

Automatic Ratings Generation System for Behavior Analysis



Poonam Sawant, Abhijeet Kaiwade

Abstract: User Behavior Analysis plays a pivotal role in finding the behavior of an individual regarding any certain business or social issue. It is helping a business to take an appropriate decision for improvement by understanding customers' opinion about their products or services in a positive or negative way. Although there are many systems have been developed and implemented till date for performing behavior analysis in different ways; still better advancement is needed due to the nature of the data. In this paper, we have developed a system to generate automatic ratings based on the comments given by the banking customers on social networking sites. These ratings then further analyzed to find out positive and negative behavior. PIG ETL tool on the top of MapReduce is used to develop and implement a proposed system and performing the analysis. AFINN sentiment lexicon is used to generate automatic ratings whereas visualization is done using D3.js. Performance evaluation of the proposed system is done by comparing it with the existing system.

Keywords: Behavior Analysis, AFINN, Pig, Hadoop, MapReduce, D3.js, Positive and Negative Polarity.

I. INTRODUCTION

Knowing customer is one of the major challenges facing by banking industry in India as well as in the world. Customers give feedbacks, comments and suggestions on social networking sites which influence others. Behavior analysis is the way to contextually mining such a data to identify social sentiments about banking products and services and helping them to optimize their business. Automatic rating generation is an effective way to find out the behavior of customers based on their comments. As banks have been storing huge amount of data in their databases and it is growing further due to digitization; it is valuable, but they are facing big data challenges as traditional data analytics are not capable of handling such kind of data. Big Data Analytics helps to turn big data into big value by allowing organizations to analyze customer behavior by improving their services [2-5]. Many big data technologies are used for building big data analytics to understand user behavior that may be used for good decision-making. We found Pig an ETL tool on the top of the MapReduce suitable for handling our data in easy and efficient manner by improving performance [6] [8-9].

Revised Manuscript Received on October 20, 2020.

*Correspondence Author

Dr. Poonam Sawant*, Asst. Professor, Sinhgad Institute of Management & Computer Application Pune, India Email: poonam25m@gmail.com

Dr. Abhijeet Kaiwade, Professor & HOD, Sadhu Vaswani Institute of Management Studies for Girls, Pune, India Email: abhijeet.kaiwade-hod@svims-pune.edu.in

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

AFINN dictionary contains list of positive and negative words with weightage helping in generating automatic ratings. Customer behavior analysis is incomplete without visualization so addition to this D3.js technology is used to plot the graphs.

II. EXISTING WORK: PIG DATA ANALYTICS FRAMEWORK FOR BEHAVIOR ANALYSIS WITH ORIGINAL RATINGS

Till date many big data Analytics have been developed and implemented for analyzing customer behavior in different ways. Every technology has its own advantages and disadvantages [1][7][10-13]. In [14] we have developed a behavior analytics using pig tool on the top of MapReduce with word count programming concept to analyze customer behavior using original ratings and reviews. The system is then further extended to find out positive and negative behavior of the customers. The overall structure of developed methodology is depicted in the following figure. Approximately 100000 comments of each SBI and HDFC banks customers regarding home loan service are used to test the system.

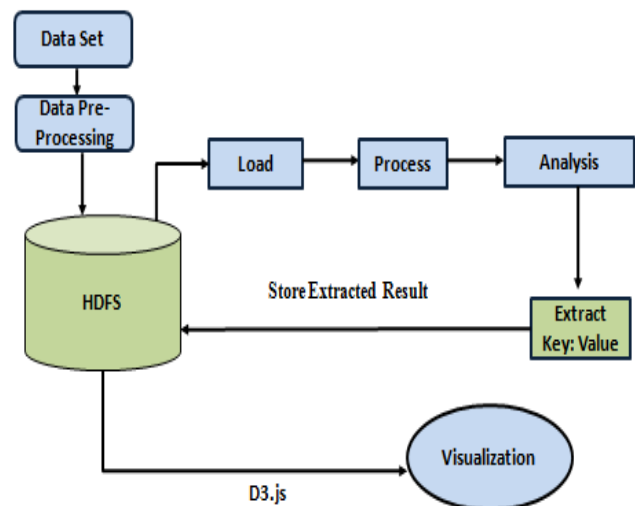


Fig.1: Proposed Data Analysis Framework with Original Rating.

