ISSN:0975-9646

Keshav R, et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 4782-4787

Content based Recommender System on Customer Reviews using Sentiment Classification Algorithms

[1]Keshav R,[2]Arvind Swaminathan S,[3]Goutham R,[4]Naren J

[1,2,3] IV B.Tech CSE, School of Computing,
SASTRA UNIVERSITY, India,
[4]Assistant Professor CSE, School of Computing,
SASTRA UNIVERSITY, India

Abstract—the paper proposes *RecoProd* - a recommender system which uses sentiment analysis techniques to provide the best products for the customers. The system uses the existing product reviews upon which sentiment classification is carried out. *RecoProd* consists of an Information Retrieval component which extracts the reviews from the e-commerce websites using the product names as queries. Sentiment Analysis algorithms like Naive Bayes and SVM are used to categorize the reviews and opinion scores are assigned to the reviews. A comparative study on the accuracy of the sentiment analysis algorithms used is also carried out. Aspect based summary of opinions for each product is carried out and visually compared. The products are then clustered and the optimal product along with the recommended products is displayed to the user.

Keywords: Sentiment Analysis, Opinion Mining, Information Retrieval, Recommender Systems

I. INTRODUCTION

In recent years, the importance of e-commerce has expanded exponentially. The opinions expressed in the Web like reviews, forum posts, social media etc., have a great effect on the buying choice of the customers. The paper utilizes the potential of the product reviews to build a recommender system for customers. Normally, a customer when faced with the task of buying a product say a mobile phone, narrows his choices to three products, then goes through all the customer reviews of these products before coming to a conclusion about which product to purchase. The recommender system helps in simplifying this task for a customer by providing the most optimal product based on the customer reviews.

Sentiment Analysis, an area of Natural Language Processing (NLP), is used to classify the reviews using the sentiment of the words into positive or negative. Using the sentiment expressed in the words, opinions

On any entity can be categorized into positive or negative. For example, the sentence, 'I am not excited by this product though it is quite cheap' expresses a negative sentiment about the product. The degree of the sentiment used is also taken into consideration. For example, 'I love this product' indicates a more positive sentiment than the sentence 'I like this product'. Apart from regular adjectives like 'good', 'bad' and 'very good', conjunctions like 'but', 'although", 'while' also have a say in the overall polarity of the sentence. Some sentences need not offer any particular opinion. For example, the sentence 'I bought this phone two days back' is a subjective opinion on the product.

The paper focuses on Document level sentiment classification, one of the key research topics of Sentiment Analysis. Document level sentiment classification as explained in [1] deals with the task of determining whether a given document expresses a positive or negative opinion on an object. Both unsupervised and supervised methods of classification exist. Unsupervised algorithms like the Pointwise Mutual Information and Latent Dirichlet allocation have been proposed and the results have been discussed in [8] and [9] respectively. Many supervised algorithms exist like the Naive Bayes Classifier and Support Vector Machines. The paper attempts to train the classifier models using these algorithms and evaluate their performance. Another important topic of Sentiment Analysis which has been discussed in this paper is the task of Opinion Summarisation. The reviews are analyzed for aspect based opinions and have been summarized. A visual comparison of the summarizations has also been executed and the results have been discussed. For the task of recommending the products to the user, the products are clustered based on the assigned polarity scores and the products belonging to the cluster of the most optimal product are recommended.

