IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified

Vol. 6, Issue 4, April 2017

Collaborative Approach based Restaurant Recommender System using Naive Bayes

Prof N G Bhojne[1], Sagar Deore[2], Rushikesh Jagtap[3], Gaurav Jain[4], Chirag Kalal[5]

Assistant Professor, Department of Computer Engineering, Sinhgad College of Engineering, Pune, India[1]

Student, Department of Computer Engineering, Sinhgad College of Engineering, Pune, India[2,3,4,5]

Abstract: In the past decades, people have gained a wide range of options as the availability of information expands. To help them make decisions, recommendation systems play an important role in all kinds of aspects, e.g. news, books, movies and so on. One such aspect is Restaurant where recommendations can be provided using user attributes and past activity. A noticeable similarity is found in people belonging to same categories based on attributes like age, native place, gender, work-type, etc. Using Collaborative approach these attributes of individuals can be analysed. Also the reviews and ratings given by customers to a restaurant play an important role in selection of an ideal restaurant. In this paper, we follow an approach based on the Simple Bayesian Classifier and apply it to user-based variant of the collaborative filtering, which makes predictions based on the user similarities. The recommended results are further refined by the review/rating analysis of individual restaurants using Text Mining. The review/rating analysis of predicted restaurants help to assess the current overall user experience of those restaurants which include the quality of food served, service, cost, ambience, etc. Our approach comprises counting positive and negative term scores to determine sentiment orientation, using Sentiment Analysis (SentiWordNet library). Finally more relevant results with positive reviews can be obtained which are passed as output recommendations to customers. In future we can also add content based filtering to recommend restaurant on the basis of the characteristics like dinning arrangement, facilities, working hours, etc. of restaurants that the particular user have already visited. By making hybrid of both content and collaborative we can increase the quality of recommendation result.

Keywords: Recommender System, Collaborative filtering, Naive Bayes, SentiWordNet.

I. INTRODUCTION

In today's world we have a large amount of information available about almost each and every item around us. The data generation rate is exponential and thereby now we face the problem of excessive information dump. We have search engine giants like Google, Yahoo, etc. which provides you with all the available data but mostly the data of actual use usually gets lost in the large data which we receive. For Example consider a person visiting a certain place wish to dine there. So the person searches for restaurants in that area on Google for which the Google provides him with the list of all the available restaurants in that specified area sorted maybe with respect to distance from the user. Here the user dining preferences or likes are not considered and all the available data is directly dumped on the user. Some restaurant specific sites like Zomato and others which may provide certain additional filters on basic attributes like price and cuisine, they however fail to consider user preferences and his dining patterns.An effective solution to this is a Recommendation Systems. Recommender systems are information filtering systems that deal with the problem of in-formation overload by filtering vital information fragment out of large amount of dynamically generated information according to user's preferences, interest, or observed behaviour about item. Recommender system has the ability to predict whether a particular user would prefer an item or not based on the user's profile. There are 3 types of Recommender system content-based recommender, collaborative recommender and hybrid based recommender [1]. In our system we have used collaborative based system. Recommender systems typically produce a list of recommendations in one of two ways through collaborative and content-based filtering approach. Collaborative filtering approaches building a model from a user's past behaviour (items previously purchased or selected and/or numerical ratings given to those items) as well as similar decisions made by other users. This model is then used to predict items (or ratings for items) that the user may have an interest in.Content-based filtering approaches utilize a series of discrete characteristics of an item in order to recommend additional items with similar properties. Combination of these two systems is Hybrid system. This paper presents restaurant recommender system that extracts user interests for food or restaurants and then makes recommendation accordingly. When searching for restaurants information and making decisions on where to eat, people rely on the review sites. But it is possible that the highly rated ones do not align with individual's tastes. Different people have different food preferences and dietary restrictions. A noticeable similarity is found in people belonging to same categories based on attributes like age, native place, gender, work-type, etc. which can help to relate individuals and their dietary preferences.A collaborative

DOI10.17148/IJARCCE.2017.6402

approach based recommendation system can be designed to provide restaurant recommendations to people based on their preferences andpast history. The purposed algorithm for Restaurants classification is Naive Bayes Classifier. Naive Bayes Classifier is statistical classifier which can predict class membership probabilities such as the probability that a given sample will belong to a particular case [2].

