Research Journal of Applied Sciences, Engineering and Technology 4(23): 5084-5088, 2012
ISSN: 2040-7467

Submitted: March 31, 2012 Accepted: April 17, 2012 Published: December 01, 2012

A Novel Opinion Mining Technique for Product Review Based on Preferences

G.R. Brindha and B. Santhi
Department of Information Communication and Technology, SASTRA University,
Thanjavur, Tamil Nadu, India

Abstract: As loading of opinions on the Internet increasing, the people have the tendency to seek advice from the Internet to find out other people's opinions about all fields. Web contains valid opinion information such as product, movie descriptive reviews etc. Because of this, the emerging research field such as opinion mining and sentiment analysis is essentially needed. Objective of this study is to provide a decision making tool to guide the consumer for purchasing two wheelers. This study aims to design decision making system based on algorithmic approach. Several methods have been proposed with some limitations. This study includes different parameters for getting accurate decision. The parameters are cost, mileage, style, comfort and performance. The strength of the proposed method is inclusion of weight factor and preference factor. Experimental analysis shows that, this method provides guidelines for both consumer and manufacturer to decide, which are missing in existing system.

Keywords: Decision classification, opinion mining, preference, ranking, weight factor

INTRODUCTION

The rapid development in technologies and web usages enables the consumers to access any type of information throughout the world. Information and/or suggestions about anything appear in different channels like blogs, forums, websites and social networks. People are giving more importance to these information, comments and suggestions, since they can access the content of real experience, which is similar to word-of-mouth publicity. In this fast moving world, everything should be in hand on time and users don't want to read all the overloaded descriptive content. When they go for search engines, they help in finding the reviews of specific destinations which are beyond visual capability of users. Here comes Opinion mining or Sentiment analysis, through which consumers can share their experience as reviews with world-wide. With this consumers can read and use the reviews as references for their decision making to purchase or for some other purposes. But till date the research about sentiment analysis provides the users whether the given product is recommended or not. Some of the researchers provide summarization of positive or negative experiences. Thus the analyses are based on qualifying the information. To make a decision one has to analyze the available ideas and he/she need a toll to give valid guidelines with quantification. This study deals with preference based suggestion, which quantifies the results. The proposed method analyzes the ratings of 243 bike reviews allows the user to enter their taste for the bike features and based on their preference the suggestions are displayed. This study also gives quantified suggestions for the manufacturer for the optimization of their product features to improve sales and brand name.

LITERATURE REVIEW

With the enormous growth of web information, the research on opinion mining is also increasing rapidly. One of the researchers (Zhu, 2008) focuses on opinion extraction and ranking with regression analysis. He created models with different feature set for four Amazon product reviews. Another researcher (Andrea and Fabrizio, 2010) speaks about the lack in traditional opinion mining analysis that is the analysis of large quantities of text must be carried out at the entity or aggregate point. They have compared the accuracy of their weighing method for camera, automobiles and movie reviews from popular websites. Earlier (Abbasi, 2008; Dang *et al*., 2010) sentiment analysis used more refined feature illustration, bag-of-words and n-grams for words. The features also includes part-of-speech, idiom patterns and lexicons which results in challenging part of separating quality from quantity. Apart from the usual feature occurrence measures like frequency and presence, Xue and Zhou (2009) effectively applied distributional measures, such as compactness and initial occurrence in topic-based classification. Mining was done (Silvana *et al*., 2007) based on prioritization of the users skill and

Corresponding Author: G.R. Brindha, Department of Information Communication and Technology, SASTRA University, Thanjavur, Tamil Nadu, India

experience about the product. Another study (Bing, 2010) compares feature based opinion summaries of two cellular phones and shows number of positive opinions and negative opinions.

In this study a corpus with 243 reviews about a bike had been taken as data set. Data may be in any form, ie., raw opinion or normalized data set. Proposed method has taken normalized data of 243 reviews. For raw reviews the following preprocess should be done to get review matrix as input for the proposed method. The review matrix contains 243 rows with 5 columns, each represent cost, mileage, style, comfort and performance of a bike. And the value for each feature is based on the review comment. The opinion is categorized into 3 types, 'Poor', 'Normal' and 'Good' which takes the values 1, 2 and 3, respectively for each feature, which is ranked based on review word extraction. This matrix is given to the proposed algorithm which takes preference from the user and provides suggestions to the user and manufacturer for decision making.