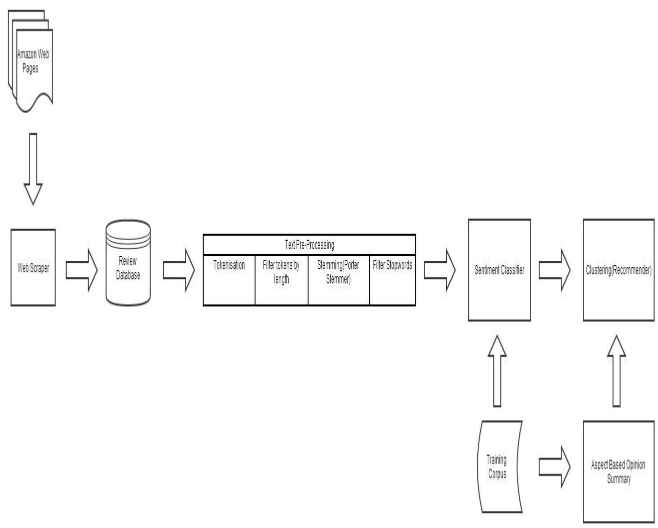

Fig. 1: System Architecture of *RecoProd*

II. LITERATURE REVIEW

In the literature survey a bird's eye view on the techniques involved in Sentiment Classification, Aspect Based Opinion Summarization are studied. The paper on Feature Extraction by Liceste [4] proposes a new methodology that retrieves aspects from a collection of customer reviews and also some salient sentiments about products.

Lie Zhang, Bing Liu et.al in [3]; speak about the methodology of ranking feature candidates by two factors viz. Feature relevance and feature frequency. Yu Zhang and Weixiang Zhu [12] focus on extracting implicit features of the products based on Customer Reviews. In the paper of Santhoshi Moringa and Kenji [13], a framework for mining the reputations of Products in the Internet is proposed. Larissa and Renata in [15] have established an Ontology Based Feature Level Opinion Mining where an evaluation is done for identifying polarity in Portuguese user generated reviews based on the description of features in Domain Ontology.

Many papers on Opinion Retrieval including that of Shuang Liu in [16] has the main aim as Document retrieval that is done by utilizing Word Net and Recognizing Phrases. Another paper by Clement Yu and Weiyi Meng [17] focus more on Opinion Polarity Classification for Opinion Retrieval. Generative Models for Opinion Retrieval is proposed by Chirag Shah Et.al in [18]. Bing Yang Li, [19] explains about an effective approach for Sentence Based Opinion Retrieval as most of the opinions in various social networking websites are sentences. There are also papers on the Statistical Approach for reviews Bo Pang and Liang Lee, [20].

Feng Wang and Li Chen,[21] say the traditional recommender techniques assume that the users would have prior experience on handling e-commerce products and hence proposed a product review similarity network. A Latent Class Regression Model is built to identify sub-communities in the Network. The Literature Survey also has couple of papers on Sentiment Classification where Julian Brooke in [22] proposes a Semantic Orientation Calculator for lexicon Based Sentiment Analysis. A paper on mining and summarizing reviews is elucidated by Bing lie in [1].

III. OVERVIEW OF *RECOPROD*

The system architecture which has been depicted in Fig. 1 shows the detailed structure of *RecoProd* , which consists of the following components:

- Information Retrieval
- Sentiment Classification
- Aspect based Opinion Summarization
- Recommender System

A. Information Retrieval and Corpus Creation

1) Information Retrieval: The goal of the Information Retrieval Component is to extract the reviews from the product websites. Three product names are taken as input queries and mine the respective products reviews from Amazon. The product URL is determined using the product name and ASIN (Amazon Standard Identification Number), which is determined from the input query. The system then runs a HTML parser to find the required tag using which the reviews are retrieved. Once all the product reviews are extracted they are stored in the local database.

2) Corpus Creation: The training corpus was prepared from the Amazon review dataset provided by J. McAuley and J. Leskovec [14]. The dataset consisted of all cell phone reviews, spanning a period of 18 years up to March 2013.

The training data was prepared by extracting the review text and review score from each review. The reviews were then categorized into positive and negative reviews. It was assumed the reviews with ratings 1 and 2 were assumed negative.

B. Sentiment Classification

The training data was in the form of two directories named positive and the other negative, each containing the respective reviews. The task of sentiment classification was done by the machine learning software RapidMiner [25].RapidMiner provides a platform for machine learning, data mining and predictive analytics.

Initially, the text present in the reviews alone is extracted, removing structured tags like HTML and XML. The task of vector creation involves the creation of a word vector from the datasets. The tf-idf (short for term frequency inverse document frequency) method is used to reflect how important a word is to a collection of data or corpus [26]. The word vector is created using the tf-idf method. The text is then pruned; the common and infrequent terms get eliminated.