By implementing the above system we have observed following results.

- A. List of ratings and reviews by customers.
- B. Total no. of customers w.r.t ratings
- C. Positive and Negative Behavior based on the ratings given.



D. Positively and Negatively Influenced Customers.

A. List Of Ratings and Reviews Given By Customers

At this point various original ratings along with reviews are identified as shown in Table-I and fig.2 which can be used for analyzing customer behavior precisely.

Table-I: Ratings and Reviews

Review & Ratings		
Sr. No.	Original Rating	Reviews
1	0.5	Unacceptable
2	1.0	Really Bad
3	1.5	Bad
4	1.8	Poor
5	2.0	Not Good Services, Expected More
6	2.5	Poor, Just Ok
7	2.8	Just Ok
8	3.0	Satisfactory
9	3.3	Satisfactory
10	3.5	Good, Pretty Good
11	3.8	Fine
12	4.0	Great
13	4.3	Good
14	4.5	Excellent, Great
15	4.8	Great, Good
16	5.0	Excellent, Blown Away



Fig.2: Ratings and reviews

Results drawn reveals that, customers ratings regarding to the services varies in the scale 0.5 to 5.0. Laterally they have given their short opinions like “Blown Away”, ”Great”, ”Bad” and so on. It has been also observed that, Each and every rating has different reviews but at the same time some of them have same. E.g. for ratings 3.5 and 4.3 “Good” remark is given. Similarly “Great” remark is given for rating 4.0, 4.5 and 4.8 whereas some ratings have multiple remarks. E.g. 2.0 having remarks like “not good services” and “expected more”.

B. Total No. of Customers w.r.t. Ratings

Here we found total number of customers who rate home loan service of SBI bank & HDFC bank respectively as shown in Table-II and Fig.3.

Table-II: Total No. of Customers w.r.t to Ratings

Ratings					
Rating	No. of Customer		Rating	No. of Customer	
	SBI	HDFC		SBI	HDFC
0.5	764	751	3.3	5197	5733

1.0	3528	3423	3.5	2286	2547
1.5	512	829	3.8	2784	2201
1.8	1532	4486	4.0	35569	34683
2.0	2547	3759	4.3	847	680
2.5	1571	1680	4.5	2327	1077
2.8	2043	2245	4.8	881	662
3.0	8017	10625	5.0	29595	24619

