IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

The Analysis and prediction of customer Review Rating and Opinion Mining.

[1]Dr. Saravanakumar,[2] Akshay Manoj, [3]Akshita Sharma,[4] Aroop Das, [5]Frank E M Dannie
[1](Associate Professor, Dayananda Sagar Academy of Technolgy and Management,Bangalore)
[2][3][4][5](Dayananda Sagar Academy of Technolgy and Management,Bangalore)

Abstract— The customer review is important to improve service for company, which have both close opinion and open opinion. The open opinion means the comment as text which shows emotion and comment directly from customer. However, the company has many contents or group to evaluation themselves by rating and total rating for a type of services which there are many customer who needs to review. The problem is some customers given rating contrast with their comments. The other reviewers must read many comments and comprehensive the comments that are different from the rating. Therefore, this paper proposes the analysis and prediction rating from customer reviews who commented as open opinion using probability's classifier model. The classifier models are used case study of customer review's hotel in open comments for training data to classify comments as positive or negative called opinion mining. In addition, this classifier model has calculated probability that shows value of trend to give the rating using naive bayes techniques, which gives correctly classifier to 94.37% compared with decision tree Techniques.

Keywords—open opinion; customer review; opinion mining; naive bayes; decision tree

I. INTRODUCTION

Nowadays, a company or organization provide a business service which needs to get feedback from customer. With the rapid expansion of company or organization have more services and products online and enhance customer satisfaction. The provider will read customer review and other customers who need to use services or products will read review to express opinions on the services. The number of customer review is increasing or huge from website, blogs, forums and social media, which the services or product is interesting. Therefore, many customers will read comment randomly which is hard to read all comments and make decision the services or products. If customer reads a few reviews, customer might get opinion review to be bias. Therefore, opinion mining is a technique of field area of information extraction from text processing, which is benefit and many opportunities to improve or develop factor to business work by this analysis. The problem is the comments from customer review about products or services, which are contrast with comments. For example, the customer

978-1-5090-5756-6/17/$31 .00 02017 IEEE SERA 2017, June 7-9, 2017, London, UK commented "Even if I stay alone, it was safe, clean and suite for seminar". However, the customer given rating only 5.7 which others customer expected more rating. Or the customer commented "Old room, dirty, and water flow in toilet was very slow. Breakfast was not delicious, services was not good", customer given value rating 4.3, it seems to bias their comments. Therefore, the opinion mining is computing value automatically which can be trended to their opinion to judge the comments negative and positive.

Many researches in sentiment analysis and opinion mining have been many languages, for example, Chinese [1],[2],[3], Arabic[4],[5], Vietnamese[6], and Thai[7]. These researches focused into 2 ways: analysis of sentence has level of sentiment from emotion word and calculated score of similarity or cluster with the kind of word as positive or negative called sentiment polarity [7],[8],[9],[10]. Secondly, the papers 12],[13]proposed and survey the classifier model to summarize sentence as positive and negative and try to apply in other case studies [14],[15]. However, the summarized from opinion sentences are unable to show the continuous value trending to negative and positive. For example, sentiment analysis summarizes as positive, however, the customer still needs to know rating the overall value of positive opinion indicating a number. Therefore, our approach will be proposed methodology in this paper that can be generating from probability of classifier models.

This paper is organized the following: the related work will be shown in section Il. Section Ill describes the proposed methodology how to calculate rating form customer review automatically. Section IV shows experimental results compared between two models, moreover, discussion for each models will be show in this section. Finally, the conclusion is explained in section V.

Il. RELATED WORKS

The opinion mining has become one of popular research area. The challenge is in process of opinion mining or sentiment analysis that is unstructured and noisy data on website. A part of opinion mining refers using of natural language processing (NLP) by proposed different method of dictionary for sentiment analysis of text as corpus, lexicon and specific language dictionary [4], [7], [8], [16]. They tried to extract word from sentences for removal stop word or IEEE

Ocomputer society

unnecessary word automatically. In addition, various dictionaries are solved by machine learning methods [12], [13], which try to rank scoring of various dictionaries. For example, the paper in [13] used fuzzy logic algorithm to collect the ranking of different dictionary into rule for classify the opinion.