A number of text pre-processing operations are performed before they can be used to train the classifier model. They include the process of tokenizing, stemming, filtering tokens by length and removal of stop words.

The task of Sentiment Classification was carried out with the help of two supervised learning approaches. Supervised Learning is a type of machine learning method which builds a predictor model based on the given training data. Three widely used supervised machine learning algorithms were used for the classification: Support Vector Machines, Logistic Regressions and Naive Bayes Classifier.

1) Naive Bayes Classifier: The Naïve Bayesian Classifier is a statistical based classifier which is based on the Bayes Theorem. The Bayes theorem calculates the posterior probability (H |X) from the prior probability P (H), P (X |H) which is the posterior probability of X conditioned on H and P(X) which is the prior probability of X where H is the hypothesis and X is the given data set.

$$P(H|X) = \frac{P(X|H)\,P(H)}{P(X)} \qquad (1)$$

Naive Bayes Classifier works on the assumption that the effect of an attribute on a class is independent of the other attributes. Such an assumption is called a nave independence assumption thus lending the classifier its name. Naive Bayesian Classifiers are very efficient since they are less computationally intensive and require a small size of training data. However they can fall behind in performance

2*) Support Vector Machines*: Another commonly used and powerful supervised learning algorithm (besides logistic regression and neural networks) is Support Vector Machine. The support vector machine is also sometimes called a “large margin classifier” because when defining the decision boundary between two classes it tries to maximize the margin between each class and the boundary.

The support vectors are what that defines the margins on either side of the decision boundary, between the boundary and the data points. SVM performs the classification operations and has a high accuracy rate. It is popular because it can be easy to use and the same algorithm resolves a diverse range of problems with minute adjustments.

3) Logistic Regression: Logistic Regression is a type of statistical classification machine learning algorithm. The basis of this algorithm is derived from the sigmoid function, $g(z) = \frac{1}{1+e^{-z}}$. Logistic Regressions are especially useful in binary classification problems. As the present classification concerned is a binary type, Logistic Regressions are highly applicable.

4) Training of Classifier Model: After the training data has been pre-processed cross validation is performed using the ‘X-Validation’ operator. Cross validation is a standard way to assess the accuracy and validity of a statistical model. The dataset is further divided into two parts, training set and the test set. This process is repeated for a number of times. Next, the training and testing parts of the cross validator is populated.

A number of classifiers including SVM, Logistic Regressions Nave Bayes are used to test the model. The model is tested by using the ‘Apply Model’ operator. To measure the model accuracy the ‘Performance Operator’ is used.

5) Applying the classifier model on the reviews: The reviews extracted for each product by the information retrieval component is now passed as data to the classifier model for sentiment classification. The model is run and the results of the classification are stored in a .csv (Comma Separated Values) file.

Sno(Review)	Confidence (Positive)	Confidence (Negative)	Prediction (Label)
1	0.506	0.494	positive
2	0.605	0.395	positive
3	0.768	0.232	positive
4	0.309	0.691	negative
5	0.842	0.158	positive
6	0.146	0.854	negative
7	0.843	0.157	positive
8	0.414	0.586	negative
9	0.208	0.792	negative

TABLE I: Classifier output with confidence values

As shown in the above table, the reviews are classified as positive or negative. The confidence attributes indicate the degree of the prediction. Basically the attributes are a measure of the probability involved in the prediction. If the positive confidence is greater than the corresponding negative confidence, the prediction label is positive otherwise its negative.

The calculation of sentiment scores for each product involves using the confidence values. Since the confidence values indicate the degree of sentiment expressed in the reviews, the overall score for each product is calculated using them.

$$R_i = 5 \sum_j^N C_p(j)/N_i \qquad (2)$$

Where,

R_i indicates the overall sentiment score for the i^{th} product,

C_p (j) indicates the positive confidence value of the j^{th} review for i^{th} product and

N_i indicates the total number of reviews for the i^{th} product.

