=4 29 ==> Q2=4 23	0.72009
75	Q2=5 Q6=5 29 ==> Q5=5 23	0.72009
76	Q5=4 Q6=4 66 ==> Q2=4 49	0.71639
77	Q3=3 Q6=3 7 ==> Q2=4 6	0.71521
78	Q1=4 Q2=4 Q5=3 7 ==> Q3=4 6	0.71521
79	Q1=4 Q2=4 Q6=5 7 ==> Q3=4 6	0.71521
80	Q1=4 Q3=5 Q6=5 7 ==> Q2=5 6	0.71521
81	Q1=5 Q2=5 Q4=2 Q5=5 7 ==> Q6=5 6	0.71521
82	Q1=5 Q2=5 Q4=3 Q5=4 7 ==> Q6=4 6	0.71521
83	Q1=5 Q6=5 21 ==> Q5=5 17	0.71004
84	Q1=3 Q2=4 Q6=4 21 ==> Q5=4 17	0.71004
85	Q1=3 Q4=1 13 ==> Q2=4 11	0.70652
86	Q1=4 Q4=3 Q5=4 20 ==> Q6=4 16	0.69745
87	Q3=3 Q5=4 26 ==> Q2=4 20	0.69129
88	Q1=4 Q3=3 16 ==> Q6=4 13	0.68844
89	Q2=4 Q4=3 Q6=4 35 ==> Q5=4 26	0.68706
90	Q2=4 Q3=4 Q6=4 44 ==> Q5=4 32	0.68456
91	Q2=3 Q4=3 19 ==> Q3=4 15	0.6839

92	Q1=4 Q2=4 Q4=3 19 ==> Q6=4 15	0.6839
93	Q3=5 Q5=5 Q6=5 19 ==> Q2=5 15	0.6839
94	Q3=5 Q6=5 25 ==> Q5=5 19	0.68008
95	Q3=4 Q4=3 Q5=4 34 ==> Q6=4 25	0.67798
96	Q2=4 Q3=4 Q5=4 45 ==> Q6=4 32	0.66894
97	Q1=2 Q3=4 24 ==> Q2=4 18	0.66808
98	Q1=2 Q6=3 6 ==> Q2=4 5	0.66576
99	Q1=2 Q6=3 6 ==> Q4=3 5	0.66576
100	Q2=2 Q3=4 6 ==> Q4=2 5	0.66576

V. CONCLUSION

The current work has explored the importance of sentiment analysis on marketing analytics. The work has utilized the benefits of one of the association rule mining algorithms, Predictive Apriori to understand the sentiments of the Varanasi customers. The algorithm has performed with an accuracy of 99.34 % which proved that they stop purchasing the products once the celebrity has had negative publicity or is involved in anti-national activities. The work has also explored the correlation between the questions being asked in the survey and found that they strongly agree that the brand is effective if it is endorsed by several celebrities over a while.

REFERENCES

Agrawal, R. (2013). Fast Algorithms For Mining Association Rules In Datamining. *International Journal of Scientific & Technology Research*, *2*(12), 13–24.

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining Association Rules Between Sets of Items in Large Databases. *ACM SIGMOD Record*, *22*(2), 207–216. https://doi.org/10.1145/170036.170072

Kumar, S., Pandey, M. K., Nath, A., & Subbiah, K. (2016a). Missing QoS-values predictions using neural networks for cloud computing environments. *2015 International Conference on Computing and Network Communications, CoCoNet 2015*, 414–419. https://doi.org/10.1109/CoCoNet.2015.7411219

Kumar, S., Pandey, M. K., Nath, A., & Subbiah, K. (2016b). Performance Analysis of Ensemble Supervised Machine Learning Algorithms for Missing Value Imputation. *Proceedings - International Conference on Computational Intelligence and Networks*, *2016-Janua*, 160–165. https://doi.org/10.1109/CINE.2016.35

MK Pandey, S. K. (2017). *Performance analysis of time series forecasting of ebola casualties using machine learning algorithm*. *2*, 885–898.

Pandey, M. K., Kumar, S., & Karthikeyan, S. (2013). *Information Security Management System (ISMS) Standards in Cloud Computing-A Critical Review*.

Pandey, M. K., & Subbiah, K. (2016). Social networking and big data analytics assisted reliable recommendation system model for internet of vehicles. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 10036 LNCS*. Springer Verlag. https://doi.org/10.1007/978-3-319-51969-2_13

Pandey, M. K., & Subbiah, K. (2018). *Performance Analysis of Time Series Forecasting Using Machine Learning Algorithms for Prediction of Ebola Casualties*. 320–334. https://doi.org/10.1007/978-981-13-2035-4_28

Pandey, M. K., & Subbiah, K. (2017). A novel storage architecture for facilitating efficient analytics of health informatics big data in cloud. *Proceedings - 2016 16th IEEE International Conference on Computer and Information Technology, CIT 2016, 2016 6th International Symposium on Cloud and Service Computing, IEEE SC2 2016 and 2016 International Symposium on Security and Privacy in Social Netwo*, 578–585. https://doi.org/10.1109/CIT.2016.86

Scheffer, T. (2005). Finding association rules that trade support optimally against confidence. *Intell. Data Anal.*

Scheffer, Tobias. (2001). Finding association rules that trade support optimally against confidence. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *2168*, 424–435. https://doi.org/10.1007/3-540-44794-6_35

Zaki, M. J. (1999). Parallel and Distributed Association Mining: A Survey. *IEEE Concurrency*, *7*(4), 14–25.

***