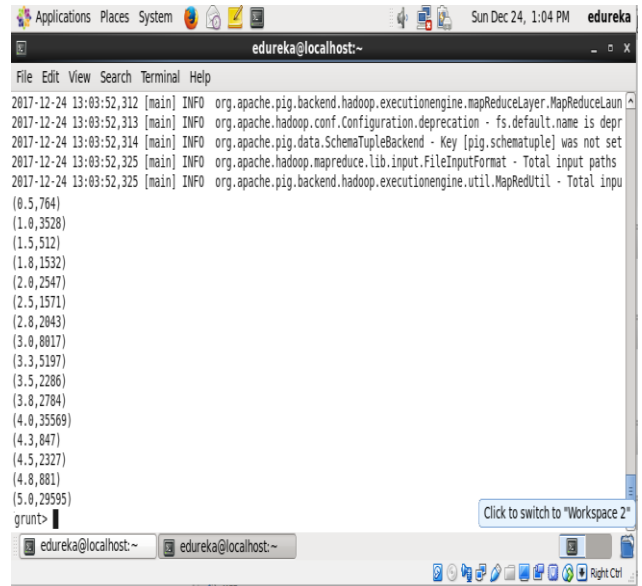


Fig. 3: Ratings Given by No. of Customers

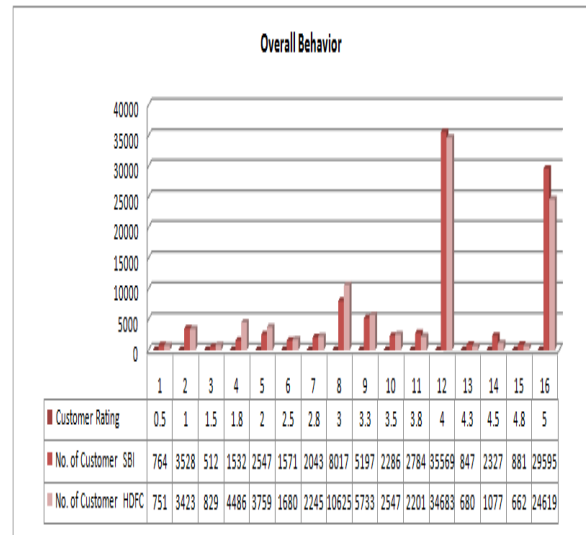


Fig.4: Ratings

From the above graph shown in Fig.4 we can conclude that, the highest number of customers have given rating 4.0 followed by 5.0 for both the banks.



C. Positive and Negative Behavior Analysis Based on Ratings

Simple K-mean clustering algorithm is applied to analyze positive and negative behavior based on original ratings. To minimizing total intra-cluster variance is the main objective of K-Means clustering.

$$J = \sum_{j=1}^k \sum_{i=1}^n \|x_i^{(j)} - c_j\|^2$$

Algorithm:

1. Clusters the data into predefined *k* groups.
2. Randomly select *k* points as cluster centers.
3. Assign objects to their closest centers according to the *Euclidean distance* function.
4. Assign objects to their closest cluster center according to the *Euclidean distance* function.
5. Until the same points are assigned to each cluster in consecutive rounds, repeat steps 2, 3 and 4.

Cluster-1 {0.5, 1.0, 1.5, 1.8, 2.0, 2.5, 2.8}
Cluster-2 {3.0, 3.3, 3.5, 3.8, 4.0, 4.3, 4.5, 4.8, 5.0}

Table-III: Positive and Negative Behavior

Positive and Negative Behavior		
Type of Behavior	SBI	HDFC
Positive Behavior	87503	82827
Negative Behavior	12497	17173
Total No. of Customers	100000	100000

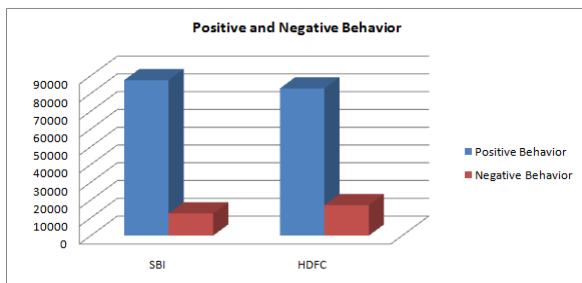


Fig. 5: Positive and Negative Behavior

Table-III and Fig. 5 give a count of positive and negative behavior of SBI and HDFC banks. It has been observed that, the positive behavior of SBI bank's customers is slightly higher than HDFC bank's customers and negative behavior of SBI bank's customers is slightly lower than HDFC bank's customers.

D. Positively and Negatively Influenced Customers

The total number of existing or non-existing customers influenced positively or negatively from the feedback given by the existing customers are observed as shown in Table-IV and Fig. 6. From the graph shown in Fig.7, we can conclude that positively influenced customers are more than negatively influenced customers in both the banks.

Table -IV: Influenced Customers

Influenced Customers		
Customers	Positively Influenced	Negatively Influenced
SBI	25236	14527