C. Evaluation and Statistical Comparison of the Classifiers

Evaluation of the classifiers can be done on many parameters. This paper attempts to use the Receiving Operator Characteristics (ROC) of the classifiers for evaluation. ROC curves measure the ratio of false positives and false negatives. A ROC curve is a plot between two parameters - TPR and FPR respectively.

TPR, also known as True Positive Rate is defined as

$$TPR = \frac{TP}{TP+FN} \qquad (3)$$

Where, TP is the number of true positives and FN is the number of false negatives. FPR, also known as False Positive Rate is defined by

$$FPR = \frac{FP}{FP+TN} \qquad (4)$$

A ROC curve as explained in [23] is primarily a plot of TPR vs. FPR as the values of the threshold θ is varied. The threshold controls the number of entities the classifier computes correctly. Normally, a value of θ = 0 yields a classifier which classifies everything as positive and a value of θ = 1 yields a classifier which classifies everything as negative. An ideal classifier has its threshold value somewhere near 0.5.

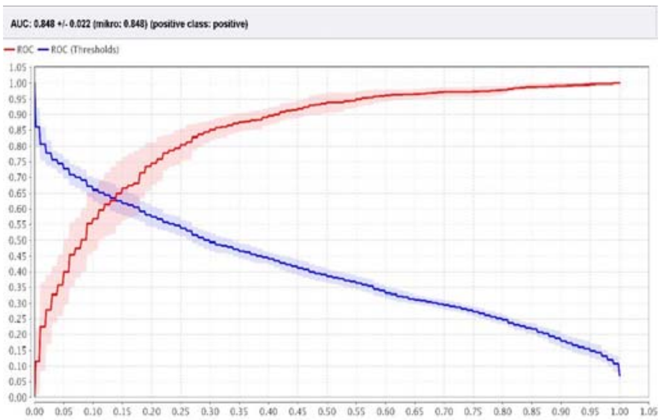

Fig. 2:ROC curve for Linear SVM

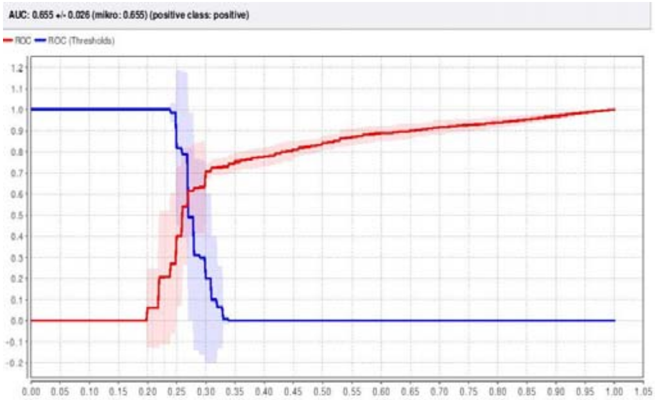

Fig. 3:ROC curve for Naïve Bayes Classifier

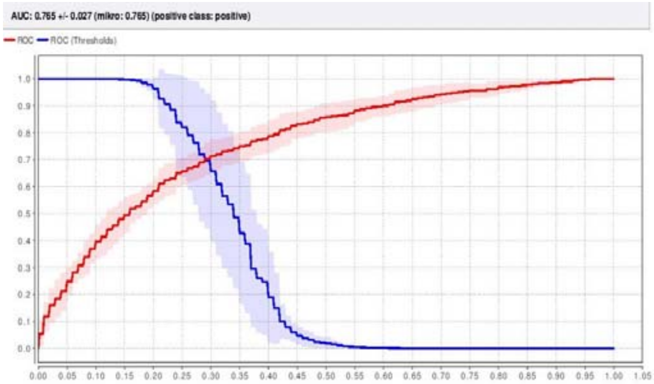

Fig. 4:ROC curve for Logistic Regression

Three different classifiers were used Support Vector Machines, Logistic Regression and Nave Bayesian Classifier.

The above figures show the ROC curves for each of the classifiers used. An ideal system achieves the point on the curve where (FPR=0, TPR=1) .An often used metric used to measure the quality of a ROC curve is the AUC (Area under Curve) metric. The AUC scores range from 0 to 1 where an ideal classifier has its AUC as 1.