After word segmentation process is removal stop words by dictionary checking. The group of researches in [l], [2], [6], [9], [17] focuses on the calculating polarity of words to trend in positive or negative in a cluster of interest's customer that are 71

extracted from texts and compared the word occurrence of whole sentence. If the word extractions have weight from dictionary of emotional words, it is calculated to answer the comment as positive or negative.

However, the customer review has different behavior with the product. The proposed classifier model is presented using association rule in [11]. The popular classifier model is naive bayes compared with other model [5], [8], [10], which there are different sources such as social media and web site. From these researches are used classifier models that are the same objective to classified opinion. Our approach is different from them, this paper use the advantage of classifier model to generate the rating value from classifier which is not only shown classify opinion as positive and negative and also factors analysis to impact the customer who posted or commented to positive and negative.

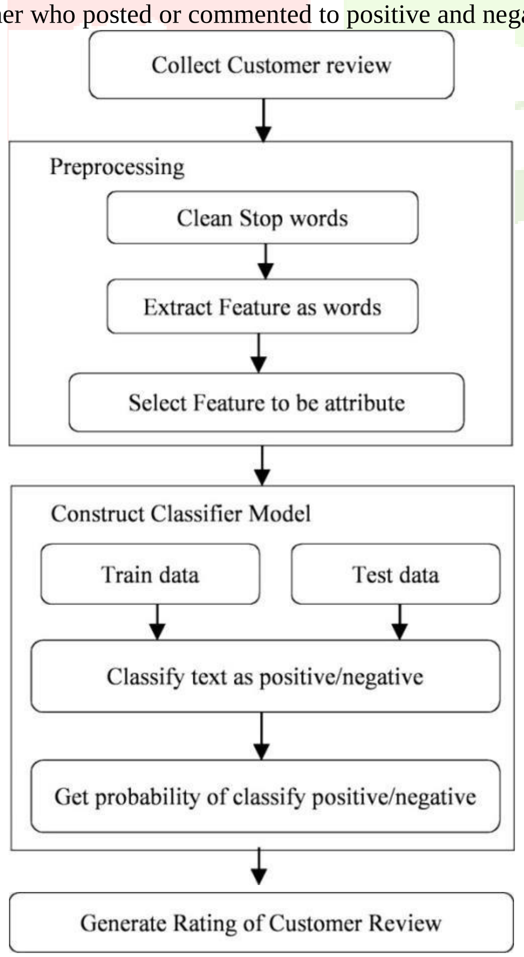


Fig. l. Proposed Methodology for generating score of customer review using opinion mining

Ill. PROPOSED METHODOLOGY

The proposed methodology used Thai customer review's hotels from a website of hotel agent service, which service in hotel reservation directly. The target of classify customer review from this website because the comment is posted from customer who is serviced checked-in and checked-out from hotel. The system has cleaned the promotion of hotel's comment which has only existed customer review given comment and rating. The numbers of open opinion texts are collected 400 customer reviews that are used service to checked-in/out the hotels in Bangkok, Thailand. The process is started from collected data and preprocessing is cleaned data by removal stop words and using the high frequency of word which will be selected into attribute for using classifier model. The classifier model will be solve the text of customer review that is positive of negative from training data and test data which are train from behavior posting from customer of hotel service group. The proposed methodology are detailed as follows,

A. Preprocessing

The feature selection is to be attributes in classifier that will be extracted words from these customer reviews as words occurred frequently to 36 words. There are positive and negative in Table I, which are ordered by descending frequent.

TABLE 1. FEATUE SELECTION FROM FREQUENT WORDS

No.	Words (Positive)	#Frequent	Words (Negative)	#Frequent
1.	Convenient	245	Small/Narrow	44
2.	Good	206	Little/Few	43
3.	Near	142	Old	32
4.	Clean	140	Not delicious	25
5.	Comfortable	62	Not Care/not impression	19
6.	Very Good	59	Dirty	14
7.	Take care	33	Far	13
8.	New	32	Not Smile	12
9.	Smile	29	Uncomfortable	11
10.	Big/Wide	26	Dark	11
11.	Delicious	25	Crowded	9
12.	Cheap/not expensive	19	Inconvenient	8
13.	Much/Many	17	Slow	8
14.	Safe	17	Expensive	8
15.	Quiet	16	Bad/Not good	7
16.	Worth	12	Not beautiful	6
17.	Beautiful/Luxurious	9	Not worth	5
18.	Fast/Quick	6	Improve	4

