© 2016 IJSRSET | Volume 2 | Issue 6 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

An Enhanced Extraction and Summarization Technique with User Review Data for Product Recommendation to Customers

Vikas B O[1], Dr. Jitendranath Mungara[2]

[1]Assistant Professor, Department of IS & E, NHCE, Bangalore, India
[2]Professor & HOD, Department of IS & E, NHCE, Bangalore, India

ABSTRACT

In the current world, recommendation system plays a major role in helping consumer find for relevant product information with summarized reviews. It contains collection of opinions, reviews, recommendation, ratings, comment and personal experience shared by different user review on a product through social networks, e-commerce websites, blogs and forums. These reviews become an opinion for consumers to learn different aspects of products like limitations, advantages, features, services and suppliers. Various methods of evaluating products/services to consumer are provided from different review sites. The proposed method extracts reviews and summarises to provide enhanced product recommendation to the consumer.

Keywords: Product recommendation, Opinion mining, Sentiment analysis, Natural Language Processing, POS tagging

I. INTRODUCTION

In this progressive world, collection of data is enormous and evolving with changing times. Each product has two types of contents. Objective content is actual data is about service or product that a business production provided to consumer. Subjective content is a collection of ratings, comments, opinions, recommendations and reviews shared by different users which are extracted from forums, blogs and social networks [1]. These reviews become an opinion for consumer to learn and share different aspects like limitations, features, merits, services and suppliers of a product. Due to this, the business of a product and consumer is related that is a very important feature for product recommendation system. These review data can be spread to larger audience with the help of online peer to peer communication that can provide awareness regarding a product or services. Large number of review recommendation systems has been designed that provide different review on a product to consumers. These include comments, numerical ratings, like and dislike, thumbs up and down to help ease of experience with product with other consumers [7].

Background

Most of the consumers in current world make use of Internet as communication for finding products as well as finding solutions for service/product related problems such as pre-purchase inquiries on a product and post-purchase services on product in real time. Numerous aspects such as reviews, mining of data, review recommendation and providing ease of facility for decision making process will be invaluable to consumers. Mining on reviews provided by users will involve extraction of review, cleaning up of data, rating the product, analyse rating quantitatively, analyse reviews through sentiment analysis, opinion mining and provide a score for a product that will help consumer feel ease whole selecting a product and differentiate product based on the user reviews [3].

Several researches have been developed, various recommendation systems that can extract reviews, mine the reviews and provide recommendations. Most common method is to summarize reviews is using the concept of sentiment analysis [3]. Sentiment analysis will classify the different of a given text at the sentence, document and express it in opinion such as positive, negative or neutral.

Advanced methods of sentiment analysis will search for emotional states such as "happy", "sad" and "angry". This methodology works at document level. A different method of using sentiment analysis (positive, negative and neutral) is to make use of scaling system where words are commonly associated with them having positive, negative and neutral sentiments with number on a scale of 1 to 5 (negative to positive) and when an input text as a review is given the text is analyzed using natural language processing, next the understanding of these words and how they relate to concept of sentiment analysis. Each word is given a score, which allows understanding of sentiment based analysis on a 5 point scale [5]. Thus, usage of review extraction and summarization technique of reviews becomes most important for a product to sell with respect to business and important for a consumer to select the right product.

II. METHODS AND MATERIAL

1. Related Works

The studies carried out on the prior works include the following method which deals with opinion mining and sentiment analysis on online reviews.

A. Studies Related with Opinion Mining and Sentiment Analysis On Online Reviews [2]

User reviews will provide well informed decisions for a service/product; also help the product production team to underhand problems such as product drawbacks, product weaknesses. The online review data information is very vast and various text processing methods, mining on opinions, tools for automated extraction and natural language processing have been in literature. In this larger consumer and business audience, moving forward with tools and techniques for online product reviews set a pattern or new era of recommendation systems. Machine learning, web search methods, mining of text, information retrieval and natural language processing methodology will provide an enhanced method to find, extract and summarize sentiment, opinions and mind relevant data required from vast information of user-review context on the web.

Document-level Sentiment Analysis: Personalized text extraction from document with sentiment analysis is notation of document level sentiment analysis.

