Adopting K-Mean Clustering Algorithm on Multiple Choices Feedback Opinion Mining

J. Sheik Mohamed*
Asst. Professor, MCA Department
SITAMS, Chittoor, A.P., Indi
Sheik_50@yahoo.co.in

P. Nirmala
MCA Department
SITAMS, Chittoor, A.P., India
pr_nimmi@yahoo.com

Dr. P. Chitti Babu
Professor
M CA Department
AITS, Rajampet, A.P., India

Dr. S. Ramakrishna
Professor & BOS
Department of Computer Science
S. V. University, Tirupathi, A.P., India

Abstract: With the dramatically increase of e-commerce product information gradually forms a massive scale. The product information that user real needs can't be searched and acquired quickly and easily, so user spends lots of time in eliminating the useless information. Product opinion mining, namely analyze the user's opinion of products, can find out some of characteristics of the product itself and whether user is satisfied with the products characteristics or not. According to the description of product characteristics one can draw some information which is fundamental to user's concern and used them for reliable recommendation in further sales decision making.

Keywords: Opinion Mining, K-Mean Clustering, Customer, Arff File,

I. INTRODUCTION

Given a set of evaluative text documents that contain opinions about an object, opinion mining aims to extract attributes and components of the object that have been commented on in each document and to determine whether the comments are positive, negative or neutral.

Before the Web, when an individual needs to make a decision, he/she typically asks for opinions from friends and families. When an organization needs to find opinions of the general public about its products and services, it conducts surveys and focused groups.

With the Web, especially with the explosive growth of the user generated content on the Web, the world has been changed. One can post reviews of products at merchant sites and express views on almost anything in Internet forums, discussion groups, and blogs, which are collectively called the user generated content. Now if one wants to buy a product, it is no longer necessary to ask one's friends and families because there are plentiful of product reviews on the Web which give the opinions of the existing users of the product. For a company, it may no longer need to conduct surveys, to organize focused groups or to employ external consultants in order to find consumer opinions or sentiments about its products and those of its competitors.

II. RELATED WORK

With the rapid expansion of e-commerce, more and more products are sold on the Web, and more and more people are also buying products online. In order to enhance customer satisfaction and shopping experience, it has become a common practice for online merchants to enable their customers to review or to express opinions on the products that they have purchased [5].

With more and more common users becoming comfortable with the Web, an increasing number of people are giving feedback. As a result, the number of feedback

that a product receives grows rapidly. Some popular products can get hundreds of feedback at some large merchant sites [1].

Furthermore, many reviews are long and have only a few sentences containing opinions on the product. This makes it hard for a potential customer to read them to make an informed decision on whether to purchase the product. With such feedback, a potential customer and producer can easily see how the existing customers feel about the particular product [1].

Customers have utilized this piece of this information to support their decision on whether to purchase the product. For product manufacturer perspective, understanding the preferences of customers is highly valuable for product development, marketing and consumer relationship management [4].

The feedback task is interested in features or objects on which customers have opinions. This is different from traditional reviews that involve reducing a larger corpus of multiple answers from customers into a single that conveys the opinion of a particular product. The product feedbacks on the Web are in three formats:

Format 1 - Pros, cons and the detailed review: The reviewers describe pros and cons in the form of short phrases and also write the detail of reviews separately.

Format 2 - Pros and cons: The reviewers describe pros and cons in the form of full sentences separately.

Format 3 - Free format: The reviewers write the reviews in the free form that no separation of pros and cons.

In format 1, pros and cons usually consist of short phrases and incomplete sentences, for example "**pros: fabulous photo quality, large LCD, great battery life, great features**".

The reviews of format 2 and 3 usually consist of long sentences and complete sentences, for example "**I have taken hundreds of photos with it and I continue to be**

amazed by their quality". However, the product features and opinions extraction from reviews of format 2 and 3 is more challenge because the complete sentences are more complex and contain a large amount of irrelevant information. The task of manually scanning through large amounts of review one by one requires a lot of time and cost for both businesses and customers [2]. Therefore, a good summarization system can help them in getting the required and relevant information without going through all the reviews present on the site.

III.   PROPOSED MODEL

A good summarization system can help them in getting the required and relevant information without going through all the reviews present on the site. Instead of collecting feedback as stated above, the feedback is collected from customers by giving multiple choices to each feature. To some extent the manufacturers grab the feedback from the customers what they think about their product that can be given as multiple choices.

A.   Merits :

Opinions are so important that whenever one needs to make a decision, one wants to hear others opinions. This is true for both individuals and organizations.

a. *Individual Customers:* If an individual wants to purchase a product, it is useful to see a summary of opinions of existing users so that he/she can make an informed decision. This is better than reading a large number of reviews to form a mental picture of the strengths and weaknesses of the product. He/she can also compare the summaries of opinions of competing products, which is even more useful.

b. *Organizations and Business:* Opinion mining is equally, if not even more, important to businesses and organizations. For example, it is critical for a product manufacturer to know how customers perceive its products and those of its competitors. This information is not only useful for marketing and product benchmarking but also useful for product design and product developments.