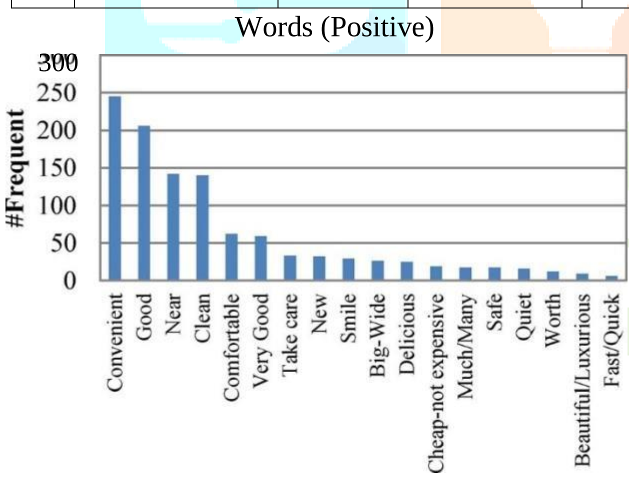


Fig. 2. Frequent word of positive opioins

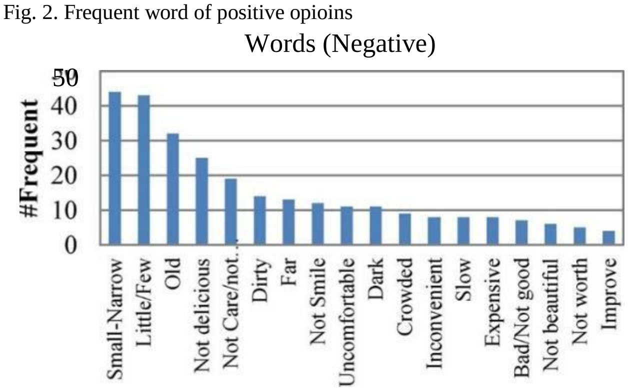


Fig. 3. Frequent word of negative opinions

The frequent words of positive are analyzed for attribute transformation individual text of customer review. The training and test data are separated into 3 sets: set I is composed 5 positive and 5 negative words; set is 2 is composed of 10 positive and 10 negative words and set 3 is composed of all positive and negative words in Table Il as follows,

TABLE 11. DATA SET S FOR CLASSIFIER MODELS

Data sets	Words
Setl (10 words)	Positive: convenient, good, near, clean, comfortable Negative: small/narrow, little/few, old, not delicious, not care/not impression
Set2(20 words)	Positive: convenient, good, near, clean, comfortable, very good, take care, new, smile, big/wide Negative: small/narrow, little/few, old, not delicious, not care/not impression, dirty, far, not smile, uncomfortable, dark
Data sets	Words
Set3(36 words)	Positive: convenient, good, near, clean, comfortable, very good, take care, new, smile, big/wide, delicious, cheap/not expensive, much/many, safe, quiet, worth, beautiful/luxurious, fast/quick Negative: small/narrow, little/few, old, not delicious, not care/not impression, dirty, far, not smile, uncomfortable, dark, crowded, inconvenient, slow, expensive, bad,/not good, not beautiful, not worth, improve

B. Model Construction

From data sets lead to model construction. The classifier models are used 2 models which are decision Tree (C4.5) and naive Bayes to classify texts as class labels: positive or negative. Each data set is trained to model and test model that given predicted class labels follows probability trending of classifier model. The classifier models are described as bellows,

• Decision Tree(C4.5)

The decision tree learning was proposed as a model of data classification for a class label, which called ID3 and developed to C4.5. In addition, decision tree is clearly represented through a tree diagram. It starts from the first node is a root node. The root node selects an attribute as words in opinion from the best value of measurement. Each attribute has its own values i.e. true/false, which are separated by branch links composed of original attributes. At the end, the data reveals a class which represents a leaf node (i.e. positive/negative).