From the three ROC curves plotted for the classifiers, the Support Vector Machine has a better AUC score of 0.848, whereas the Logistic Regression has a score of 0.765 and the Nave Bayesian Classifier giving an AUC score of 0.655.

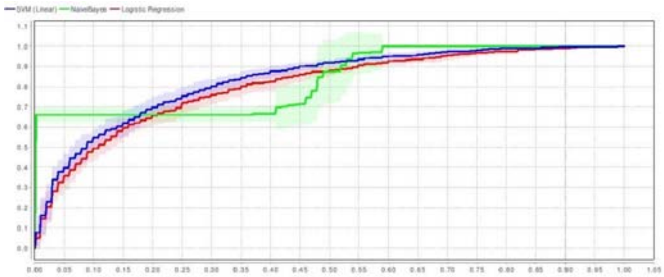

Fig 5. ROC comparison of the classifiers

The above figure shows the ROC comparison of the three classifiers used. From the plot, using the AUC metric it is concluded that the Support Vector Machine is the best classifier of the three.

A single measure for the classifier evaluation may not lead to the right results. Another method of evaluating classifiers is by comparing the confusion matrices of the three matrices.

accuracy: 77.35% +/- 2.12% (mikro: 77.35%)

	true negative	true positive	class precision
pred. negative	705	158	81.69%
pred. positive	295	842	74.05%
class recall	70.50%	84.20%	

Fig. 6:Confusion Matrix of SVM

accuracy: 70.45% +/- 2.41% (mikro: 70.45%)

	true negative	true positive	class precision
pred. negative	726	317	69.61%
pred. positive	274	683	71.37%
class recall	72.60%	68.30%	

Fig. 7:Confusion Matrix of Naïve Bayesian Classifier

accuracy: 70.05% +/- 2.40% (mikro: 70.05%)

	true negative	true positive	class precision
pred. negative	672	271	71.26%
pred. positive	328	729	68.97%
class recall	67.20%	72.90%	

Fig. 8:Confusion Matrix of Logistic Regression

From the three confusion matrices it is observed that the accuracy of the Support Vector Machine is the highest with an accuracy of 77.35% whereas the accuracies of the other two classifiers are around 70%.

D. Aspect Based Opinion Summarisation

Aspect based opinion summarization captures the essence of opinions: opinion targets (entities and their aspects) and sentiments about them. Also this task is quantitative, that is it gives the number or percentage of positive or negative opinions about the aspects. The reviews are mined for the product aspects and the corresponding opinions about each aspect are extracted. Using the sentiment model trained earlier the opinions are classified as positive or negative.

Using the results obtained a visual comparison on the opinions about the different aspects for each product is designed. Various types of summarized have existed including statistical summaries, text selection methods, aggregated ratings and summaries with timelines. One such visualization used in [1] has been presented here.

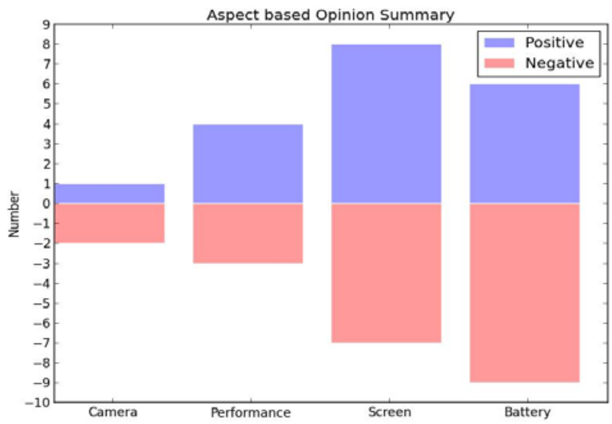

Fig. 9:Aspect Based Opinion Summary

Here, the number of positive and negative opinions for each aspect of the product is plotted. This offers a perspective of how the various aspects have been reviewed.

E. Content based Recommender System

The recommender system provides the most optimal product among the selected products and also displays the recommended products to the user. Using the sentiment classification model, the sentiment scores for each product are calculated. The product with the highest sentiment score is selected as the optimal product and shown to the user.

