JIBCA: Jaccard Index based Clustering Algorithm for Mining Online Review

Nihalahmad R. Shikalgar
PG Scholar,
Department of Computer,
PVPIT, Pune,

Arati M. Dixit
Associate Professor,
Department of Technology,
University of Pune,

ABSTRACT
Sentiment analysis, also known as opinion mining, is the analysis of the feelings (i.e. attitudes, emotions and opinions) behind the words. Sentiment analysis involves classifying the opinions as positive, negative, or neutral. Classification of textual objects in accordance with sentiment is considered to be a more difficult task than classification of textual objects in accordance with the content because opinions in natural language can be expressed in subtle and complex ways containing slang, ambiguity, sarcasm, irony, and idiom. This paper investigates the problem of sentiment analysis of online review. A Jaccard index based clustering algorithm (JIBCA) is proposed to support mining online reviews and predicting sales performance. The information gain is the change in information by considering number of datasets. The performance of *information gain* varies depending on the dataset. It is observed that the information gain performed better in JIBCA than existing methods for the movie review dataset. It is therefore recommended that JIBCA can be a good feature selection method for sentiment classification tasks. This paper also proposes a new approach for movie reviews classification based on extraction and analysis of appraisal groups such as action, thrill, comedy, and romantic.

General Terms
Data mining.

Keywords
Sentiment Analysis, Review Classification, Opinion Mining, Review Mining, Prediction.

1. INTRODUCTION
The increasing use of the internet has changed the way people shop for goods, which includes the purchase of online movie tickets. Traditional user also sells or buys by using internet that shows increasing usability of same. Number of product review forum invites people to express their experience of the said product. Review forum has been playing a vital role in increasing sales performance of a product. A user experience is one of key term to get knowledge of certain product before purchasing. So consequently it is important in decision support. Online people increasingly rely on alternative sources of information such as movie review provided by different social sites.

Data available on the internet comes in a variety of different formats. The simplest and most common format is plain textual data. One example of textual data is online reviews. Online reviews are broadly described as either positive or negative reviews [8]. An objective review tends to contain mostly facts, while a subjective review contains mostly opinions. There are two common review formats [9]: the restricted review format and the free format. In a restricted review, the reviewer is asked to separately describe the pros and cons [1], and to write a comprehensive review. In a free format review, the reviewer writes freely without any separation of the pros and cons. There is a potential issue with using only numeric ratings as being representative of the information contained in movie reviews. By compressing a complex review to a single number, here implicitly assume that the product quality is one-dimensional, whereas economic theory tells us that any movie has multiple attributes and different attributes can have different levels of importance to people. Thus, unless the person reading a review has exactly the same preferences as the person who wrote the review, a single number like an average movie rating might not be sufficient for the reader to extract all information relevant to the watching decision.

In movie review domain, recent studies have been focused on the reviews for finding the relationship between the sales performance of the movie and reviews. The actionable knowledge developed by using average number of the quality reviews and the number of people rated the reviews in the blogs and IMDB (http://www.imdb.com) websites. Not only one time review but also need to consider reviews form date of movie release to one week after release. So that one of find a pattern and determines future sales performance of movie. Online reviews [2], [3], [4], [5], [6], [7], [8] are readily available publically all over world on internet. Predicting sales performance is completely a domain driven task, it considers the public sentiments and box office revenues. Some negation word also affects the prediction so that there is need to find such word. The proposed system can analyze such word and handle it, so that it can improve quality of prediction.

In summary, this paper makes the following contributions:

- Using the movie domain as a case study, it approaches the problem of predicting sales performance using online reviews, by finding different factors involved in review mining.
- The proposed JIBCA model with use of appraisal groups, provides a Jaccard framework to analyze sentiments in reviews.
- It models sentiments in reviews as a hidden sentiment factor involved in expression. It also accomplishes complex nature of sentiment.
- It discusses how actionable knowledge can be derived by utilizing the proposed models, explaining the practical impact of the proposed approach.
- It also proposes novel approach to predict the type of movie like - Action, Thriller, Comedy, Romantic, etc.