The advantage of the decision tree is for ordering attributes that are the best measurement as Eq.(l).

$$I(s_1, s_2, \ldots, s_n) = -\sum_{i=1}^{n} \frac{s_i}{s} \log_2 \frac{s_i}{s} \quad (1)$$

where,

n is the number of class label.

S is the number of data Sl of class i.

After the distinguished information of attribute is calculated, the entropy value is also calculated to define the summary of each branch needed be clearly separated from attribute A as Eq.(2)

$$E(A) = \sum_{j=1}^{m} \frac{s_{1j} + \ldots + s_{nj}}{s} I(SIJ \quad S2J) \qquad (2)$$

where, m is the number of branch of attribute A.

The highest gained value of the attribute A results in the difference the number of feature are extracted as 10, 20 and 36 best attribute to classify data set which is calculated and range words respectively. The accuracy of naive Bayes is given between 0 and I by Eq.(3) values that are higher than decision tree all of data sets. Moreover, the highest of accuracy value is 94.37% with 20 words and also average of naive Bayes is higher than decision

$$\text{Gain }(A) - I(s\ s\ _n) - E(A) \qquad (3)$$

tree to 93.61% in Table 111.

TABLE 111.

ACCURACY OF CLASSIFIER MODELS

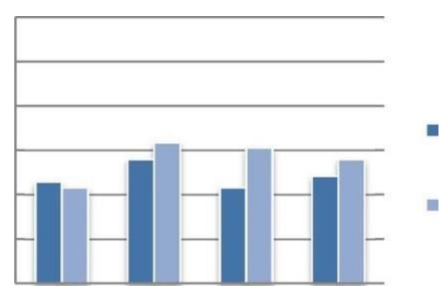
Attributes	Correctly Classifier (%Accuracy)	
	Decision Tree(C4.5)	Naive Bayes
10 words	92.58	92.33
20 words	93.61	94.37
36 words	92.33	94.12
Average	92.84	93.61

- Naive bayes

Naive bayes is an algorithm of probability based on Bayes theorem of learning. It aims to create a model in the form of probability. The advantage of naive bayes is an effective method which is easy processing. The probability of the classification data with prior knowledge is denoted by P(ail y), where ai refers to the attribute i and vj refers to class label j. Therefore, the classification has been calculated for this probability. The highest

each class is trend to answer of classification. probability of ai The range of probability is between 0 and I as is depended on y for Eq.(4).

$$V_{NB} = \arg\max P(V_J) \prod_{i=1}^{n} P(a_i \mid V_J) \qquad (4)$$



C. Evaluation Model

The evaluation model is used k-fold cross validation with test data which are generated all training data. The k defines the number of grouping data. For example, k is 10-fold cross validation of 400 training data, means each group as 40 records and 10 groups, whereas the testing data will be groups I of 40 records and evaluation this groups to calculate average of the accuracy collected until N as 10 groups,

$$\text{Accuracy} = \frac{\sum_{j=1}^{N} \delta_{ij}}{N} \qquad (5)$$

where,

1 = predicted class label is correct

0 = predicted class label is incorrect

In addition, the results is evaluation by rating, the root mean square error is used in this case. The comparison results are generated rating with classifier model and rating from actual customer review as Eq.(6).

related and priority following the entropy value. For example of decision tree with 10

Rule2: IF not care — true and near — true THEN Positive

where, Rule3: IF not care — true and near = false and convenient = true Pi is prediction from probability value of classifier model. THEN Positive

Oi is actual score from customer review. Rule4: IF not care = true and near — false and convenient false and good = true THEN Positive

IV. EXPERIMENTAL RESULTS Rule5: IF not care = true and near = false and convenient=

The experimental results are tested with open opinion texts false and good — false THEN Negative from 400 customer reviews from a website of hotel agent service. The results are compared percentage of accuracy between decision tree model (C4.5) and naive Bayes [16] and

$$RMSE (6) = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(P_i - O_i)^2}$$