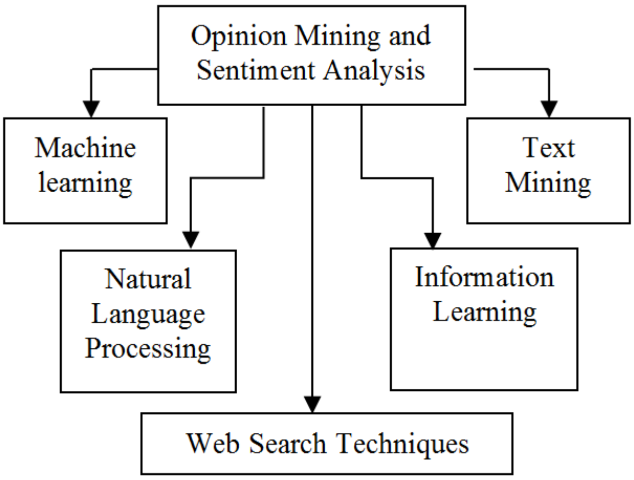


Figure 1. Opinion Mining and Sentiment Analysis [2]

B. Sentiment identification can be performed on two levels of granularity [3]

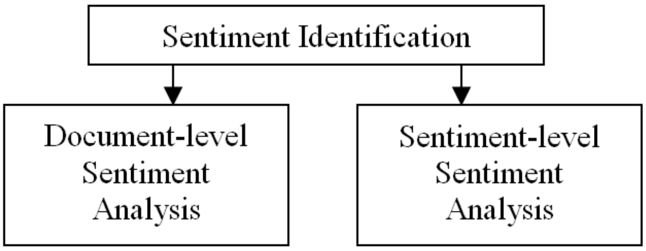

Sentiment Identification
Document-level Sentiment Analysis
Sentiment-level Sentiment Analysis

Figure 2. Sentiment Identification [3]

Sentence-level Sentiment Analysis: Sentence level sentiment analysis produces an inference that different sentences might possess different opinions about a product.

C. Specific context of opinion mining can be performed on two types of parameters [4]

Parameter such as inclination review, review via graph, word frequency review are categorized as objective parameters. While expressing an opinion on product and providing particular feature of product on emotions of reviewers are categorized as subjective parameters. Need of natural language processing for studying the subjective parameters and numerical computation for objective parameter will tell the opinion extracted from reviewer is negative or positive. NLP technique will extract all essential features of reviews.

In general, mining of user-generated reviews involve following steps [8]

i. Mechanized extraction of review element from e-commerce sites

ii. Text reviews, classify reviews as positive or negative depending upon adjectives with NLP technique

iii. Recognition and extraction of product characteristics from the text reviews

iv. Specific product characteristics opinion extraction

v. Arrive at enhanced product score with combination of subjective and objective parameters.

The outcome of the review problems identified are classified according to the techniques used based on the functionality of different mining techniques for review recommendation.

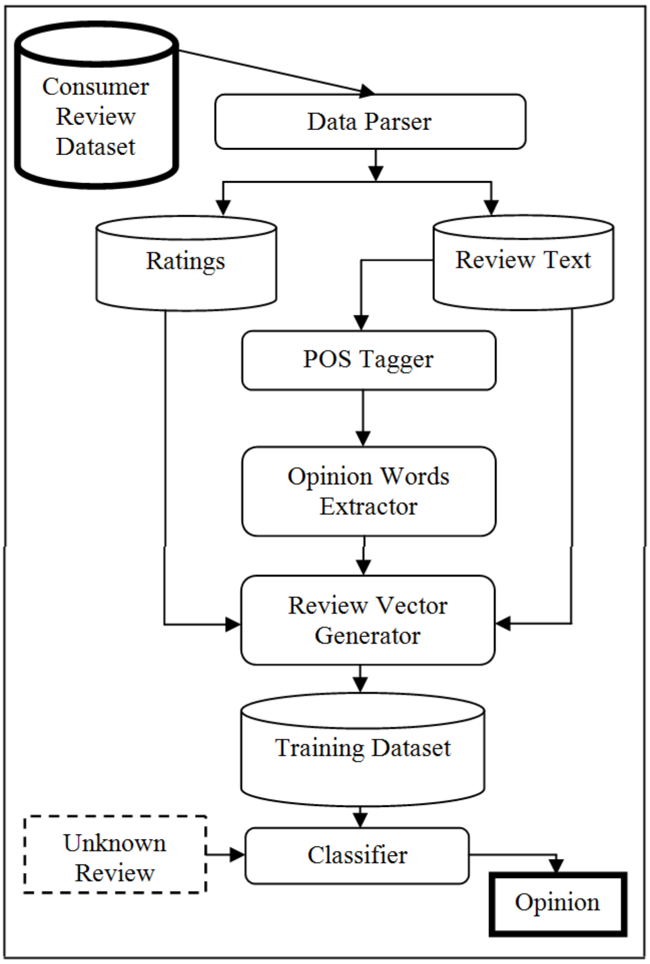


Fig. 3 System Architecture

2. Implementation

The implementation consists of the four modules which are to be implemented.