2. RELATED WORK

2.1 Online Review

A critical appraisal of a book, play, film, etc. is published on social networking web sites as well as peer review websites. Typically review is a sentiment expressed in consideration of any product and thus is true for films. Review can be expressed in two ways - offline review and online review. Reviews expressed in magazine, news paper are termed as offline review. A formal assessment of something with the intention of instituting change is necessary. Review expressed in various websites is termed as online review. Online reviews [2], [3], [4], [5], [6], [7], [8] have greater impact on the process of knowing product sentiments expressed by user. It is also useful for producer or stock holder to measure performance of the system. It is very convenient to collect peer review by the expert as well as user of the system. Online reviews are (at least in principle) contributed by people who have watched the movies being rated. It is expected that their early volume will exhibit a strong correlation with the corresponding box office revenues.

2.2 Online Review Mining

By considering user experience as a matter of fact has an impact on giving preference to purchase any kind of product or watching movie. Review mining is one of the important components of it. Many online blog and social networking websites are available, where many people have been expressing their review with respect to product and movie. By considering these reviews, it is very supportive to increase sales performance. Numbers of social sites are available to user for sharing their views. Due to this trend these days online review mining [22] has gained considerable attention by the researchers. In early days most of the work focused on semantic analysis. Different classifiers like Bayesian [1], [10], Support vector machine [3], [13] are used to treat review as positive, negative, neutral. This kind of online review mining is considered at document level, node level, and word level. In document level, review mining focuses on the reviews analysis of a whole document. In node level, single statement is considered to analyze reviews. In word level, single word is in considered for analysis of reviews. Some of the studies have resulted in evolution of learning involved in a positive/negative classifier at the document level. It is very analytical to carry out review analysis.

2.3 Domain Driven Data Mining

Domain Driven Data Mining (D3M) overcomes the traditional data-centered review mining framework. Since last few years domain-driven data mining has emerged as an important new standard for knowledge finding. Domain Driven Data mining is composed of Human intelligence, Domain intelligence, and Network intelligence. Human intelligence consists of opinions of user and information of other comments posted on number of social sites. Domain intelligence consists of movie database or IMDB sites available to impart movie review analysis [21].

Knowledge discovery from database is a part of KDD [6]. Knowledge discovery from database discovers knowledgeable data by mining raw data. Raw data is considered as input to the system. By using this data, further apply data mining algorithm so as to discover knowledge from it. This knowledge is of the pattern or a collective analysis report. This is very important for business to increase earnings of an organization. Here actionable knowledge [6] is considered for the knowledge discovery. A KDD is an iterative optimization process toward the actionable model, considering surrounding business environment and Problem state.

2.4 Sentiment PLSA

In Sentiment Probabilistic Latent Semantic Analysis [1] a review can be considered as being generated under the authority of a number of hidden sentiment factors. For the said purpose numbers of blogs are available. Separating number of blogs from other blog is a very difficult task. Categorize each one blog and consider only blogs which are related with a movie review for analyzing a movie. If word review pair is considered, then each word expresses the positive and negative comments. Initial task is to separate the sentiment in different categories like positive, negative, and average. The numbers of other words are present into a sentiment; those unnecessary words are increasing calculation. Optimization is done to eliminate unnecessary calculation. The people can observe their opinions and increase their feel to watch the movie, if the reviews are good. If the movie is bad and the reviews are dreadful, then it will have a serious impact on the business of the said movie. Reviews posted online play a significant role on the movie's success in terms of sales. It is also observed that the bad movie reviews results in failure of movie to get the right place in the box office in the movie database site.

Step wise procedure of SPLSA model described below,

- Feature Extraction: Initially word review pair is calculated,
- For a word-review pair (w, b), S-PLSA models the co-occurrence probability as a mixture of conditionally independent multinomial distributions by separating word in blog and taking into consideration.
- Word Blog analysis: To this end, extracted feature is to be considered. Analyze each word review pair. We compare and test many blog and review from first step.
- Heuristics Algorithm: An Expectation step, a heuristic is developed to generate optimal strategy.
- Optimization of sentiment: Optimization is done to eliminate unnecessary calculation,
- Display Analytics: Last Stage based on classification and display of the consumer sentiment.