II. RELATED WORK

Recommender systems help users deal with data overload by recommending to them items that they would like. There has been a lot of work done on designing recommender systems during the last two decades. Amazon.com [3] and Netflix [4] are two popular applications of recommender systems. [5] Presents an online social network-based recommender system that extracts user's interests for jobs and then makes recommendations to them accordingly. It is focused on two very popular social networks Facebook and LinkedIn. [6] Implements Naive Bayes to retrieve hidden data from stored database and compares the user values with trained data set. Then mapping of patient's attributes with stored database entries is done and probabilistic values are analysed for decision making. It can answer complex queries for diagnosing heart disease and thus assist healthcare practitioners to make intelligent clinical decisions which traditional decision support systems cannot. Sentiment analysis or opinion mining, an imperative research area of natural language processing, involves the extraction and identification of the attitude of a speaker or writer about a certain subject matter [7]. Opinion is generally combination of words, sentences, or documents. Opinion mining is based on the reviews of the other users. Sentiment analysis is used to classify each opinion as positive or negative.[8] Research paper proposed a novel document quality classification approach, which extracts sentiment value from SentiWordNet and accumulates the different sentimental influence of each word based on a document level. According to the experimental results, this proposed approach, which extract sentimental knowledge from SentiWordNet, outperform the approach in which SentiWordNet is not used for all categories with an exception, which is spam category. [9] Proposed system uses SentiWordNet library. The data from the reviews first removing stop words, then stemming by Porter Stemmer algorithm and then that reviews are tagged by their respective parts of speech. Then the score of review is calculated by pair of part of speech and rank in SentiWordNet.

III.DATA GATHERING

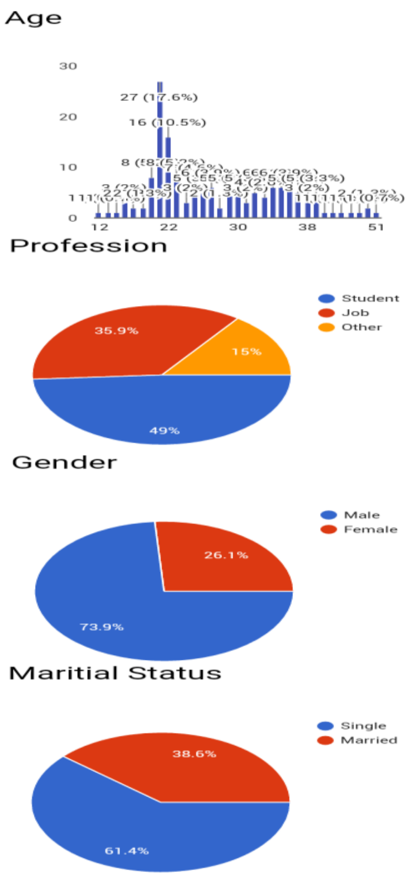

Fig. 1.Survey Data Screenshot

DOI10.17148/IJARCCE.2017.6402

At a time when really knowing customer preference is crucial to many successful business operations, data mining should be at the forefront of technology toolbox. Customer attraction, retention and prediction are important marketing concepts in the most of industries and central components of data mining. Thus these concepts can be smartly used for Restaurant Industry. Foodservice establishments have long known there is a need to exceed customer expectations in order to stimulate current sales while creating the opportunity for repeat business. Also available customer data can help analyse customer food item relations if any and help to target certain group of customers. The data mining process is designed to identify relationships, patterns and trends that may be present among data, but are not obvious.

The data mining process is intended to turn data into information and information into insight. The mining process demands large data sets to train the available algorithms. Such data requirements are fulfilled by conducting survey through different survey forms, websites, etc. The similarities between customers are analyzed on the basis of attributes like their age group, marital status, home town, price range, personal preferences, and past dining history. Hence the forms were designed so as to effectively gather information about these attributes from different customers visiting restaurants at different locations. Also text analysis of the reviews received for restaurants is done to get a real time idea about the overall current experience of restaurants which include the quality of food served, service, cost, ambience, etc. So an effort is made to collect some initial reviews about restaurants, but the most vital source of reviews are those generated through user application interaction after project deployment. The screenshots of survey form analysis is shown below:

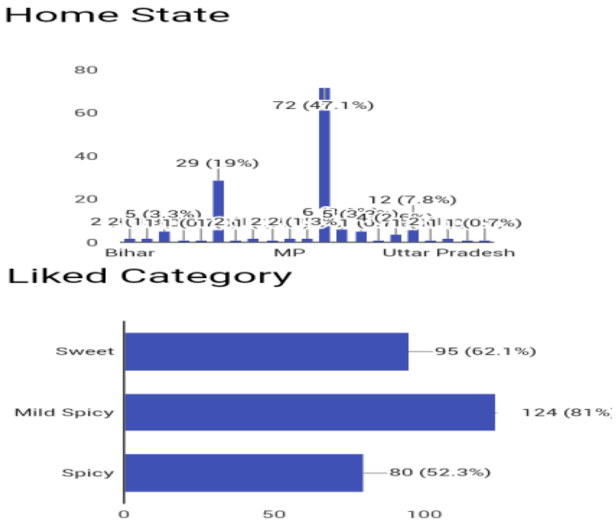

Fig. 2. Survey Data Screenshot

IV. PROPOSED SYSTEM

The proposed system aims at presenting the end users a personalized restaurant recommendation list with help of Data Mining and Text Mining techniques.

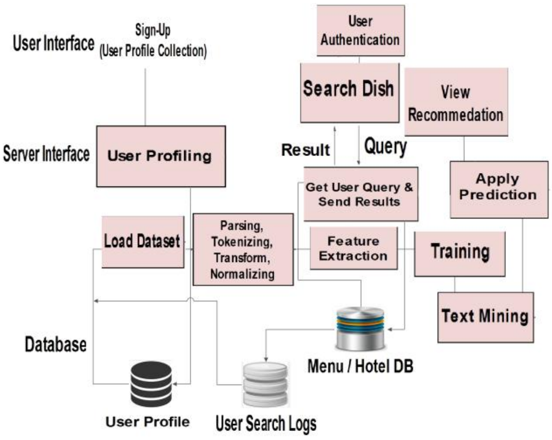

Fig. 3. Architecture Diagram

DOI10.17148/IJARCCE.2017.6402

The end user interacts with the system through a mobile application wherein he can register himself on application and request for recommendations. The Data Mining and Text Mining algorithms run at a centralized server. The server also holds the training dataset gathered for training the Mining algorithms. The system provides separate Admin and Restaurant manager modules through which the system administrators and actual restaurant owners can communicate with system. The following figure shows basic system architecture.

4.1 Naive Bayes

A **Bayes classifier** is a simple probabilistic classifier based on applying Bayes theorem with strong independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model". In simple terms, a naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 4" in diameter. Even if these features depend on each other or upon the existence of the other features, a naive Bayes classifier considers all of these properties to independently contribute to the probability that this fruit is an apple. Depending on the precise nature of the probability model, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood; in other words, one can work with the naive Bayes model without believing inBayesian probability or using any Bayesian methods. In spite of their naive design and apparently over-simplified assumptions, naive Bayes classifiers have worked quitewell in many complex real-world situations. Analysis of the Bayesian classification problem has shown that there are some theoretical reasons for the apparently unreasonable efficacy of naive Bayes classifiers [10]. Still, a comprehensive comparison with other classification methods showed that Bayes classification is outperformed by more current approaches, such as boosted trees or random forests [11]. An advantage of the naive Bayes classifier is that it requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix.

4.1.1 Bayes Theorem

Bayes' theorem is stated mathematically as the following equation.
whereA and B are events and $P(B) \neq 0$.

- P(A) and P(B) are the probabilities of observing A and B without regard to each other.
- P(A | B), a conditional probability, is the probability of observing event A given that B is true.
- P(B | A) is the probability of observing event B given that A is true.

4.1.2 Naive Bayes Probabilistic Model

Abstractly, the probability model for a classifier is a conditional model$P(C|F_1,...,F_n)$over a dependent class variable with a small number of outcomes or classes, conditional on several feature variables through . The problem is that if the number of features is large or when a feature can take on a large number of values, then basing such a model on probability tables is infeasible. We therefore reformulate the model to make it more tractable
Using Bayes' theorem, we write

$$P(C|F_1,...,F_n) = \frac{P(C)P(F1,..,Fn|C)}{P(F1,...,Fn)}$$

In practice we are only interested in the numerator of that fraction, since the denominator does not depend on and the values of the features are given, so that the denominator is effectively constant. The numerator is equivalent to the joint probability modeli.e,$P(C|F_1,...,F_n)$using repeated applications of the definition of conditional probability this can also be expressed as

$$P(C|F_1,...,F_n) = P(C)\ P(F_1|C)\ P(F_2|C)\ P(F_3|C)\ ...$$

This means that under the above independence assumptions, the conditional distribution over the class variable can be expressed like this

$$P(C|F_1,...,F_n) = \frac{1}{Z}\ P(C)\prod_{i=1}^{n} P(F_i|C)$$

where (the evidence) is a scaling factor dependent only on
$F_1,\ldots,F_n$ i.e. a constant if the values of the feature variables are known.