As explained in [24],the concept of similarity among the feature vectors is used to build the content based recommender system. To calculate the similarity between the different entities, the distance between the entities when expressed as a feature vector is calculated. When two feature vectors are close to each other, they have a high degree of similarity.

In a multi-dimensional space the distance between two feature vectors is calculated using the Minkwoski Distance. Given two points, P($x_1, x_2, x_3 \ldots x_n$) and Q($y_1, y_2, y_3 \ldots y_n$)) $\in$ R ,the Minkwoski Distance is calculated as $\left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{\frac{1}{p}}$. When p=2, this distance is called as Euclidian Distance.

The recommender system uses two attributes for the feature vectors: the sentiment scores and the costs of the products.

The data can be represented in the form of a matrix, where the rows correspond to a particular product and the columns correspond to the score and cost for each product:

$$\begin{pmatrix} 3.21 & 7462 \\ 3.45 & 10735 \\ 3.12 & 17620 \\ 3.14 & 8981 \\ 3.25 & 12998 \\ 2.64 & 14890 \\ 3.22 & 8499 \\ 3.03 & 11596 \\ 2.73 & 13320 \\ 3.31 & 18649 \end{pmatrix}$$

Similarly the feature vectors can be represented as:

$X^{(1)} = (3.21,7462)^T$, ... and $X^{(10)} = (3.31,18649)^T$ where each feature vector refers to a product. The products that are similar to the selected three products are plotted using the Euclidian Distance measure.

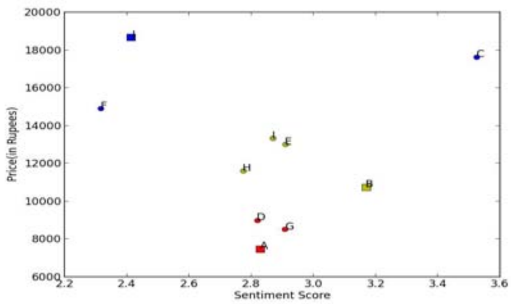

Fig. 10: Sentiment Score vs Price Graph based on Similarity

The above graph shows the products grouped together based on their similarity. The products similar to the optimal product (chosen before) are thus displayed to the user as the list of recommender products.

IV. CONCLUSIONS AND FUTURE WORK

A novel recommender system using sentiment classification algorithms to provide the best products to the customer was been proposed. Various sentiment classification algorithms including Naive Bayes Classifiers and Support Vector Machines were been used for the task of classification.SVMs were found to be the most accurate classifier using various statistical evaluative measures with an accuracy of 77.35 %.Using the ROC parameter too, the finding was consistent with SVMs having the highest AUC score of 0.848.Using the concept of similarity, a content based recommender system was built to provide the optimal products to the user.

RecoProd has been developed only for mobile phones and it is found to work efficiently. In future it is possible to extend the scope of the system to any type of consumer product with added customization options for the users. A limited number of classifiers were used for the classification. More number of classifiers can be used to provide a greater depth into the performance evaluation of the classifiers. Various factors involved in the field of Sentiment Analysis like sarcasm in opinions, detection of fake opinions etc. can be used further to get a better sentiment score of the products thus improving its efficiency.

REFERENCES

[1] Bing Liu,*Sentiment Analysis and Opinion Mining*,Morgan Claypool Publishers,May 2012

[2] K. Dave, S. Lawrence, and D. M. Pennock,*Mining the peanut gallery: Opinion extraction and semantic classification of product reviews*,Proceedings of WWW, pp. 519528,2003

[3] Lei Zhang, Bing Liu, Suk Hwan Lim, and Eamonn O'Brien- Strain. *Extracting and ranking product features in opinion documents*,In Proceedings of the 23rd International Conference on Computational Linguistics: Posters (COLING '10) Association for Computational Linguistics, Stroudsburg, PA, USA, 1462-1470,2010