3. JIBCA

A Jaccard index based clustering algorithm (JIBCA) is proposed to support mining online reviews and predicting sales performance. It is a clustering and regression based algorithm for sentiment prediction of online data. Clustering [9] can be considered the most important unsupervised learning problem, like every other problem of this kind, it deals with discovering a *structure* in a collection of unlabeled data. A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way". A *cluster* is therefore a collection of objects which are "similar" in a group/cluster and are 'dissimilar' to the objects belonging to other clusters. So problem solution is not to use a kind of supervised learning. By using clustering, the records of the movies data is collected simultaneously and made accessible for analysis. Fig. 1 indicates the JIBCA System architecture. User can express their sentiment on various social networking website like IMDB sites, where movies related data is stored. System should use all review collectively for further processing. Blog review can be collectively used for preprocessing. Preprocessed data is useful for feature selection. Here it extracts feature mined from review, which is useful for review analysis. Word blog

analysis is one of the important factors. Review word pair should have to find it out from a blog. From that word pair Testing Set (TS) is formed. Testing set is a collection of word pair collected from blog review. TS are useful for JIBCA algorithm. Training set (TrS) and Testing set are made available as an input for CRBSP algorithm. TrS is a number sample set review file which is already classified according to different categories. TrS and TS - are the sample set available for further processing. Jaccard coefficient is calculated by comparing testing set with training set. Jaccard coefficient is very important in this entire concept. Jaccard coefficient is used to find out similarity and dissimilarity measure from the set of reviews. Negation of similarity measure will find out dissimilarity. In this case average number of similarity measure with set of positive domain is treated as a positive review whereas average number of similarity measure with negative domain or set is treated as negative comments. By consideration of this here formulate different categories of comment as shown in Table 1. Jaccard index based clustering algorithm (JIBCA) and the mathematical model is discussed in the next sub-section.

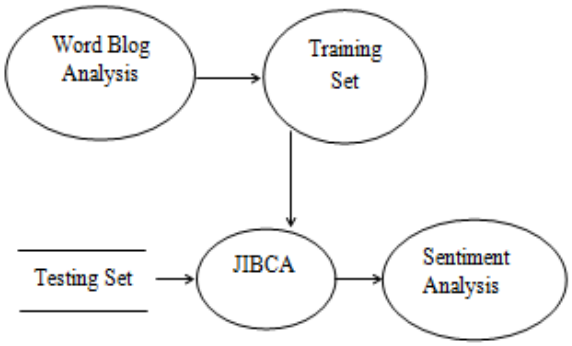

Fig 1: JIBCA Architecture

3.1 Mathematical Model Design

The Jaccard index [13] also known as the Jaccard similarity coefficient, is a statistic used for comparing the similarity and diversity of sample sets. The Jaccard coefficient measures similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets.

Let $B = \{b_1, b_2, ..., b_n\}$ be the number of blog containing online reviews.

Let $W_i = \{w_1, w_2, ..., w_m\}$ be the words in blog b_i,

Where, $i = \{1, 2, 3, \ldots, n\}$

Let $A = \{a_1, a_2, \ldots, a_{2000}\}$ be the set of appraisal words used for analysis.

The Jaccard index similarity is given by,

$$\therefore S = \frac{2|X \cap Y|}{|X| + |Y|} \quad \ldots\ldots\ldots\ldots\ldots [1]$$

Where X & Y are two samples, S should be $0 \leq S \leq 1$,

For negativity, formulate equation as defined below,

$$\therefore N = 1 - S$$

where, N is a set of negation words in movie review pair.

For this formulate problem with the given set of blogs B & Wi as set of words in each blog bi to classify the blogs into various sentiments.

From the JIBCA Algorithm graphical data analytics shown this indicates the movies sales performance.

3.2 Jaccard Index based Clustering Algorithm

Jaccard coefficient is used for similarity and diversity measure of textual data. So here in this paper, Jaccard coefficient measure is used for sentiment analysis.