4.1.3Parameter Estimation

All model parameters (i.e., class priors and feature probability distributions) can be approximated with relative frequencies from the training set. These are maximum likelihood estimates of the probabilities. A class' prior may be calculated by assuming equiprobable classes (i.e., priors = 1 / (number of classes)), or by calculating an estimate for the class probability from the training set (i.e., (prior for a given class) = (number of samples in the class) / (total number of

DOI10.17148/IJARCCE.2017.6402

samples)).To estimate the parameters for a feature's distribution, one must assume a distribution or generate nonparametric models for the features from the training set. If one is dealing with continuous data, a typical assumption is that the continuous values associated with each class are distributed according to a Gaussian distribution.
Consider the case where user profiles are analyzed to suggest restaurant recommendations to other similar profiles. A noticeable similarity is found in people belonging to same home town, to same age group etc. Such attributes thus can be used to define a user profile which can be further used as parameters for Naive Bayes. Data can be maintained with a set of user attributes along with his opinion about restaurants they visited, i.e. they like it or dislike it (yes/no). The selection of user attributes to be selected as parameters plays the vital role in the accuracy of the results obtained. Survey needs to be carried and users profiles studied in deciding which attributes to be selected.
The user attributes considered are:

- Age
- Gender
- Price range
- Home town
- Marital status
- Preferred food category/type

4.1.4 Sample Correction

If given class and feature value never occurs together in the training set then the frequency-based probability estimate will be zero. This is problematic since it will wipe out all information in the other probabilities when they are multiplied. It is therefore often desirable to incorporate a small-sample correction in all probability estimates such that no probability is ever set to be exactly zero**.**

4.1.5 Example

Consider the following tables describing sample data about attributes of some customers and their likes/dislikes.

TABLE I Hotel Good Luck

Gender	Age Group	Preference	Origin	price	Marital status	Feedback
M	Young	Mild Spicy	Western	average	single	Dislike
M	Old	Sweet	Western	high	single	Like
F	Young	Spicy	Southern	low	single	Like
M	Middle Age	Spicy	Western	low	single	Dislike
M	Middle Age	Sweet	Eastern	average	single	Like
M	Old	Spicy	Western	high	married	Like
F	Young	Sweet	Eastern	low	single	Dislike
M	Young	Mild Spicy	Southern	average	single	Like
F	Middle Age	Spicy	Northern	high	married	Dislike
F	Old	Sweet	Eastern	average	married	Dislike
F	Old	Sweet	Northern	low	married	Like
M	Middle Age	Spicy	Southern	high	single	Like

TABLE II Restaurant Vaishali

Gender	Age Group	Preference	Origin	price	Marital status	Feedback
M	Young	Mild Spicy	Northern	Low	single	Like
F	Old	Sweet	Western	average	married	Like
F	Young	Sweet	Southern	High	single	Like
M	Middle Age	Spicy	Southern	average	single	Dislike
F	Middle Age	Mild Spicy	Eastern	High	married	Dislike
M	Young	Spicy	Western	Low	single	Like
M	Old	Sweet	Eastern	average	married	Like

M	Young	Mild Spicy	Northern	Low	single	Dislike
F	Middle Age	Mild Spicy	Northern	High	single	Dislike
M	Old	Sweet	Southern	average	married	Like
F	Old	Mild Spicy	Western	High	single	Dislike
F	Young	Spicy	Southern	Low	single	Dislike