10 20 36 Average words words words

Fig. 4. Comparision of decision tree and naive Bayes

However, the advantage of decision tree is model shown structure of words has

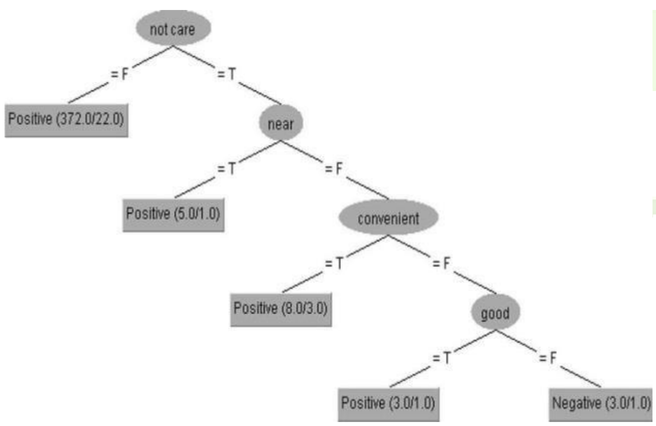


Fig. 5. Decision tree from training data (10 words)

The decision tree with 20 word training data shown the first service is dirty and lower level are far which is related to clean and good, moreover, not care word is related to smile and far again. These keywords are translated into rules, for example, Rule3: If customer complains far but have other good convenient, customer still gives positive score. And Rule4: If customer complain far but do not have any good convenient, customer still give negative score. Moreover, Rule 5-8: dirty room is first factor to decide of negative score. All relationship of word has IF-THEN rules as follows,

Rulel : IF dirty — false and far — false THEN Positive

Rule2: IF dirty — false and far — true and clean — true THEN

words training data, the hotels should take care of customer, location is near tourist attraction, convenient in room are ready, therefore, the review is trend to good and positive. The words relationship is able to translate to IF-THEN rules as follows, IF not care = false THEN Positive

Rulel:

Positive

Rule3: IF dirty — false and far _ true and clean — false and good — true THEN Positive

Rule4: IF dirty — false and far _ true and clean — false and good= false THEN Negative

Rule5: IF dirty = true and not care — true THEN Negative

Rule6: IF dirty = true and not care — false and smile — true then Negative

Rule7: IF dirty = true and not care = false and smile = false and far = true then Negative

Rule8: IF dirty = true and not care = false and smile = false and far — false and uncomfortable = true then Negative

Rule9: IF dirty = true and not care = false and smile = false and far = false and uncomfortable = false then Positive.

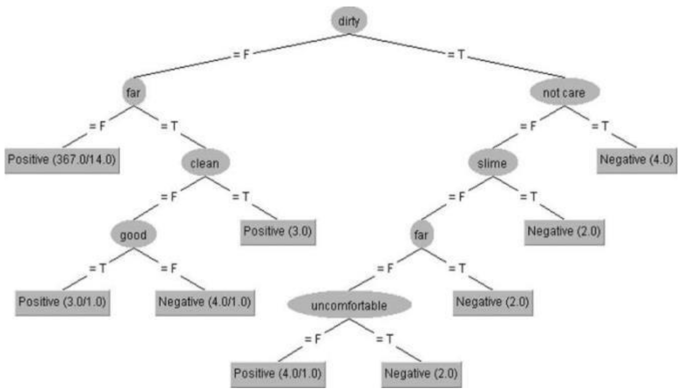


Fig. 6. Decision tree from training data (20 words)

The decision tree with all word training data shows word relationships such as dirty, far, not care, not good, many/much, smile, expensive, good, far, near, uncomfortable in form of tree. In this experimental results show that some words has effected to class label. For example, Rule 5, even if customer review in text as expensive and near, customer still has opinion as positive, whereas, Rule6 has expensive and good word in customer review, customer given negative rating to service. All relationship of word has IF-THEN rules as follows,

Rule l : IF dirty false and far true and many/much — false THEN Positive

Rule2: IF dirty — false and far true and many/much true THEN Negative

Rule3: IF dirty — false and far — false and not good — true and good — true THEN Positive

Rule3: IF dirty — false and far — false and not good — true and good - false THEN Negative

Rule4: IF dirty — false and far — false and not good = false and expensive — false THEN Positive

Rule5: IF dirty — false and far — false and not good = false and expensive = true and near = true THEN Positive