Let us use an example to illustrate a feature-based feedback .Assume that the reviews of a particular laptop_1. The summary looks like the following.

Feature: Picture Quality, Screen Size
Positive: Fabulous, Amazing, Excellent, Very Good
Neutral: Average, Best
Negative: Not Good, Poor, Bad, Worst

In the above, picture quality and screen size are the product features. The customers have given feedback in multiple ways about that product with respect to each feature and feedback can be grouped based on similarity as positive, neutral, negative opinions.
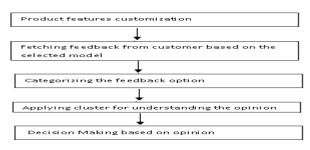
IV.   PROBLEM ARCHITECTURE

Figure 1 Problem Architecture

A.   Description:

a. Applying cluster for understanding the opinion
b. In this approach, first the customer has to register by giving their personal details. A registered customer selects a particular product from various products and views all features of that particular product.
c. The feedback is collected from various customers for the selected model and he/she also fetched feedback based on selected model that can be given by several customers.
d. Then it categorizes the feedback option after getting the feedback from customers based on the combination of choices.
e. After collecting feedback from several customers, the feedback has been mined by applying K-Mean clustering technique.
f. After mining feedback, the feedback has been analyzed based on similarity of choices with highest rank of clusters..
g. Finally, it helps the manufacturer to take right and suitable decision for their improvement of product sales

V.   ALGORITHM AND ANALYSIS

Algorithm: K-Means:

The K-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.
Input
    K: Number of clusters
Here Four Clusters are categorized based on the four choices of all 10 questionnaires. i.e.
Choice A of all questionnaires is grouped cluster 1
Choice B of all questionnaires is grouped cluster 2
Choice C of all questionnaires is grouped cluster 3
Choice D of all questionnaires is grouped cluster 4
So k=4

D: A data set containing n objects:

Here feedback is collected from 20 Customers which indicates 20 objects. The 10 choices of each questionnaire for each customer is grouped together as single object and the same data will be passed as a input for the "k" clusters which in turn each choice is categorized to each clusters and calculates mean value for the each clusters. This step would be repeated for "n" objects.
So, n=20
Output: A set of k clusters
        (A set of Four clusters)
Method
    a.   Arbitrarily choose K objects from D as the initial cluster centers;
    b.   Repeat
    c.   (Re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
    d.   Update the cluster means, i.e., calculate the mean value of the objects for each cluster;
    e.   **Until** no change;

VI.   EMPIRICAL RESULTS AND DISCUSSION

The feedbacks are collected from "n" customers by giving questionnaires to each feature for a particular product. The Questionnaires used for feedback are given below,

A. How would you rate the Performance of the Processor?
   a. Excellent
   b. Good
   c. Moderate
   d. Poor
B. How would you rate Display Quality of the Screen?
   a. Fabulous
   b. For Better
   c. Medium
   d. Below Average
C. What is the life stand of the Battery?
   a. Excellent
   b. Good
   c. Moderate
   d. Poor
D. How would you feel about Pricing Strategy of the Product?
   a. Too Expensive
   b. Quite Expensive
   c. Quite Fair
   d. Very Fair
E. How would you rate the Quality of Services after Sales?
   a. Exceeds than Expectations
   b. Meet with Expectations
   c. Meet lesser than Expectations
   d. Out bound Expectations
F. How do you feel about Data Storage Device?
   a. Exploding Performance
   b. Better Performance
   c. Performance with Satisfaction
   d. Average Performance
G. What kind of Security Measures required in all level while using the Product 'Laptop'?
   a. Sufficient security measures are available
   b. Minimum security measures are available.
   c. Partial security measures are available
   d. Inadequate security measures are available
H. How would you rate the other Accessories of the Product?
   a. Excellent Accessories
   b. Good Accessories
   c. Average Accessories
   d. Poor Accessories
I. How would you rate overall level of Satisfaction with Us?
   a. Highly Satisfied
   b. Some What Satisfied
   c. Neutral
   d. Some What Dissatisfied
J. Will you recommend our Brand to your Colleagues / Friends / Relatives?
   a. Very likely
   b. Some What likely
   c. Neither likely nor unlikely
   d. Some What Unlikely

The above 10 questionnaire are simulated for collecting the feedback from "n" customers which the collected dates' are converted in to text format by using arff file. The screen for data collection and arff file conversion are shown Fig 2 & 3
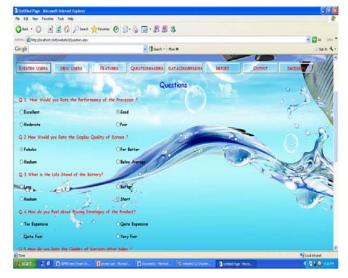

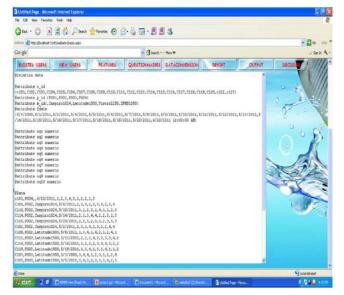Figure. 2 Sample Questionnaire

Figure. 3 Arff File

The data which is in arff file format are given as a input to **Weka tool**. After loading arff file into weka tool and applying clustering technique by using Simple K-means algorithm.