TABLE III Restaurant Sukanta

Gender	Age Group	Preference	Origin	Price	Marital status	Feedback
F	Young	Mild Spicy	Western	High	single	Dislike
M	Old	Sweet	Eastern	Low	married	Dislike
F	Old	Sweet	Southern	Average	married	Like
F	Middle	Spicy	Western	High	single	Dislike
M	Middle Age	Sweet	Eastern	High	married	Like
M	Old	Sweet	Northern	Average	married	Like
M	Young	Mild Spicy	Eastern	Low	single	Like
M	Old	Mild Spicy	Northern	Low	married	Like
F	Middle Age	Spicy	Northern	Average	single	Dislike
F	Young	Sweet	Eastern	High	single	Dislike
M	Old	Sweet	Western	High	married	Like
M	Middle Age	Spicy	Southern	High	single	Dislike

Now we have a person with the attribute tuple as:
C= {age: young, gender: M, preference: spicy, origin: southern, price range: average, marital status: single}
We want to determine the posterior for which restaurant(Good Luck/Vaishali/Sukanta) is greater.
posterior (Good Luck) = P(like|C)
=P(like | young)
* P(like | M)
*P(like | spicy)
*P(like | southern)
*P(like | average)
*P(like | single)
=3/12 *4/12 * 1/12 * 2/12 * 3/12
=0.00028935
posterior (Vaishali) = P(like|C)
=P(like | young)
* P(like | M)
*P(like | spicy)
*P(like | southern)
*P(like | average)
*P(like | single)
=5/12 * 2/12 *3/12 * 3/12 * 2/12
* 5/12
=0.00361689

posterior (Good Luck) = P(like|C)
=P(like | young)
* P(like | M)
*P(like | spicy)
*P(like | southern)
*P(like | average)
*P(like | single)
=5/12 * 1/12 * 1/12 * 1/12 * 2/12 * 1/12
=4.774783 * e^{-6}

Since the posterior numerator (Vaishali) > posterior numerator (Good Luck)> posterior numerator(Sukanta), Restaurant Vaishali can be recommended to the user with profile C.

4.2 Review Analysis

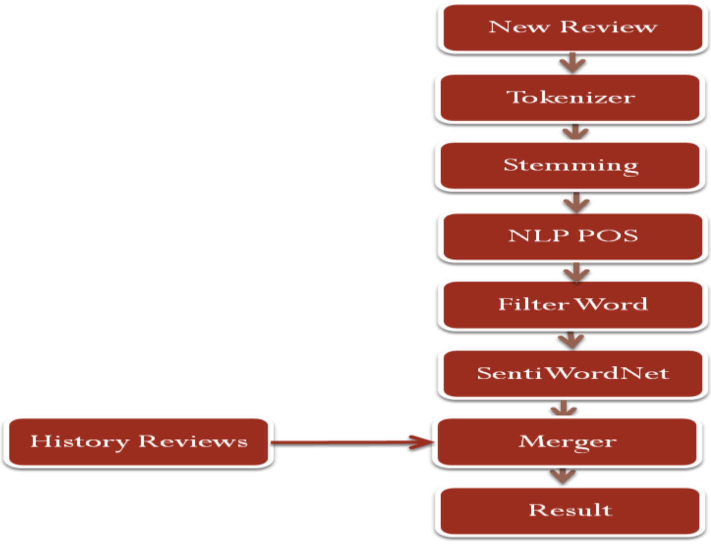

Fig. 4. Review Analysis Steps

4.2.1Tokenizer

This will convert the string of words i.e., a sentence in list of tokens. It will separate out each and every word by using any word separator, that are“ (space)”, “.(full stop)” ,“(comma)” Etc.

4.2.2Stemming or Slang Identification

Slangs are shorthand words that are used in informal texts in order to reduce the length of text. We have replaced slangs with complete words in order to perform efficient sentiment scoring and classification. Slang dictionary is used to find slangs and its definition and then it is replaced such words with repeated letters. For example we use gr8 for great.

4.2.3Part of Speech Tagger

SentiWordNet provides synset mutual information scores based on the part of speech tag of each term. Therefore, it is necessary to perform POS tagging.The part of speech (POS) can be one of the following:

- Adjective ‘a’
- Verb ‘v’
- Adverb ‘r’
- Noun ‘n’

4.2.4Filter Word

We need to filter out common words which have less or no meaning. Stop words are common words of a language, which tend to have little meaning. For example is, am, are, and, before, while, etc. are stop words and they need to be filtered out because there is not meaning or sentiment attach to those words.

4.2.5SentiWordNet

SentiWordNet is a publicly available lexical resource. It provides information about polarity identification as well as for subjectivity detection. We are using SentiWordNet 3.0. Each synset in SentiWordNet 3.0 is uniquely identified by ‘POS & Synset#rank’ pair.