MATERIALS AND METHODS

Usually reviews are collected from web pages and blogs for input data set. The blanks were removed from those sentences and made plain text. Key phrases are identified which describes about bike features. This content becomes input to get building word vector, which contains following process (Christopher *et al*., 2009).

- **Tokenization:** Tokenization is the process of cutting a stream of text into meaning full elements called token such as phrases, words, symbols. For this unigram, bi-gram or n-gram is used.
- **Stop word removal:** The corpus usually filled with articles and prepositions, which are called stop words. Removal of stop word will minimize the corpus which ease further searching and process
- **Collection frequency:** Number of times term appears in the whole corpus.
- **Case folding:** Making upper case letters into lower case to form proper noun
- **Stemming:** Chopping derivational afflictions. For example the word 'costs' becomes 'cost' after stemming. Poter's algorithm is the popular algorithm for stemming.
- **Lemmatization:** Vocabulary and morphological analysis that returns the base dictionary word.

Now the word vector is passed as input to train classifier. Here additional filters like tagger and shallow parser are done. Further, correlation and parameter testing makes the entire matrix fit to be an input for classifier (such as SVM-Support Vector Machine, NB-Naïve Baise).

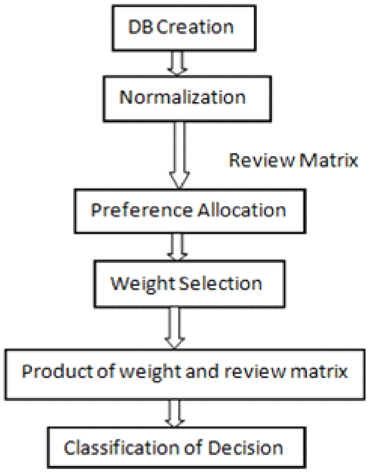

Fig. 1: Process flow of proposed method

Now the process is sentiment lexicon in which scoring, negation detection and sentiment orientation are being done. After this, for five features ranking is posted based on 243 review words. This becomes the input for the proposed algorithm.

Proposed work flow: The ranked review matrix is given as input to the algorithm. Now the system is ready to accept review matrix with 'Poor', 'Normal' and 'Good' rating (1, 2 and 3, respectively) for each 5 feature set (cost, mileage, style, comfort and performance). Hence 3^5 = 243 combinations of reviews are possible. Here proposed method is given with the input of 243 reviews. But there may be redundancy, since any user can have same taste of another one. Also user's taste may differ from one person to other. Some may give preference to style and some to cost. Hence the system takes the preference between 1 and 5 for each feature, listed above. Based on the preference, weight value is assigned. Multiplying the input with weights provides a final input matrix. After this, mean value for the rows are found and based on mean value range is set for 'Poor', 'Normal' and 'Good' rating count. Then these three counts are displayed, which gives the final result. Based on column wise mean value for review matrix, manufacturer gets idea about the product for further enhancement and decision. Figure 1 shows the proposed work flow.

Algorithm:

```
Assign review_matrix as primary input
feature_values = mean (review_matrix)
Get preference [cost, mileage, style, comfort,
    performance]
for 5 columns do
    Arrange the review_matrix according to the preference
end
```

```
Assign weight = [0.5 0.3 0.1 0.075 0.025];
for 243 rows do
    preference_matrix = weight*review_matrix
end
classified_matrix = mean (preference_matrix')
for 243 rows and 5 columns do
    Find the ranks (1, 2, 3) position in classified _matrix
    Assign it to position_matrix
end
set range for poor, normal and good
use the range in position_matrix to get count of poor, normal and good ranks
```

- Consumers get suggestion from the count of position_matrix
- Manufacturers get ideas from feature values

RESULTS AND DISCUSSION

The result analysis includes parametric quantified interpretation based on user's age. Here the parameters for product features are cost, mileage, style, comfort and preference. The preference based reviews are ranked with three parameters, 'poor', 'normal', 'good' with values 1, 2 and 3, respectively. From the following charts and tables, interpretation about users taste and suggestions for the manufacturer to improve their quality and sales are listed.

From a corpus, 243 reviews were analyzed and ranked 'Poor', 'Normal' and 'Good' for the bike performance. Table 1 shows the preferences of 9 users for 5 features and the corresponding review count for 3 categories. The table also contains the mean for those three ranks. Figure 2 shows the comparative chart for 3 categories of 9 preferences.

Case 1: Though the corpus is same, based on preferences the result will vary. And the result finalizes the ranking for the bike performance as 'Good' (Fig. 3).
Figure 4, 5 and 6 shows how same corpus results in different rating based on reviewer's age which in turn depends on the preferences.