The sample clustering output for given input are shown below
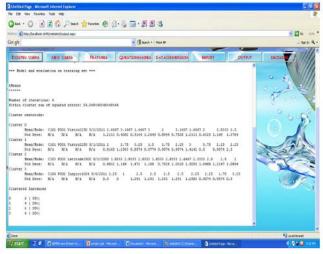

Figure. 4 Clustered Outputs

After obtaining the results of clustering, the feedbacks is analyzed based on the highest value among the clusters and finally gives a decision as,

D1: Highly recommended to continue the same procedure for the product

D2: Slightly recommended to concentrate on minor changes in the procedure of the product

D3: Moderately recommended to concentrate on the slight and major changes on the procedure of the product

D4: Strongly recommended to re modifying the procedure of the product

If cluster **C1** has highest value among the 4 clusters, then decision **D1** would be considered as opinion or If cluster **C2** has highest value among the 4 clusters, then decision **D2** would be considered as opinion or If cluster **C3** has highest value among the 4 clusters, then decision **D3** would be considered as opinion or If cluster **C4** has highest value among the 4 clusters, then decision **D4** would be considered as opinion

The screen below shows the result of the opinion for 20 customers which is given as input and it gives the opinion **D1, since cluster C1 has highest among all 4 clusters.**
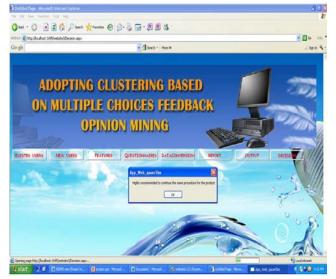

Figure 5 Decision results based on cluster

VII.   CONCLUSION

In this paper, we proposed a set of techniques for mining and summarizing product feedback based on data mining technique. The objective is to make effective decision for product and sales improvement in manufacturing business sector with the help of feedback opinion collected from various customers. The experimental results indicate that the proposed techniques are very promising in performing their tasks. we believe that this problem will become increasingly important as more people are buying and also giving feedback on the web. Summarizing the reviews is not only useful to common shoppers, but also crucial to product manufacturers.

VIII.   FUTURE SCOPE

The above procedure is tested only for same product with different models, it is not tested for different products. In future the same procedure can be enhanced for different products with different models.

IX.   REFERENCES

[1]. Ana-Maria, "Extracting Product Features and Opinion from Reviews," Human Language Technology Conference and Conference on Empirical Methods in Natural Language, Vancouver , Canada, pp. 339-351, 2005.

[2]. Bing Liu," Opinion Observer: Analyzing and Comparing Opinions on the Web", WWW 2005, Chiba, Japan, pp.342-251, 2005.

[3]. Minqing Hu and Bing Liu., Mining and Summarizing Customer Reviews In KDD04, August 22–25, 2004, Seattle, Washington, USA., pp. 168-177.

[4]. Thomas Y. Lee, "Needs-based Analysis of Online Customer Reviews," ICEC'07, Minneapolis, Minnesota, USA, pp. 303-309, 2007.

[5]. Wei Haung, XinChen and Haibo Wang, Product Information Retrieval based on Opinion Mining., 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2010), pp. 2489-2492

Short Biodata of the Author's

J. Sheik Mohamed has obtained MCA Degree from Bharathidasan University in the year 2000, who is currently pursuing Ph.D(PT) Computer Science at S.V.University, Tirupathi, Andhra Pradesh. He has published 8 papers in international and national journals. He is working as an Asst. Professor in SITAMS, Chittoor, A.P. with the experience of 11 years. He is a life member of CSI and ISTE.

P. Nirmala has obtained MCA degree from JNTUA, Anantapur through MCA Department, SITAMS, Chittoor, A.P.

Dr. P. Chitti Babu has obtained MCA & Ph.D(Computer Science) Degree from S.V.University, Tirupathi and M.Tech.(CSE) from Nagarjuna University, Guntur. He is working as a Professor MCA Department, AITS, Rajampet, Andhra Pradesh with the experience of 11 years**.** He has published 6 research papers both in international and national journals of computer science

Dr. S. Rama Krishna has obtained M.Sc., M.Phil., Ph.D., Degrees from S.V.University, Tirupathi and M.Tech (IT) from Punjabi University. He has guided 10 Ph.D scholars in Mathematics and 2 Ph.D. scholars in Computer Science. Currently he is guiding 10 Ph.D. scholars in Computer Science in S.V.University, Tirupathi, Andhra Pradesh. He is working as a Professor & BOS Chairperson in Computer Science Department, S.V.University, Tirupathi, Andhra Pradesh with the experience of 25 years**.** He has published 35 research papers both in international and national journals of mathematics and computer science