HDFC	24523	19555
------	-------	-------

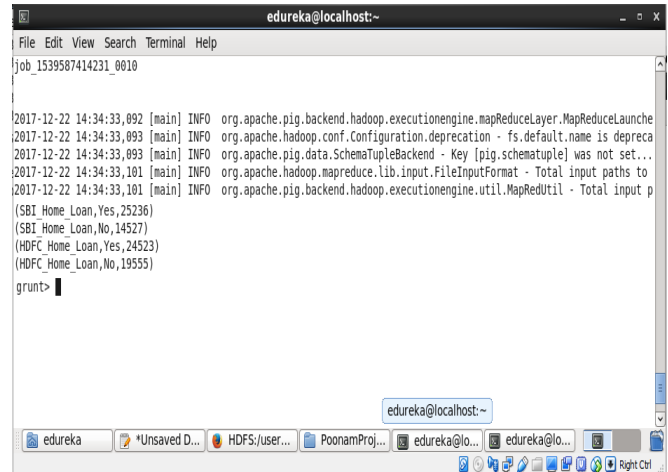


Fig. 6: Influenced Customers

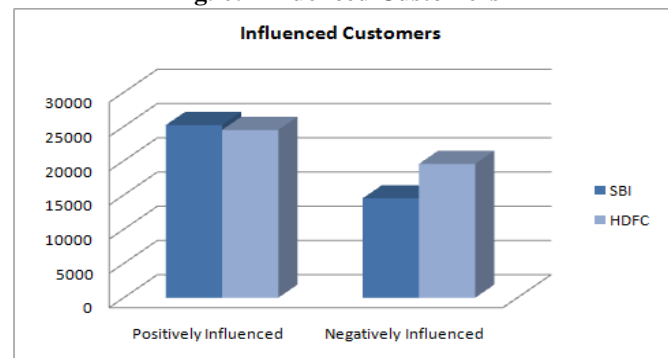


Fig.7: Influenced Customers

III. LIMITATIONS OF EXISTING ANALYTICS

By considering the positive behavior analysis as shown in Table-III, on an average 85% customers are seems to be happy with the services. Though they are pleased with the service we cannot verdict about their continuation and discontinuation as a lots of comments and reviews are in contrast; some of them we have listed in the below table.

Table-V: Ratings and Comments in Contrast

Polarity	
Ratings	Comments
2.0	Excellent Service
4.0	Poor Service
5.0	Should Improve Services
1.0	Fine
2.5	Great
4.5	Provide better services
3.5	Bad Services

Such kind of reviews and comments confuse the new customers. Also, some datasets contains only comments so it is very difficult to find out the ratings and exact behavior of the customers. Every data set contains different ratings and reviews. In a conclusion, data is not in a standard format. To find out exact positive and negative behavior of customers generation of automatic ratings is important.

Therefore, in the next session we have proposed Automatic Rating Generation System to generate automatic ratings based on the positive and negative words in the comments using AFINN dictionary and Pig Behavior Analytics.

IV. OBJECTIVES OF THE PROPOSED STUDY

To extend the research further, following objectives are defined.

1. To develop a system for generating automatic ratings and finding positive and negative behavior
2. To compare Existing Analytics and proposed Analytics

V. HYPOTHESIS

To check the effectiveness of the proposed system following hypothesis is formed.

H1: There is a significance difference between the customer behavior analyzed using original ratings and automatic ratings with Pig Tool.

VI. PROPOSED SYSTEM

In this section, we have proposed “Automatic Rating Generation System” for generating automatic ratings using comments and analysing customer behavior based on that. We have proposed an algorithm based on the proposed framework mentioned in Fig.8 with addition of AFINN dictionary and then implemented the algorithm to find out desired output.

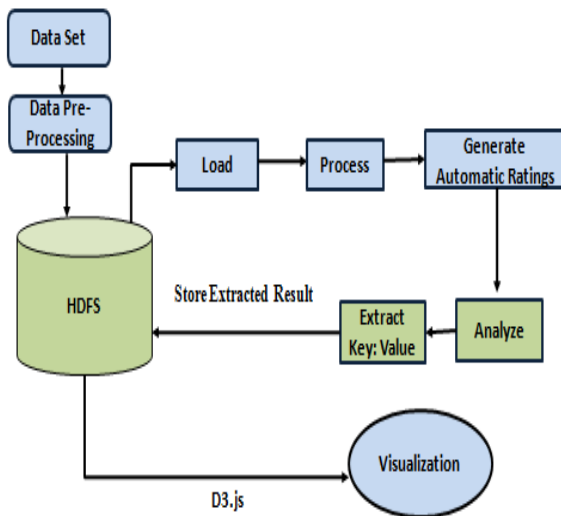


Fig.8: Proposed Framework for Automatic Ratings Generation and Data Analysis

A. Data pre-processing:

The data set collected from the internet sites was consists of noise and incomplete data, using pre-processing techniques noisy and inconsistent data is reduced and converted into CSV format. In the consideration of nature of data, in the initial stage, researcher has used multiple data sets individually to draw conclusions.