Rule6: IF dirty — false and far — false and not good = false and expensive = true and near — false and good = true THEN Negative

Rule7: IF dirty — false and far — false and not good — false and expensive = true and near — false and good — false THEN Positive

Rule8: IF dirty = true and not care = true THEN Negative

Rule9: IF dirty = true and not care = false and smile = true THEN Negative

Rule10: IF dirty = true and not care false and smile — false and far = true THEN Negative

Rulel I : IF dirty = true and not care false and smile__ false and far __ false and uncomfortable = true THEN Negative

Rule12: IF dirty = true and not care = false and smile __ false and far = false and uncomfortable = false THEN Positive

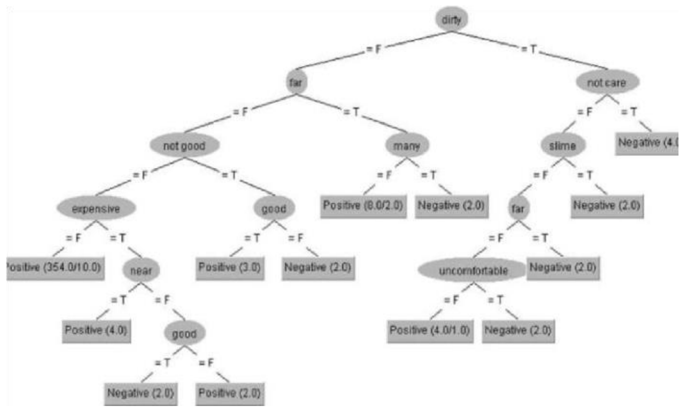


Fig. 7. Decision tree from training data (36 words)

However, the rating generating is testing by naive bayes by probability trend to predict class label in Table IV. The table IV shows RMSE of different data sets. The lowest of RMSE is 36 words testing data that give rating that are similar to actual score from customer review to 0.2326. The rating of 20 words and 10 words are slightly higher value than 30 words to 0.2390 and 0.3669 respectively. The average of naive bayes model generates rating value that is similar actual rating as 0.2792 and median as 0.2390.

TABLE IV. ROOT MEAN SQUARE ERROR OF NAIVE BAYES

Attributes	Root Mean Square Error (RMSE)
10 words	0.3660
20 words	0.2390
36 words	0.2326
Average	0.2792

The Table V. show testing data using naive bayes generating. For example, the comment no. l, the customer posted the comment good words and trend to positive the predicted of naive bayes is the similar as 8.5 from 8.3. and naive bayes classifier give opinion positive. The system is better in commented no.3. The customer commented the hotel is safe and clean which the other customer read will be make decision the comments as positive, but comment is given point as 5.7. The same as comment no.5, the customer posted trend to negative but rating is natural. Therefore, our approach will be generated the rating value in consistency with opinion with their decision automatically.

TABLE V. EXAMPLE OF TESTING DATA TO GENERATE RATING FROM NAIVE BAYES MODEL

No.	Customer Review's comment	Rating Value		Opinion
		Manual Rating	Predicted Rating by Probability's Naive Bayes x10	
1.	Toilet was clean, comfortable bed, it's near downtown. Staff smile, take care of customer but staff is a few to stand by services, not many car parking. However, it is worth com ared with rice.	8.3	8.5	Positive
2	The location is near business center but narrow road. Laxury room, good service but bedroom is too small.	8	8.5	Positive
3	Even if I stay alone, it is safe, clean and suite for seminar	5.7	8.5	Positive
4	Old room, dirty, and pie in toilet was slow. Breakfast was not delicious and service was not good.	4.3	2.2	Negative
5	I read from the Internet and booked it. I feel disappoint, different from imagine, bad smell, di and darkness.	5.3	1.6	Negative

V. CONCLUSION

The opinion mining of customer review is very important to improve service, which the model is compared between decision Tree and naive Bayes. The advantage of the classification model is calculated from probability that is trended to predicted class label. However, the advantage of the decision tree shown the factors ordered by level of tree to help analyzing service improvement and priority factors. In additional, naive Bayes model is able to use probability which is similar value rating, which the system is computing automatically. Even customer will be read comments, but the system can be summarized whole rating consistency with the comments. Therefore, the customers can make decision rapidly. In the future work, we focus on preprocessing data automated extract words from a sentence using machine learning method in order to solve different sentiment polarity.