Algorithm 1: Jaccard Index Based Clustering Algorithm
Input: Set C = {C1, C2, C3}, where C is set of appraisal words. Set S = {S1, S2, S3, ……, Sn}, where S is the set of online review words **Output:** Set of positive & negative words.
Step 0: Start Step 1: For a given blog b, Calculate top frequently used words and form the vector. Step 2: Select k randomly and form k clusters. Step 3: Assign the vector of frequently used words to one per each cluster. Call this as centroid. Step 4: Assign the remaining vectors of other blogs to every cluster by comparing Jaccard similarity of equation 1. Step 5: Update the centroid & assign vectors as per new centroid. Step 6: Repeat step 4 & 5 until centroid don't move. Step 7: Compare Set C & S to find matching words. .Declare the sentiment for every cluster by using this method. Step 8: Declare overall sentiment for all clusters.

4. RESULT ANALYSIS

This section discusses the results obtained from a set of experiments conducted on a movie data set in order to validate the effectiveness of the proposed model, and compare it against alternative methods.

4.1 Experiment Methodology

The movie data used for experimentation consists of the Cornell movie review dataset (sentiment polarity dataset v2.0) and the congressional-speech corpus [20]. The movie review dataset is a popular dataset that has been previously used for research in sentiment classification. For the sake of input, the movie review from rotentammatoes social sites [24] was manually collected for finding out types of movie. It contains a social site which consists of online review of movies. In execution of the experiment, the following procedure was followed:

- Step 1. Choose randomly half of the movies for training and the other half for testing. The movie reviews are partitioned correspondingly.
- Step 2. Using the training blog entries, train JIBCA model. For each blog entry b, the sentiments toward a movie are summarized using a vector of the posterior probabilities of the hidden sentiment.
- Step 3. Feed the probability vectors obtained in Step 2 and predict sales performance.

- Step 4. Then evaluate the prediction of types of movies according to classification.

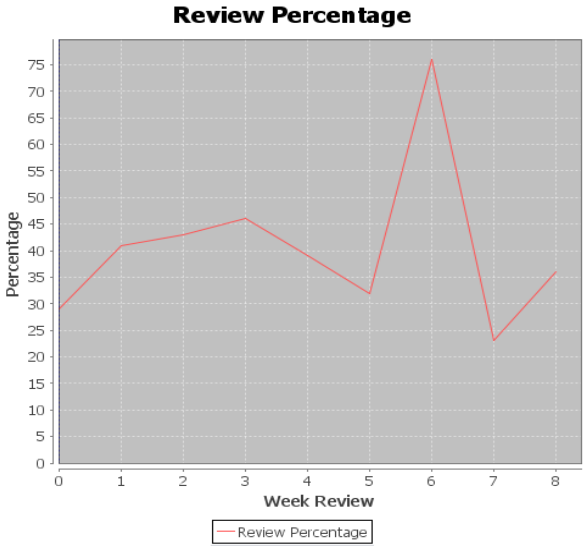

Fig 2: User Trend towards movie.

4.2 Sentiment Prediction

Sentiment prediction is with reference to the positivity and negativity present in document. It is predicted by analyzing and comparing with appraisal word set [23]. In this scenario, if system has given an input to movie review then by using this system it can calculate the percentage of positive and negative feedback. For example if providing movie review from Cornell dataset gives resultant output as following:

Positive Feedback: 04%

Negative Feedback: 70%

Here in this example 70 % of movie reviews are negative. Negative percentage is more than positive percentage so this movie fits in negative category.