Case 2: From Fig. 4, 5 and 6 aged group and youngsters are maintaining consistency in rating, but middle aged group is inconsistent. Reason for this inconsistency is they cannot compromise style and cost factor.

Now Fig. 7 shows the age wise mean difference in taste.

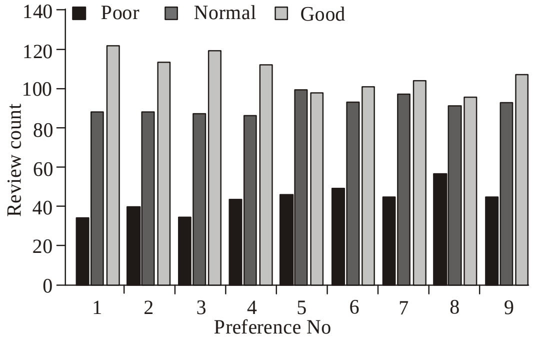

Fig. 2: Preference ratings for 243 reviews by 9 users

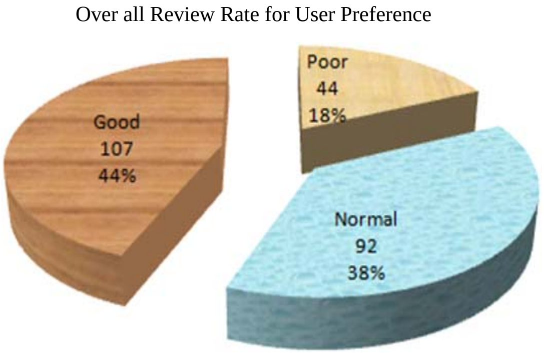

Fig. 3: Average preference rating

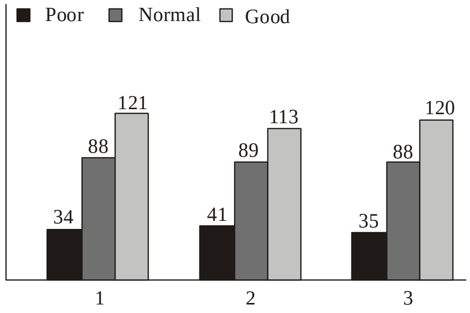

Fig. 4: Old age person's taste

Table 1: Preferences of 9 users for 5 features and their ratings

Preference No.	Cost	Mileage	Style	Comfort	Performance	Poor	Normal	Good
1	1	2	5	4	3	34	88	121
2	1	3	5	2	4	41	89	113
3	1	2	5	3	4	35	88	120
5	3	4	5	2	1	44	87	112
6	5	4	2	1	3	46	100	97
7	4	5	3	2	1	49	93	101
8	5	4	1	3	2	44	97	102
9	4	5	1	2	3	56	92	95
10	3	2	1	4	5	44	93	106
					Mean	44	92	107

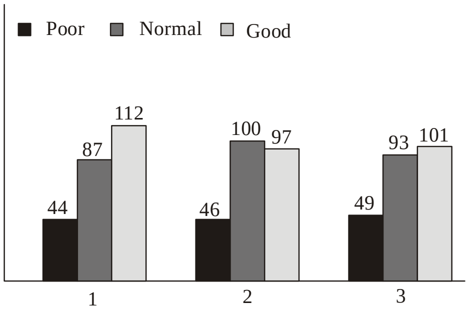

Fig. 5: Middle aged person's taste

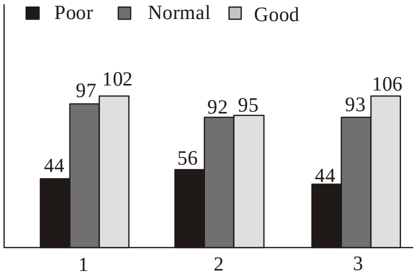

Fig. 6: Youngster's taste

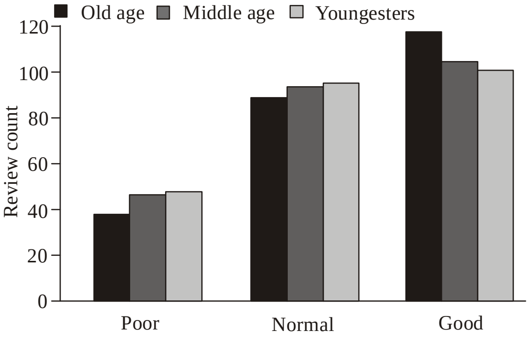

Fig. 7: Over-all age-wise preference comparison

Case 3: Aged people marked 'Poor' with less rating and 'Good' with more rating compared to others, therefore old aged people like this bike more compared to other age groups.