[4] Lisette Garca-Moya, Henry Anaya-Snchez, and Rafael Berlanga-Llavori, *Combining probabilistic language models for aspect- based sentiment retrieval*,In Proceedings of the 34th European conference on Advances in Information Retrieval (ECIR'12), Ricardo Baeza-Yates, Arjen P. Vries, Hugo Zaragoza, B. Barla Cambazoglu, and Vanessa Murdock (Eds.). Springer-Verlag, Berlin, Heidelberg, 561-564,2012

[5] Christopher Manning and Hinrich Schtze,*Foundations of Statistical Natural Language Processing*,The MIT Press, Cambridge, Massachusetts. May 1999

[6] www.sentiment.christopherpotts.net

[7] Steven Bird, Ewan Klein, and Edward Loper,*Natural Language Processing with Python*,The MIT Press, Cambridge, Massachusetts, OReilly Media,2009

[8] Peter.D.Turney, *Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews*,Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL),Philadelphia,July 2002, pp.417-424.

[9] C. Lin and Y. He,*Joint Sentiment/Topic Model for Sentiment Analysis*.,The 18th ACM Conference on Information and Knowledge Management (CIKM),Hong Kong, China,Nov. 2009

[10] D. Thenmozhi, Ajay Shrivatsav V.P, Bhawani Pratap Singh Manhas and Kishorekumar P.,*Extracting Customer Needs for Electronic Gadgets using Opinion Mining.*,International Journal of Advanced

Engineering Research and Studies. Vol. I/ Issue III/April-June,2012/01-04

[11] Lisette Garca-Moya, Henry Anaya-Snchez, and Rafael Berlanga-Llavori,*Retrieving Product Features and Opinions from Customer Reviews,*IEEE Computer Society,May/June 2013

[12] Yu Zhang and Weixiang Zhu. 2013,*Extracting implicit features in online customer reviews for opinion mining*,In Proceedings of the 22nd international conference on World Wide Web companion (WWW '13 Companion). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 103-104.

[13] Satoshi Morinaga, Kenji Yamanishi, Kenji Tateishi, and Toshikazu Fukushima,*Mining product reputations on the Web*,In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '02). ACM, New York, NY, USA, 341-349.

[14] J. McAuley and J. Leskovec.,*Hidden factors and hidden topics: understanding rating dimensions with review text,*RecSys. 2013

[15] Larissa A. Freitas and Renata Vieira,*Ontology based feature level opinion mining for portuguese reviews,*In Proceedings of the 22nd international conference on World Wide Web companion (WWW '13 Companion),International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 367-370,2013

[16] Shuang Liu, Fang Liu, Clement Yu, and Weiyi Meng,*An effective approach to document retrieval via utilizing WordNet and recognizing phrases*,In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '04). ACM, New York, NY, USA,266-272,2004

[17] Wei Zhang, Lifeng Jia, Clement Yu, and Weiyi Meng,*Improve the effectiveness of the opinion retrieval and opinion polarity classification*,In Proceedings of the 17th ACM conference on Information and knowledge management (CIKM '08),ACM, New York, NY, USA, 1415-1416,2008

[18] Koji Eguchi and Chirag Shah,*Opinion Retrieval Experiments using Generative Models*,2006

[19] Bin-Yang Li,Lan-Jun Zhou,Shi Feng,Kam-Fai Wong,*An efficient approach for sentence-based opinion retrieval,*Machine Learning and Cybernetics (ICMLC), 2010 International Conference on , vol.6, no., pp.3281,3286 ,11-14 July 2010

[20] Bo Pang and Lillian Lee,*Using very simple statistics for review search: An exploration,* 2008

[21] Feng Wang and Li Chen,*Recommendation based on Mining Product Reviewers Preference Similarity Network,*2012

[22] Taboada, Maite, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Mandred Stede,*Lexicon-based methods for sentiment analysis,* Computational Linguistics: 17 (2)

[23] Kevin P. Murphy,*Performance evaluation of binary classifiers*, Technical Report, University of British Columbia, 2007

[24] Hiroshi Shimodaira, *Similarity and recommender systems* School of Informatics,The University of Edinburgh, 2014

[25] www.rapidminer.com

[26] http://en.wikipedia.org/wiki/Tf