Table 1 Classification of Review

Sr. No.	Range of Percentage	Category
1	85-100	Pure Positive
2	75-85	Positive
3	50-75	Average Positive
4	25-50	Average
5	15-25	Average Negative
6	0-15	Pure Negative

Beyond the percentage of positivity and negativity, classify it further according to sub classes like in Table 1. If movie reviews having percentage between the 85-100 % then this movie has a *purely positive* response. If movie review is having percentage between 75 to 85 then this movie is having *positive* response. If movie reviews obtain percentage in range of 50-75 then this movie has obtained *average positive* response. On similar lines the various categories on basis of the *Range of Percentage* as shown in Table 1 is defined as a set:

Category = C = {Pure Positive, Positive, Average Positive, Average, Average Negative, Pure Negative}

By considering movie reviews between *date of movie release* and date one week after the movie released, resultant output is shown in Fig. 2. In this analysis the review of particular kind of movie after released is represented. It shows that User trends towards the movie per week. Typically the percentage of positivity and negativity is calculated in the reviews post movie release for few weeks. Accordingly the graph is plotted to represent the processing of per week user online review.

4.3 Movie Classification According to Type

The result of prediction of movie review categorization on basis of their type includes comedy, Romantic, Thriller, and Action. Here classifier is used for categorization of review into these types. Different classes should consist of set of appraisal words [23]. It is classified on basis of both the review set and the appraisal set.

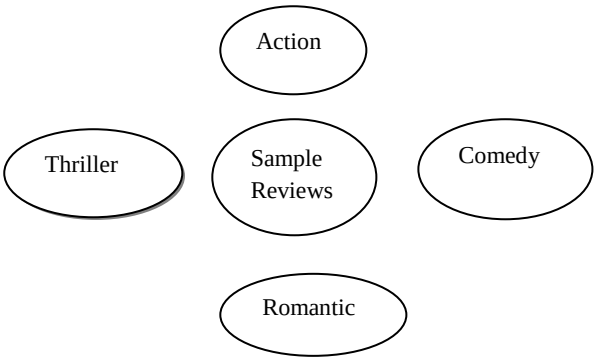

Fig. 3: Appraisal set [23]

Let us consider a case –study with following details:

Input: Movie Review Set

"This film is stylish, intelligent and, one hopes, influential on the next generation of superhero movies; it should leave a lasting legacy."

Output: Predicted Type of Movie.

Set of Appraisal words found: influential next generation superhero movies; should leave lasting legacy.

Type: Thriller

It is observed that from the given review set, an appraisal set is extracted to predict the type of movie to be – 'Thriller'.

5. PERFORMANCE ANALYSIS

To evaluate the performance of proposed JIBCA system, an empirical study is conducted on both blog documents and IMDB movie reviews. As similar trends are observed on those two data sets, only the experimental results for the blog documents are demonstrated from Sections 4. To verify the effectiveness of proposed quality-aware model, only the IMDB movie reviews are adopted, as each post in this data set is associated with a clear helpfulness score which can be used for training and testing. However, such information is not available for the blog data set. It is observed that the performance of JIBCA system is better than preceding work. Fig. 3 shows graph of JIBCA Vs Naïve Bayes Classifier. Experimental result shows the JIBCA accuracy is 68% whereas; Naïve Bayes classifier accuracy is around 60%. So it is more predictive to have JIBCA system.

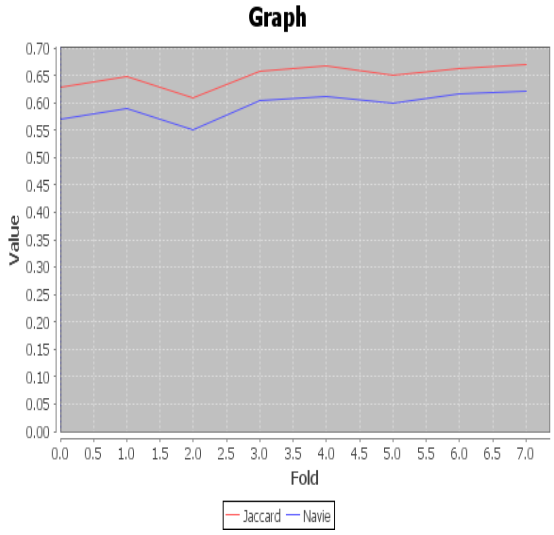

Fig 4: Comparisons of JIBCA Vs Naïve Bayes

The Cornell movie review dataset is consisting of 200 to 300 reviews. This review can be divided into each fold value. Fold value is concerned with total number of reviews divided by some constant value. Constant value can take any one value so that movie review is divided into some groups. Every time training set trains each of folds. Fold act as testing set. So that it can be flexible to watch results in different fold value. It's easy to see predictive Jaccard coefficient value.