Table 1. Functionality of Review Mining Algorithms

Techniques used	Functionality
K-Nearest Neighbor Classifier	Estimate the nearest neighbor values of categories of a given text document, extract words and provide category.
Naive Bayes Classifier	It takes input data /reviews and performs decision into two classes.
Classification by Association Rules	Sentiment classification is done by constructing a text classifies by extracting category of words by association rules.
POS Tagging	POS Tag will consider only opinion sentence containing at least one detected sentiment phrase for opinion extraction.

III.METHODOLOGY

The enhanced extraction and summarization technique involves following stages to provide product recommendations

A. User Reviews: Product or service made by user who has obtained the product/service is user review. It is a form of response on online shopping sites and purchases made by user. These reviews response in text are extracted and applied with POS tagging that person processing.

B. POS Tagging: Lexical set or part-of-speech(POST) are used to find syntactic words in any document of text like adjective, verb, noun, synonym etc. This processing can be done with the basis of definition for example nouns are names like Bangalore, or on the similar or adjacent words depending upon relationships among the words extracted. POS tags are also called as word classes. The use of POS tags is usually for performing an extraction of specific sentence or words. Extraction of specific words and passes it to NLP processor is performed with the help of POS tagging.

C. Apply NLP Rules: An interaction between human language and computers is the Felipe of artificial intelligence and computer science is called natural language processing. A relationship between human-computer interactions is NLP. The process of authorizing computer to fetch meaning from human language input for natural language understanding can be done using NLP. Hence, various rules can be defined to extract only sentiment, opinion- oriented, emotion-based text from a set of review information data.

D. Find Opinion Words: Opinion words are classified as emotions, facts, statement, expressions, point of view, judgment, attitude, and way of thoughts on any specific product. NLP rule mining can be applied to extract these opinion words that lead for an enhanced way of finding s specific opinion words on product review data information.

E. Summarized Review Result: Process of fragmenting a complex data into smaller parts to obtain a better understanding of the review information is called summarized review. As NLP rules are applied to obtain opinion words will be further summaries for product recommendation at ease.

For this the above mentioned modules are utilized to provide review recommendation in the following two ways [1].

a. Single Review- Sentence Level Analysis.
b. Document Level Analysis.

MODULE 1: POS TAGGING

Objective : To apply the POS tag-set for the datasets.
Input : Datasets.
Output: Classify the tag-set according to the POS.

Algorithm

Step 1	Input a string of words
Step 2	A set of specified tag-set would be created automatically.
Step 3	String of words is compared with specified tag-set.
Step 4	The corresponding words present in the tag-set are extracted.
Step 5	The POS Tagging words are classified.

Working Procedure:
Input:

Quality hotel at great price Very clean. Free breakfast with good selections. Staff friendly and most helpful.

Output:

Quality/NNP hotel/NN at/IN great/JJ price/NN Very/RB clean/JJ ./. Free/NNP breakfast/NN with/IN good/JJ selections/NNS/. Staff/NN friendly/JJ and/CC most/RBS helpful/JJ.

MODULE 2: NLP RULES

Objective: To apply the NLP rules for the datasets.
Input: Datasets.

Output: To derive meaning from human or natural language input.

Algorithm

Step 1	For a set of input words tokenization is done for the particular tag set using NLP.
Step 2	Morphological and semantic tagging is done for the input word which are read from the tag set.
Step 3	Extracting the opinion words from the pattern set using NLP.
Step 4	Producing the canonical form of the words and using lemma.
Step 5	The NLP rules are applied to get the desired output

MODULE 3: FIND OPINION WORDS

Objective: To find the opinion words
Input: Output of NLP words
Output: Opinion Words.

Working Procedure:

Input:

Quality hotel at great price Very clean. Free breakfast with good selections. Staff friendly and most helpful.

Output:

great/JJ clean/JJ good/JJ friendly/JJ helpful/JJ.