TABLE IV Sample of SentiWordNet

POS	PosScore	NegScore	Synset#rank
V	0.25	0.125	Blaze#3
A	0.5	0.125	Living#3
R	0.625	0	Mordaciously#1

There are three types of sentiment scores: positive, negative and objective. The positive and negative scores are represented by ‘PosScore’ and ‘NegScore’ whereas the objectives score ‘ObjScore’ is calculated by the equation.

$$\text{ObjScore} = 1 - (\text{PosScore} + \text{NegScore})$$

The synset score is calculated as

$$\text{Synset} = \text{PosScore} - \text{NegScore}$$

The synsets are weighted according to their usage ranks and the final score for each term is calculated by the equation.

$$\text{Score} = \sum_1^n \text{synsetScore}(r)/r$$

Where r is the rank of the synset.

V. CONCLUSION

Decision Support System for restaurant selection is developed using Naive Bayesian Classification technique. The system extractshidden knowledge from a historical restaurant customer database. This can prove to be an effective model to predict restaurants mostly likely to be liked by users. Also the review analysis of the customer reviews is done to keep track of the recent performance of restaurants. The reviews are analyzed with the help of SentiWordNet library which help in accessing positive/negative comments and reviews and furthermore asserts a score to the restaurants. This scores help to rank the most likely restaurants given by the Naive Bayes on the basic of their recent quality. Thus a user attribute based restaurant recommendation system is proposed using Data Mining (Naive Bayes) and Text Mining (SentiWordNet Analysis) techniques.

ACKNOWLEDGMENT

We would like to express our sincere gratitude towards our guide **Prof. N. G. Bhojne** for his invaluable guidance and supervision that helped us in our research. He has always encouraged us to explore new concepts and pursue newer research problems. We credit our project contribution to him. Collectively, we would also like to thank our project committee members **Prof. G. G. Chiddarwar** and **Prof. S. D. Wable** for their time, suggestions, and always making themselves available. We cannot thank them enough.

REFERENCES

[1] GediminasAdomavicius and Alexander Tuzhilin, “Toward the Next Generation of Recommender Systems: A Survey of the State-of-Art and Possible Extensions”,IEEE transactions on knowledge and data engineering, Vol. 17.

[2] Shweta Kharya and SunitaSoni, “Weighted Naïve Bayes Classifier: A Predictive Model for Breast Cancer Detection”, International Journal of Computer Applications, Vol. 133, No.9, Jan 2016.

[3] G. Linden, B. Smith, and J. York, “Amazon.com recommendation: Item-to-item collaborative filtering”, IEEE Internet Computing, vol. 7, no. 1, pp. 76-80, Jan. 2003.

[4] J. Bennett, S. Lanning, and N. Netflix, “The Netflix prize”, in In KDD Cup and Workshop in conjunction with KDD, 2007.

[5] MamadouDiaby, Emmanuel Viennet, Tristan Launay, “Toward the Next Generation of Recruitment Tools: An Online Social Network-based Job Recommender System”, 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.

[6] Shadab Adam Pattekari and AsmaParveen, “Prediction system for heart disease using naive bayes”, International Journal of Advanced Computer and Mathematical Sciences ISSN 2230-9624, vol. 3, Issue 3, 2012.

[7] Farhan Hassan Khan, Usman Qamar, Saba Bashir, “SentiMI: Introducing point-wise mutual information with SentiWordNet to improve sentiment polarity detection”, Applied Soft Computing, vol.39, Nov. 2015.

[8] Chihli Hung, Chih-Fong Tsai, Hsinyi Huang, “Extracting Word-of-Mouth Sentiments via SentiWordNet for document quality classification”, Bentham Science Publishers, 2012.

[9] Shoiab Ahmed and AjitDanti, “A Novel Approach for Sentimental Analysis and Opinion Mining based on SentiWordNet using Web Data”.

[10] Harry Zhan,”The Optimality of Naive Bayes”,Faculty of Computer Science,University of New Brunswick, Fredericton, New Brunswick, Canada.

[11] Rich Caruana, Niculescu-Mizil, “An Empirical Comparison of Supervise Learning Algorithms”, Department of Computer Science, Cornell, University, Ithaca, NY 14853, USA

DOI10.17148/IJARCCE.2017.6402