iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 **Issue:** IV **Month of publication:** April 2018

DOI: http://doi.org/10.22214/ijraset.2018.4678

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue IV, April 2018- Available at www.ijraset.com

Mining E-commerce Websites to Provide Efficient Methodology for Sentiment Analysis

Divyanshu Rai[1], Anubha Tripathi[2], Sumbul Siddiqui[3], Dr. Mahesh Pawar[4], Dr. Sachin Goyal[5]
[1, 2, 3, 4, 5] *Department Of Information Techonology, Uit Rgpv, Bhopal*

Abstract: ***The e-commerce industry is developing very rapidly and buying products online has become a trend because of its easy access, lower costs, availability of various options, latest technology available, comparison of products, cash on delivery option, and fast delivery system. Therefore many people tend to buy product online. However some websites tend to sell fake products at very low costs and people tend to buy it. So to buy a product online people's comments and feedbacks becomes essential. People often express their opinions about a product or service by posting reviews. In e-commerce websites, people usually make comments about product's properties, vendor's attitude and delivery information. Customers often tend to rely on such texts to know about the product. The provided information serves as an important reference for people who buy products from the website thus analysing large volume of online reviews available could produce useful actionable knowledge which makes it essential that this information should be available in systematic manner. The study of these opinions is referred to as sentiment analysis. We use these reviews to form opinion by deciding the subjectivity/objectivity and negative/positive attitude of the buyer. In this paper we propose a hybrid algorithm by combining various text classification and clustering technique.***

Index Terms- Sentimental analysis, opinion mining, web mining, naive bayes, decision tree, text classification, data mining, page content mining, search engine mining, hyperlink mining.

I. INTRODUCTION

World Wide Web (WWW) is a space of information on the internet which connects documents with one another with the help of hypertext links. Web has become an indispensable component of any organization. Huge amount of information can be extracted not only from the web but also from the communication of users and the web. Over the last decade this WWW has been flooded with information. Billions of people are storing and retrieving information from web daily. This causes network trafficking thereby resulting in delay of requested pages. Due to this evolution of WWW there is rapid increase in information volume. While this provides a wider range of options and services to the users, on the other hand it also causes some trouble to the same users in finding the right and interesting information from this sea of information. Simultaneously the e-commerce is also developing at a very high pace. The main reason behind the advancement of e-commerce is due to its various advantages like cost, variety, fast services etc. Therefore more people connote to online shopping. The quality of the product being uneven, the opinion of users expressed via their comments becomes an important source of information to judge the product's quality. Also when the organization wants to benefit by obtaining the opinion of public to market its product and identify new opportunities it needs to deal with an overwhelming number of available comments. Now the question arises how should we extract useful knowledge from the whole lot and how can we improve the rate of utilization of information without being lost in this bulk of data. Web mining has come up with a solution to this problem. A technique called semantic analysis which makes it possible to analyze large amount of data and extract the actual opinions expressed through them that help customers. In this paper we are going to mine the reviews of a product on an e-commerce web site by applying an unique combination of conventional techniques of classification and clustering of data.

A. Web Mining

Web mining can be referred to as the reformation and transformation of information obtained from the world wide web using data mining techniques. Mining is basically extraction of something useful and valuable from raw substances. It is used to discover patterns and latent information from world wide web. In web mining data can be obtained from the server system, client system, proxy servers, or extracted from a company's database. The data obtained can be of any type such as image, text, video, audio and meta data. Therefore, web mining simply manages scrapping huge and hyperlinked base of information having or possessing various properties. Web Mining taxonomy includes Web content mining(WCM), Web structure mining(WSM), and Web usage mining(WUM).

B. Web Content Mining (Wcm)

Web mining refers to the finding of important contents from the web. It can be text, images, audio, video. Web content mining comprise of finding resources from web, web documentation, clustering and information extraction from the web pages. The data available on internet can be of three types: Unstructured data, Semi structured data and Fully structured data. Mining in such texts is termed as text mining or data mining[8]. Symbolic knowledge extraction can be referred from[9]. A concept based knowledge discovery methods from web extracted texts is given in [10]. Here, concept extraction takes place instead of word or attribute value analyzation. Semi structured HTML pages which have hyperlinks are mined through hyperlink mining. In hyperlink mining a key role is played by supervised learning or classification [11]. For example like in an email, newsgroup management and maintaining web directories. Web content mining can be applied in two ways : (web page content mining) directly mining the document contents and (search engine mining) improving on content search of other stuff like search engines. It can be applied by using two approaches : IR approach [12] and DB approach [13].