MODULE 4: SUMMARIZED REVIEW RESULT

Summarized Review Result provides review recommendation in the following two ways.

a. Single Review- Sentence Level Analysis.
b. Document Level Analysis.

Working Procedure:

In this module after the training data is classified, evaluation of the testing data happens in cross validation with training data for the accuracy of the modules. The accuracy of the modules are measured using Confusion matrix to determine the factors like Precision, Recall and F-Measure.

Table 2: Confusion Matrix

PREDICTED

	Yes	No
Yes	TP	FP
No	FN	TN

i) Confusion Matrix:

A confusion matrix is a table that is often used to describe the performance of a classification model or "classifier" on a set of test data for which the true values are known.

a. TP Rate (True Positive): Where actual is positive and predicted is also positive.
b. FP Rate (False Positive): Where actual is Negative but predicted is Positive.
c. FN Rate (False Negative): Where actual is Positive but predicted is Negative.
d. TN Rate (True Negative): Where actual is Negative and predicted is also Negative.

The factors Precision, Recall and F-measure are evaluated using the above conventions from the confusion matrix.

ii) Precision:

Precision is the harmonic mean of the True positive (TP) and the false positive (FP)

$$\boxed{\text{Precision}=\frac{TP}{TP+FP}} \quad \ldots\ldots\ldots (1)$$

iii) Recall:

Recall is the harmonic mean of the True positive (TP) and False negative (FN).

$$\boxed{\text{Recall} = \frac{TP}{TP+FN}} \quad \ldots\ldots\ldots (2)$$

iv) F-Measure:

A measure that combines precision and recall is the harmonic mean of precision and recall.

$$\boxed{\text{F-Measure} = 2 \text{ x } \frac{Precision\ x\ \ Recall}{Precision+Recall}} \quad \ldots. (3)$$

III. RESULTS AND DISCUSSION

A **TESTING : TESTING FOR SUMMARIZED REVIEW RESULTS**

Summarized Review Result is the process of breaking a complex topic into smaller parts in order to gain a better understanding of it.

A. Testing for Naïve Bayes Classifier

The corresponding tables refer to the details of the confusion matrix and the performance metrics for 20 datasets.

Table 3: Confusion Matrix for (20 Datasets)

A	B	Classified as
619	138	A=POSITIVE
115	308	B=NEGATIVE

Table 4: Performance Metrics for (20 Datasets)

TP RATE	FP RATE	FN RATE	PRE-CISION	RE CALL	F-MEASU RE	CLASS
0.818	0.272	0.182	0.843	0.818	0.830	POSITI VE
0.728	0.182	0.272	0.691	0.728	0.709	NEGATI VE
0.786	0.240	0.214	0.789	0.786	0.787	←WEIG HTED AVG

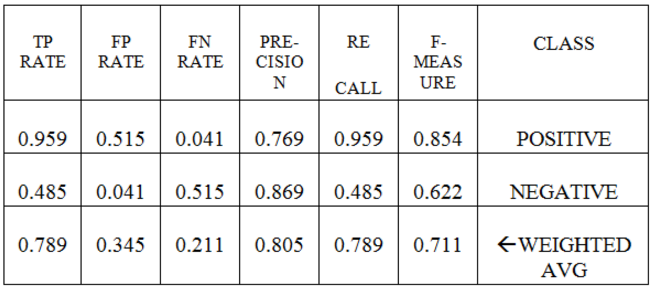
B. Testing for K-Nearest Neighbor Classifier

The corresponding tables refers to the details of the confusion matrix and the performance metrics for 20 datasets.

Table 5: Confusion Matrix for (20 Datasets)

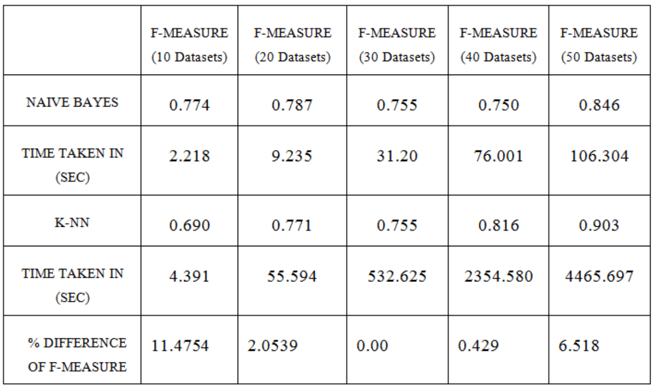
A	B	Classified as
726	31	A=POSITIVE
218	205	B=NEGATIVE