B. Proposed Algorithm

In this section we have proposed an algorithm based on the system.

Input: Input file in CSV Format

Output: Behavior Type

Begin

Step-1: Transfer data to HDFS using Hadoop *put* command.

Step-2: Use pig command to start grunt shell.

Step-3: Load data from HDFS using PigStorage () in A

Step-4: If variable = Not comments then

uc1 = select variables from A

uc2 = filter uc1 by criteria(condition)

uc3 = group uc2 by Column_Name;

test = group uc3 and count word

store test into HDFS

else

uc1= Select comments from A, split and change structure of tuple

Dictionary= Load dictionary of positive and negative words

auto_rating = join uc1 and Dictionary using left outer join

rating = select comments and rating from auto_rating

word_group = group rating by (comments);

avg_rate = generate group for each

word_group, and average ratings

if avg_rate>0

 recordpositive_behavior

if avg_rate<=0 then

 recordnegative_behavior

store positive_behavior and negative_behavior into HDFS

Step_5: Visualize Results using D3.js

End

The process of automatic rating generation and data analysis is accomplished in 5 steps as mentioned in above algorithm. Data is stored in HDFS using Hadoop put command. Pig command is used to start grunt shell. Then load command is used to load the data from HDFS using PigStorage() and comma delimiter in a SQL format. The transformation statements like foreach, generate, filter and group are used to extract required information. Function COUNT () is used to count the number of occurrences, TOKENIZE () is used to split a string, AVG () is used to count average and FLATTEN operator is used to change the structure of tuple. Finally dump and store commands are used to read and save processed data respectively.

C. Experimental Results

By implementing an algorithm mentioned in above section following results have been drawn,

1. Automatic Rating to find out Positive Behavior shown in table 6 and figure 9
2. Automatic Rating to find out Negative Behavior shown in table 7 and figure 9 and
3. Positive and Negative Behavior with Missing values shown in table 8

1. Automatic Rating to find out Positive Behavior

Table-VI shows sample comments which are positively rated.



Table-VI:Auto-Generated Positive Rating

Positive Comments with Automatic Ratings	
Comment	Rating
From SBI bank i took the housing loan which was disbursed within 15 days Documentation was easy as submitted only the house plan and the reregistration proof The process was More)	1
SBI offered me the best home loan with an interest rate of 835% for my home loan application along with zero processing fee Recently I had approached them for home loan They More)	2
SBI executives are not cordial to the customerThey are not understanding the situation and giving a proper solutionsThey offer better rates compared with a private bankCharges More)	2
SBI helped to get my dream house They done very fast process on my documentations and they disbursed the amount within 20 days of time SBI offered very attractive interest rate More)	1
I took the home loan processed by State bank of India in the year 2004 The executives took around 45 days to process the home loan I felt that the response was not good from the More)	3
It takes long time for SBI to process the Home loan applicationThere is communication gap between the customer and Bank based on the Home loanThey always offer the best rate for More)	3
It was almost close to perfect with SBI based on the Home loan processThey have kept to the words which was promised by themThey have taken around 4-5days to process the loanThe More)	2
The state bank of India service is always good based on the Home loan processThey are able to send the mail regarding IT certificate in a immediate mannerThey are require lesser More)	3

2. Automatic Rating to find out Negative Behavior

Table-VI shows sample comments which are negatively rated.

Table-VII: Auto-Generated Negative Ratings

Negative Comments with Automatic Ratings	
Comment	Rating
Interest rate offered by SBI bank for the home loan was 975% which was less comparing with hdfc bank and the processing fees charged is Rs 5700The response was very prompt More)	-3
My home loan application with SBI Bank have not processed throughIt takes long time for their process to be completedThere is no clear communication between their employees More)	0

Availed home loan for the processing charge of just 02% With SBI there is only one problem as customer has to run behind them to complete the documentation procedures Pretty low More)	-2
Within one week my home loan was processed by sbi bank as i approached the bank directly Rate of interest charged was 935% and the collected less processing fee All the required More)	-1
There was no proper response from the SBI bank when i took a home loan They made us visit the branch many times by not guiding properly with the required documents which made very More)	-1
I approached them SBI Bank for an Home loan last 5 to 6 months back They are providing long term tenure and also charges are very less Good thing is compare to other banks SBI Bank More)	-2
Very critical to take a loan from SBI bank because they more documentation process and verification My home loan disbursement took 25 months Interest rate charged was 1025% and the More)	-3
I was postponed the decision to take home loan right now because of slow processing. I am planning to take it after two or three months later.	-1