REFERENCES

sentiment, Proceeding of the 2nd International Conference on Big Data Computing Service and Application, 2016, pp. 52-57.

[1] S. J. Wu, R.D. Chiang and Z.H. Ji, Development of a Chinese opinion- mining system for application to Internet online forum, The Journal of [11] N. Kumari and S. N. Singh, Sentiment analysis on E-commerce

Supercomputing, Springer US[Online], 2016. application by using opinion mining, Proceeding of the 6th International

Z. Li, L.Liu and C.Li, Analysis of customer satisfaction from chinese Conference-Cloud System and Big Data Engineering(Confluence), 2016,

[2] pp. 320-325.

reviews using opinion mining, Proceeding of the 6th IEEE International

[12] V.B. Raut and D D. Londhe, "Survey on opinion mining and

Conference on Software Engineering and Service Science(ICSESS). summarization of user review on web", International Journal of

2015, pp.95-99. Computer Science and Information Technology, Vol.5(2), 2014, pp.

[3] Q.su, X.Xu, H.Guo, Z.Guo, X. wu, X. Zhang and B.Swen. Hidden Sentiment 1026-1030.

association in Chinese web opinion mining. Proceeding of the 17th International Conference on World Wide Web, 2008, pp.959-968. [13] J. Fiaidhi, O. Mohammed, S. Mohammed, S. Fong, and T.H, Kim,

Opinion Mining over twiiterspace: Classifying tweets programmatically

[4] S.Atia and K. Shaalan, Increasing the accuracy of

opinion mining in using the R approach.

Proceeding of the 7th International Conference on

Arabic. Proceeding of the 1 st International

conference on Arabic Digital

Information Management, 2012, pp. 313-319.

computing linguistics, 2015, pp. 106-113.

[14] M. R. Islam and Minhaz F. Zibran, "Exploration and Exploitation of

[5] R.M. Duwairi and I. Qarqaz, Arabic Sentiment Analysis using Developers' Sentimental Variations in Software Engineering",

Supervised Classification. Proceeding of 2014 International Conference Internation Journal of Software Innovation,

Vol.4(4), 2016, pp.35-55. on Future Internet of Things and Cloud. 2014, pp. 579-583. Y. Yokoyama, T. Hochin and H. Nomiya,

"Estimation of Factor Scores

[15]

[6] H.S. Le, T. V. Le and T. V. Pham, Aspect Analysis for Opinion Mining of Impressions of Question and Answer

Statements", Internation Journal of Vietnamese Text. Proceeding of Internation Conference on Advance of Software

Innovation, Vol. I (4), 2013, pp.53-66. Computing and Application, 2015, pp. 1 18-123.

[16] L. Lin, J. Li, R. Zhang, W. Yu and C. Sun,

Opinion mining and [7] T. Chumwatana, Using sentiment analysis technique for analyzing Thai sentiment analysis in social

networks: A retweeting structure-aware customer satisfaction from social media. Proceeding of the 5th approach. Proceeding

of the 7th International Confernece on Utility and International Conference on Computing and Informatics, 2015, pp.659-

Cloud Computing, 2014, pp.890-895.

664.

[17] A.H. Al-hamaami and S. H. Shahrour,

Development of an opinion blog [8] S.Ahmed and A.Danti, A novel Approach for sentimental analysis and mining system,

Proceeding of the 4th International Conference on opinion mining based on sentiwordnet using web data. Proceeding of

Advanced Computer Science Application and Technologies, 2015, pp.

International Conference on Trends in Automation, Communications and 74-79.

Computing Technology, 2015, pp. 1-5.

[9] R.K. Bakshi, N. Kaur, R. Kaur and G.Kaur, Opinion mining and sentiment analaysis, Proceeding of the 3 rd International Conference on Computing for Sustainable Global Development, 2016, pp. 452-455.

[10] P.Barnaghim, J.G. Breslin and P. Ghaffari, Opinion mining and sentiment polarity on Twiiter and correlation between events and