This problem is non- deterministic polynomial hard problem. Hidden semantic analysis is very tedious to be uncovered. So this problem is not completed between deterministic time periods.

6. PRACTICAL IMPACT OF PROPOSED MODELS

JIBCA System accurately predicts orientation of sentiment. Some of the reviews may consist of some complex word pair. In that case analysis of such word pair is very important. Such word pair generally comes in movie review. So let's consider word pair saying: *It is not good movie*. Which means that movie is bad. But according to set of appraisal, sentence consists of word good which comes from positive set of appraisal. Because of this, review is considered to be in positive comment type. In fact a wrong result will be produced. It is necessary to replace such word with opposite word. Include Opposite word with those words that comes with negation word. Here negation word may be denoted by '*not*'. After replacing such word recall word pair and then find correct output.

Proposed system is working very well in relational database management system and non RDBMS. It is having an impact on both types of datasets, where it can accurately predict sales performance of movie. There is no need of supervised learning for proposed system. There is no need of supervisor to analyze and monitor whole system, without supervisory control it can accurately work. It is very beneficial for user to understand about not only orientation i.e. positive/negative but also more details like *type of movie*, because user may have interest in one particular kind of movie. For the purpose of same domain, the JIBCA system successfully diagnoses accurate type of movie.

In experimentation a set of appraisal word for analyzing hidden sentiment factor of review is used. Thus the prediction depends upon that set appraisal words. Text sentiment analysis is also in considered, so that textual review can be taken for processing. Meanwhile some review may consist of audio and/or video format. Our system doesn't work with audio and video data. Many parameters used to build mathematical model are not considered e.g. semantic analysis.

7. CONCLUSIONS AND FUTURE WORK

The escalating use of online reviews as a way of conveying views and comments has proved to be a distinctive approach to find sales performance and derive business intelligence. In this paper, the problem of predicting sales performance using sentiment information mined from reviews is studied and a novel JIBCA Algorithm is proposed and mathematically modeled. The outcome of this generates knowledge from mined data that can be useful for forecasting sales.

A sentiment prediction in JIBCA is achieved by using Jaccard Similarity. Jaccard Similarity is useful for analysis of sentiment that assists in the process of classification of different categories of sentiments in blogs. A generative model for sentiment analysis helps us move from simple "negative or positive" classification towards a deeper comprehension of the sentiments in blogs. The proposed algorithm successfully and correctly predicts sales performance.

Predicting sales performance is a challenging task for anyone without the proper information. This paper explored predicting sales performance in the movie field. For future work, by using this framework it can extend it to predicting sales performance in the other domains like customer electronics, mobile phones, computers based on the user reviews posted in the websites etc.

8. REFERENCES

[1] Xiaohui Yu, Yang Liu, Xiangji Huang, Aijun, "Mining Online Reviews for Predicting Sales Performance: A Case Study in the Movie Domain", A Knowledge and Data Engineering, IEEE Transactions on 24, 2012.

[2] Deng Bin, Peiji, Shao, Zhao Dan, "E-Commerce Reviews Management System Based on Online Customer Reviews Mining Innovative Computing & Communication", 2010Intl Conf on and Information Technology & Ocean Engineering, 2010 Asia-Pacific Conf on (CICC-ITOE)2010

[3] Soliman, T.H.A. Elmasry, M.A.Hedar, A.R., Doss, M.M., "Utilizing support vector machines in mining online customer reviews" Computer Theory and Applications (ICCTA), 2012 22nd International Conference on2012

[4] Peng Jiang, Chunxia Zhang, Hongping Fu, Zhen dong Niu, Qing Yang, "An Approach Based on Tree Kernels for Opinion Mining of Online Product Reviews" ,Data Mining (ICDM), 2010 IEEE 10th International Conference on2010