3. Positive and Negative Behavior

The auto-rated positive and negative comments then compiled in Table-VIII to extend the study for behavior analysis of the customer’s w.r.t. SBI and HDFC banks respectively. Along with auto-generated positive and negative ratings table contains missing values too which are not rated. From the graph shown in Fig.9, it has been concluded that, the positive behavior of SBI customers is slightly higher than HDFC customers and negative behavior of SBI customers is slightly lower than HDFC customers.

Table-VIII: Positive and Negative Behavior using Automatic Ratings

Comments	Automatic Ratings	
	SBI	HDFC
Positive Comments	60441	58357
Negative Comments	34713	37219
Missing comments	4846	4424

(Compiled by Researcher by referring the automatic positive and negative ratings as shown in Table-VI and Table-VII)

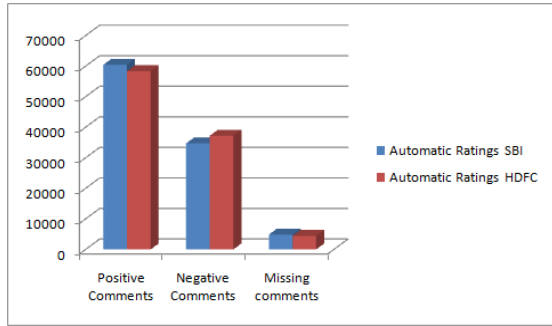


Fig. 9: Positive and Negative Behavior with automatic ratings

VII. COMPARATIVE STUDY OF BEHAVIOR ANALYSIS

This section deals with the comparative study of behavior analysis using existing analytics and proposed analytics w.r.t. SBI and HDFC banks. Hypothesis testing is done using chi-square test.

A. Results and Interpretation

The main aim of the study is to predict and analyze the customer behavior regarding home loan service provided by SBI and HDFC banks.

Table-IX is compiled by researcher by referring Table-II and Table-VIII from section II and VI respectively, presents positive and negative behavior of bank customers. Table-X shows the compiled data from Table-IX which represents the result inpercentage. Fig.10 shows the graphical representation of positive and negative behavior of SBI and HDFC bank’s customers through original rating and automatic ratings.

From the Fig.9, Table-IX and Table-X, it has been concluded that, the positive behavior is decreased by 27.1% in SBI and 24.5% in HDFC banks whereas negative behavior is increased by 22.2% and 20% in SBI and HDFC respectively. Also the positive behavior of SBI customers is slightly higher than HDFC customers.

Table-IX: Comparative Behavior with original and automatic ratings

Comparative Behavior						
B	O.R		A.R		E.R	
	S	H	S	H	S	H
P	87503	82827	60441	58357	27062	24470
N	12497	17173	34713	37219	22216	20046

(Data Compiled by researcher by referring Table-III and Table-VII)

(*B=Behavior, *O.R.= Original Ratings, * A.R. Automatic Ratings, * E.R.= Error Rate, * S=SBI, * H=HDFC, *P=Positive, *N= Negative)

Table-X: Comparative Behavior with original and automatic ratings compiled from Table-IX

Compiled Data						
B	O.R (%)		A.R (%)		E.R (%)	
	S	H	S	H	S	H
P	87.5	82.8	60.4	58.4	27.1	24.5
N	12.5	17.2	34.7	37.2	22.2	20.0