Table 6: Performance Metrics for (20 Datasets)



TP RATE	FP RATE	FN RATE	PRE-CISION	RECALL	F-MEASURE	CLASS
0.959	0.515	0.041	0.769	0.959	0.854	POSITIVE
0.485	0.041	0.515	0.869	0.485	0.622	NEGATIVE
0.789	0.345	0.211	0.805	0.789	0.711	←WEIGHTED AVG

Table 7: Comparison between NAÏVE BAYES and K-NN



	F-MEASURE (10 Datasets)	F-MEASURE (20 Datasets)	F-MEASURE (30 Datasets)	F-MEASURE (40 Datasets)	F-MEASURE (50 Datasets)
NAIVE BAYES	0.774	0.787	0.755	0.750	0.846
TIME TAKEN IN (SEC)	2.218	9.235	31.20	76.001	106.304
K-NN	0.690	0.771	0.755	0.816	0.903
TIME TAKEN IN (SEC)	4.391	55.594	532.625	2354.580	4465.697
% DIFFERENCE OF F-MEASURE	11.4754	2.0539	0.00	0.429	6.518

IV. CONCLUSION

The literature survey shows about opinion mining and user sentiments. With user reviews, findings and understanding sentiment is an emerging trend for which enhanced method has to be developed. Sentiment analysis categorize revues as positive or negative. The technique of POS tagging used for sentiment analysis by applying summaries review technique can provide clear differentiation between positive product and negative product. With the help of K-NN and Naive Bayes classification algorithms, demonstration and comparison of review recommendation system is formulated. The input review data is parsed and comparison between both algorithms as shown in table 7 shows that Naive Bayes algorithm is efficient compared to k-nn algorithm for proving an enhanced method for clear differentiation between positive and negative product respectively.

An enhanced approach for extraction and summarization technique proposed will produce efficient way of product recommendation to customers. Matching accuracy achieved with F-Measure by naive bayes algorithm is above 75% with varied number of datasets as input.

V. REFERENCES

[1] Venkata Rajeev P, Smrithi Rekha,V "Recommending Products to Customers using Opinion Mining of Online Product Reviews and Features" in International Conference on Circuit, Power and Computing Technologies ICCPCT],2015.

[2] Goyal, Ms Jayanti, and Ms Anjali Dadhich,"Procedure of Opinion Mining and Sentiment Analysis Techniques: A Survey." in machine learning, pp - 4-9, 2015.

[3] Vinodhini, G., and R. M. Chandrasekaran."Sentiment analysis and opinion mining: a survey." in International Journal 2.6 ,2012.

[4] Liu, Bing, Minqing Hu, and Junsheng Cheng. "Opinion observer: analyzing and comparing opinions on the web." in Proceedings of the 14th international conference on World Wide Web, pp - 342-351, 2012.

[5] Pak, Alexander, and Patrick Paroubek. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." In LREC, vol. 10, pp - 1320-1326, 2010.

[6] Pang, B., Lee, L., & Vaithyanathan, S. "Thumbs up?:Sentiment classification using machine learning techniques." in Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10, pp- 79-86, 2010.

[7] Pang, B., & Lee, L."Opinion mining and sentiment analysis. Foundations and Trends" in Information Retrieval, 2(1-2), pp- 1-135, 2012.

[8] Bakliwal, Akshat, Piyush Arora, Ankit Patil, and V. Verma. "Towards enhanced opinion classification using NLP techniques." In Proceedings of the 5th international joint conference on natural language processing (IJCNLP). Chiang Mai, Thailand, pp. 101-107. 2011

VIKAS B O is an Assistant Professor in the department of Information science and engineering, New Horizon College of Engineering, Bangalore, India. He has received his B.E. degree in Computer Science and Engineering in 2013 and M.Tech degree in Computer Science and Engineering in 2015. His research interest includes Information security, Data mining and Internet of things. He is a member of Computer society of India, Mumbai, India. He has published papers in national and international journals and proceedings.

Dr. Jitendranath Mungara is a Double Doctorate in Computer Science and System Engineering and Electronics. He is working as a Prof. & HOD of Information Science and Engineering department in NHCE Bangalore. He is pioneering in data mining and IoT core research. He has published 75 papers in International Journals and Conference. Obtained three doctorates under his guidance and currently guiding three Ph.D students in MANETS and IoT.