[5] Lai C.L., Xu K. Q., Lau R.Y.,K.Yue feng Li, Dawei Song, "High-Order Concept Associations Mining and Inferential Language Modeling for Online Review Spam Detection", Data Mining Workshops (ICDMW), 2010 IEEE International Conference, on 2010

[6] Weishu Hu, Zhiguo Gong, Jingzhi Guoe ,"Mining Product Features from Online Reviews" , Business Engineering (ICEBE), 2010 IEEE 7th International Conference ,on 2010

[7] Samsudin N., Puteh M., Hamdan A.R., "Bess or xbest: Mining the Malaysian online reviews" Data Mining and Optimization (DMO), 2011 3rd Conference ,on 2011

[8] XuXueke, Cheng Xueqi, Tan Songbo, Liu Yue, Shen Huawei, "Aspect-level opinion mining of online customer reviews Communications", China102013

[9] Lin E., Shiaofen Fang, Jie Wang ,"Mining Online Book Reviews for Sentimental Clustering", Advanced Information Networking and Applications Workshops (WAINA), 2013 27th International Conference on 2013

[10] Anwer, N. Rashid, A. Hassan, "Feature based opinion mining of online free format customer reviews using frequency distribution and Bayesian statistics", Networked Computing and Advanced Information Management (NCM), 2010 Sixth International Conference on2010

[11] Algur S.P., Patil A.P., Hiremath P.S., Shivashankar, "Conceptual level similarity measure based review spam detection", Signal and Image Processing (ICSIP), 2010 International Conference on 2010

[12] Ghose, A.Ipeirotis, P.G. ,"Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics", Knowledge and Data Engineering, IEEE Transactions on 23, 2011

[13] Wenying Zheng, QiangYe, "Sentiment Classification of Chinese Traveler Reviews by Support Vector Machine Algorithm", Intelligent Information Technology Application, 2009. IITA 2009. Third International Symposium on32009

[14] Hai Wang, Shou hong Wang, "A Purchasing Sequences Data Mining Method for Customer Segmentation" Service Operations and Logistics, and Informatics, 2006. SOLI '06. IEEE International Conference on 2006

[15] Rahayu, D.A., Krishnaswamy S., Alahakoon O., Labbe C., "RnR: Extracting Rationale from Online Reviews and Ratings", Data Mining Workshops (ICDMW), 2010 IEEE International Conference on 2010

[16] S.Loster, M.Lofi, C. Balke, "Will I Like It? Providing Product Overviews Based on Opinion Excerpts Homoceanu", W-T Commerce and Enterprise Computing (CEC), 2011 IEEE 13th Conference on 2011.

[17] Li Shi, Luo Siqing ,"Improving the performance of features extraction from Chinese customer reviews" Communication Systems, Networks and Applications (ICCSNA), 2010 Second International Conference on 2010.

[18] Lingyan Ji, Hanxiao Shi, Mengli Li, Meng xia Cai, Peiqi Feng ,"Opinion mining of product reviews based on semantic role labeling", Computer Science and Education (ICCSE), 2010 5th International Conference on 2010.

[19] Bo Pang, Lillian Lee, "Opinion Mining and Sentiment Analysis", ACM Journal**,** Foundations and Trends in Information Retrieval, Volume 2 Issue 1-2, January, 2008.

[20] Minqing Hu, Bing Liu," Mining and summarizing customer reviews", Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004

[21] Nihalahmad Shikalgar, Deepak Badgujar, "Online Review Mining for forecasting sales". In International Journal for research in Engineering & Technologies (IJRET), Volume 2 issue 12, December 2013.

[22] Nihalahmad Shikalgar, Arati Dixit, "Clustering & Regression based Sentiment Prediction for Online data. In Third post-graduation conference for Computer Engineering, held at MCOERC, Nashik, March 2014.

[23] Nihalahmad Shikalgar, Arati Dixit, "Prediction of Appraisal groups for movie review analysis". International Journal of Science & research (IJSR), Volume 3 issue 6, June 2014.

[24] http://www.rottentomatoes.com, accessed on dated 15 January, 2014.