P	87.5	82.8	60.4	58.4	27.1	24.5
N	12.5	17.2	34.7	37.2	22.2	20.0

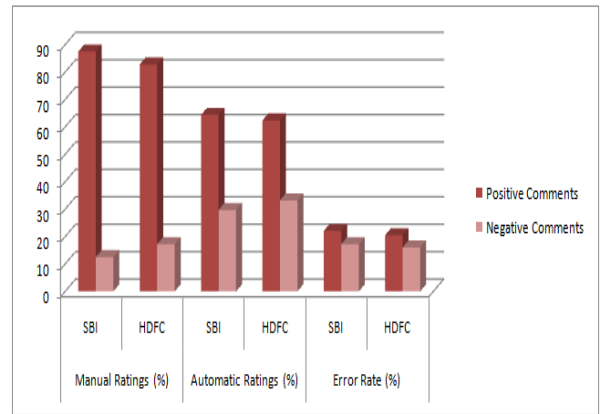


Fig. 10: Comparative Study of SBI and HDFC with Original and Automatic Ratings

B. Hypothesis Testing

To perform the comparative analysis hypothesis testing is done using data from Table-X. To test the hypothesis statistically Chi-Square test was performed,

- To test the relationship between categorical variables.
- To test the significant relationship among two or more than two categorical variables.

Formula of chi-square test is as follows:

$$\chi^2 = \sum (f_o - f_e)^2 / f_e$$

where f_o = the observed frequency (the observed counts in the cells) and f_e = the expected frequency (if, NO relation between variables).

$$\chi^2 = \frac{\sum (\text{Observed} - \text{expected})^2}{\text{expected}}$$

This test is based on the difference between what is to be observed in the data and what is expected if there is no relationship between the two variables.

H1: There is a significance difference between the customer behavior analyzed using original ratings and automatic ratings with Pig Tool.

H0: There is a no significance difference between the customer behavior analyzed using original ratings and automatic ratings with Pig Tool.

To test the above hypothesis, the data from table 10 is converted into categorical data and chi-square test is applied. Here researcher has added SBI and HDFC data together for both the categories and both the groups.

Table-XI: Compiled data from table 10

Compiled Data for Hypothesis			
Behavior	Original Rating	Automatic Rating	Row Total
Positive	170	119	289
Negative	30	72	102
Column Total	200	191	391 (G.T)



Table-XII: Testing of Hypothesis

Chi-Square Statistic							
Behaviors	Original Rating			Automatic Rating			Row Total
	Observed Value.	Expected Value	Chi-Square Statistic	Observed Value	Expected Value	Chi-Square Statistic	
P	170	147.83	3.33	119	141.17	3.48	289
N	30	52.17	9.42	75	49.83	9.87	102
C.T	200			191			391 (G.T)

(Source:

<https://www.socscistatistics.com/tests/chisquare2/Default2.aspx>)

The Chi-square statistic is 26.1008. The p-value is .000000. The result is significant at $p < .05$

(*P=Positive, *N=Negative, *C.T.= Column Total, *G.T= Grand Total)

Table-XI shows the compiled data for category original rating and automatic rating and positive and negative behavior groups. Table-XII shows the testing of hypothesis. From table 12 we can see that, Chi-square statistic is 26.1008 and p-value is 0.000000. If the $p\text{-value} < 0.05$ then the null hypothesis H_0 is rejected and the alternate hypothesis H_1 is accepted. In above case the $p\text{-value} 0.000000 < 0.05$ which shows the significant result. Therefore, here null hypothesis H_0 is rejected and H_1 is accepted and concludes that, there is a significance difference between the customer behaviors analyzed using original ratings and automatic ratings with Pig Tool.

VIII. EXPERIMENTAL RESULTS

The present research is confined to analyze customer behavior based on original reviews given as well as by generating automatic ratings using the comments given by customers on product review sites. Analysis of original ratings (section II) concludes that,

- Though average 85% customers are happy with the services, almost 35% to 40 % customers have recorded their reviews and ratings in contrasts which may influence new customers negatively.

Analysis of automatic generated ratings (section VI) states that,

- Approximately 59% customers are happy with the services and approximately 35% customers are seems to be unhappy whereas approximately 6% customers are neither evaluates as positive not negative.
- Proposed system performs better than existing system as it works on positive and negative words.

In General proposed system is able to,

- Explore customer's hidden behavioral patterns from large amount of unstructured data.
- Help banks to bridge gap between what customers want and what they actually get.

- Help customers to select best bank for taking home loan.