Case 4: For Youngster's 'Normal' and 'Good' ratings are equal, thus we can conclude, that for them bike is convincing to some extent. This interpretation suits also for middle aged people

Case 5: As a whole the bike is rated 'Good' by all age group people. Table 2 and Fig. 8, reveals the reason behind this interpretation

Table 2 contains the mean values for five feature set of 243 reviews and Fig. 7 gives the comparative chart for the table, which depicts some important points. It gives valid suggestion to the manufacturer for further enhancement.

Case 6: The bike is known for its mileage. Cost and Performance is good; they have to increase comfort satisfactorily and style needs much improvement.

In connection to this interpretation the chart tells the reason for the differences in Fig. 7.

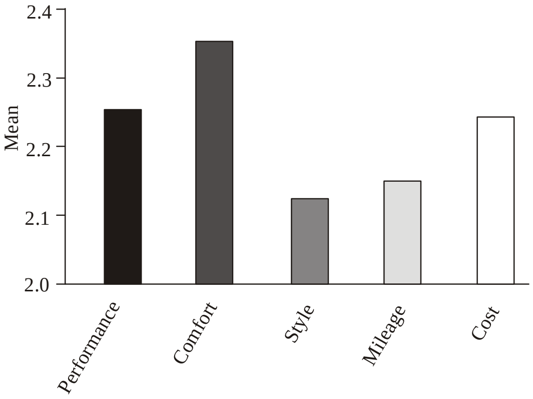

Fig. 8: Feature set comparison for 243 reviews

Table 2: Mean for 243 reviews on 5 features

Corpus size: 243	Cost	Mileage	Style	Comfort	Performance
Mean	2.2428	2.3457	2.1235	2.144	2.2346

Case 7: In general mileage is the prime factor for old age group. Cost and performance are significant factors. This study reveals, mileage gets top position, which justifies overall decision assigned with less 'poor' rating.

According to youngsters style is the prime factor. This study brings style in least position. Hence youngsters cannot give good rating. This point reveals 'good' rating has less difference.

Case 8: Sine style gets less rating, the manufacturer should improve style which is preferred by youngsters and middle aged people.

From this the interpretation is style and comfort should be taken care to attract both youngsters and middle aged people.

Case 9: Since the bike is good for its mileage, cost and performance, the 'Normal' rating is more or less equal by all age groups and 'Good' rating is equal by middle aged and youngsters.

Case 10: Though style is comparatively less, mileage, cost and performance leads the final result as 'Good'.

Thus the manufacturer should optimize these five features based on the result and taste of the reviewers, by which the product will get good ranking and they can improve the sales.

CONCLUSION AND FUTURE WORK

The study explains a method to get a suggestion from the review set, based on user preference. Existing methods concentrate on recommendations with summary, but not on preferences. The proposed method is a novel one which adds weight according to the preference and based on their taste they receives guidelines. Further it considered preferences and ranking the features then provides valid suggestions to the consumer. Further this research can also be done as a model with statistical regression, in which user can enter their review and get suggestions.

REFERENCES

Abbasi, 2008. Affect analysis of web forums and blogs using correlation ensembles. IEEE T. Knowl. Data Eng., 20(9): 1168-1180.

Andrea, E. and S. Fabrizio, 2010. AI and opinion mining, Part 2, trends and controversies. IEEE Intell. Syst., 10: 1541-1672.

Bing, L., 2010. Sentiment analysis: A multifaceted problem. IEEE Intell. Syst., pp: 76-80.

Christopher, D.M., R. Prabhakar and S. Hinrich, 2009. An Introduction to Information Retrieval. Cambridge University Press, Cambridge, England.

Dang, Y., Y. Zhang and H. Chen, 2010. A lexicon-enhanced method for sentiment classification: An experiment on online product reviews. IEEE Intell. Syst., 25(4): 46-53.

Silvana, A., Z. Debbie, S. Simeon and D. John, 2007. Informed recommender: Basing recommendations on consumer product reviews. IEEE Recomm. Syst., 07: 1541-1672.

Xue, X.B. and Z.H. Zhou, 2009. Distributional features for text categorization. IEEE Trans. Knowledge Data Eng., 21(3): 428-444.

Zhu, Z., 2008. Weighing stars: Aggregating online product reviews for intelligent E-commerce applications. IEEE Intell. Syst. NLP WEB, 08: 1541-1672.