IX. CONCLUSION

This paper deals with the understanding of customer behavior using reviews and ratings available on websites. To achieve this goal researcher has first developed Pig Data Analytics Framework mentioned in section II to analyse original ratings and then positive and negative behavior is counted using K-Mean clustering. Due to contrast comments and ratings this framework is unable to analyse data correctly. To overcome this problem researcher has proposed Automatic Rating Generation System mentioned in section VI using Pig and AFINN dictionary and predicted automatic ratings to find out positive and negative behavior of customer. This framework performs well than existing system but along with positive as well as negative ratings it provides some missing values too. The AFINN dictionary does not generate standard ratings. In a result correct prediction is not possible by proposed system. Therefore for accurate prediction and analysis researcher has proposed hybrid predictive analytics using Pig, Spark and Machine Learning Algorithms as a future work.

REFERENCES

1. Joseph O. Chan (2013), An Architecture for Big Data Analytics, communications of the IIMA, Volume 13 Issue 2
2. Vignesh Prajapati (2013), Big Data Analytics with R and Hadoop, PACKT, ISBN-978-1-78216-328-2
3. Tom White (2014), Hadoop The Definitive Guide, O'REILLY, ISBN 13:978-93-5023-756-4
4. Nada Elgendy, Ahmed Elragal (2014), Big Data Analytics: A Literature Review Paper, ICDM, LNAI 8557, 214-227
5. Utkarsh Shrivastava, Santosh Gopalkrishnan (2015) Impact of Big Data Analytics on Banking Sector: Learning for Indian Banks, Procedia Computer Science, Elsevier ScienceDirect, 643-652
6. J. Ramsingh, Dr. V. Bhuvanewari (2015), An Insight on Big Data Analytics Using Pig Script, IJETCS, Volume 4, Issue 6, ISSN 2278-6856
7. Vikhyat Gupta, Tarik Taeib (2015), Analyzing User Behavior Using MapReduce, JMEST, Volume 2, Issue 3, ISSN 3159-0040, 384-387
8. MapReduce vs. Pig vs. Hive - Comparison between the key tools of Hadoop, 01, sept 2015, <https://www.dezyre.com/article/mapreduce-vs-pig-vs-hive/163>
9. Subramaniya swamy V, Vijayakumar V, Logesh Rc and Indragandhi V (2015), Unstructured Data Analysis on Big Data using Map Reduce, Procedia Computer Science, 456 – 465
10. MohdRehan Ghazi, Durgaprasad Gangodkar (2015), Hadoop, MapReduce and HDFS: A Developers Perspective, Elsevier ScienceDirect, 45 – 50
11. Rajendra Akerkar, Sajja (2016), Intelligent Techniques for Data Science, Springer, ISBN 978-3-319-29205-2
12. Anindita A Khade (2016), Performing Customer Behavior Analysis using Big Data, Procedia Computer Science, Elsevier ScienceDirect, 986-992
13. A. Shrivastava, Chandan Kumar, Neha Mangla (2016), Analysis of Diabetic Dataset and Developing Prediction Model by using Hive and R, Indian Journal of science and Technology, Vol 9(47), ISSN 0974-6846
14. Poonam Sawant, Dr. Abhijeet Kaiwade, Dr. S.D. Mundhe (2018), Pig: Novel Approach to Analyze Customer Behavior in Banking

AUTHORS PROFILE



Dr. Poonam Sawant is an Assistant Professor of Computer Application in Sinhgad Institute of Management and Computer Application, Pune, India. She has pursued her B.Sc. degree from Mumbai University, MCA from Shivaji University, M.Phil in Computers from Shivaji University and Ph.D in Computers from Savitribai Phule Pune University. Her research interests are AI & Machine Learning, NLP, Big Data Analytics and Web Mining. Her Papers are published in peer review journals including International Journal of Innovative Research in Computer and Communication Engineering, International Journals of Multidisciplinary Research Academy (IJMRA), International Journal of Emerging Technology and Advanced Engineering, IPASJ International Journal of Computer Science (IJCS) and many more.



Dr. Abhijeet Kaiwade is Professor and HOD in Computer Department of Sadhu Vaswani Institute of Management Studies for Girls, Pune, India. He is a Ph.D Guide in Savitri Bai Phule Pune University and has guided numerous students. He has published his research in various International Journals and Conferences. His area of Specialization is Data Mining and Big Data.