C. Web Structure Mining (Wsm)

Web structure mining is the layout underlying the link structure of the web. Information is then extracted using these layouts. Good quality and high relevance is indicated by hyperlinks. The relationship between information linked web pages is identified using this tool. This link permits a web page to extract data relating to the problem statement directly to the linking web page from web site the content is actually posted on. The main objective is to correlate the underlying latent relationships between web pages. Thus, this can provide strategic results for marketing of a web page. In WSM inter document structure is mined using hyperlinks where as in WCM intra document structure mining takes place.

D. Web Usage Mining (Wum)

Web usage mining is motivated from large chunks of raw data of various search engine log files or any other log files to find some amusing patterns using the conventional data mining techniques . Information regarding every visit to the page is contained in web server logs. Thus, by interpreting the web usage log data, web mining systems are able to establish patterns and gather knowledge regarding system's usage characteristics and therefore, they get to know the interest of the users. The browsing behaviour at a web site helps us to discover patterns. Evaluation of the data used provides the needed information to the companies to deal effectively and efficiently with their customers. Developing strategies through this type of mining provides more efficiency to the databases by having easier access paths. Collaborative filtering also plays a key role in finding out other users with similar interests. It allows e-commerce web sites to group the users based on their similarities and dissimilarities. Association rule mining is used along with collaborative filtering techniques to develop a collaborative recommender system [14]. Classification rules, clusters, sequential patterns, and association rules are discovered through pattern mining [15].

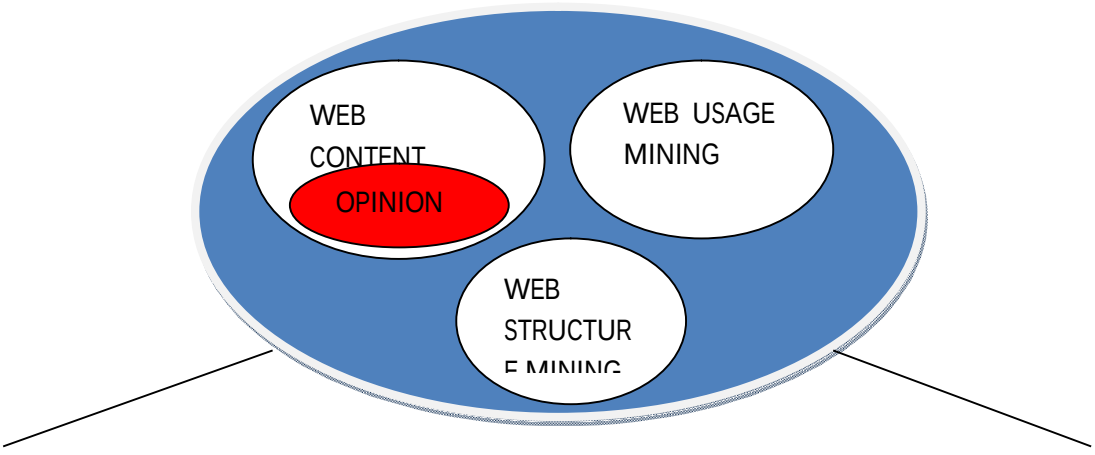


Fig. 1 WEB MINING

II. RELATED WORK

Gurneet Kaur et al in their paper titled "Sentimental analysis of flipkart using Naive Bayes and decision tree algorithm" presented a factual study showing the capability of classifying product review by semantic analysis. The fundamentals of opinion mining re analysed in their work. In the process of doing semantic analysis the author has give complete evaluation setup where web crawler

fetches the comments from a particular web page , comments are made sensible by spelling correction , removing stop words , naive bayes algorithm classifies polarity of comments , decision tree calculates overall polarity and finally providing a personalised experience to end users. Nitu Kumar et al in their paper titled "Sentiment analysis on e-commerce web sites using opinion mining " depicted that internet is no only about buying or selling but is also helpful in improving the efficiency to compete with other giants in the market. Opinion mining methodology is projected comprising of three stages : pre-processing, mine association rule and summarisation. This reviewed paper described that by opinion mining we can look properly before arriving at a decision, based on the ratings, reviews and stars given to a product. Weigang Zuo et al in their paper "The application of web data mining in e-commerce" presented web mining courses in e-commerce and some techniques dealing in web data mining. Web data mining includes data preparation where data is integrated, settled and pre-processed then actual data mining operates at mining stage. After this information is filtered and results are expressed. Various methods of web data mining are proposed. Finally the paper shows the applications of web data mining in e-commerce. The combination of web mining and e-commerce will help the enterprise gain competitive advantage. K. Mouthami et al in their paper "Sentiment analysis and classification based on textual reviews" proposed an algorithm named Sentiment Fuzzy Classification algorithm. It is a tough task for humans to adumbrate a movie review. For this, classification of sentiment on document level is already being used in existing systems. It actuates whether a movie review is negative or neutral. Classification accuracy is improved by the usage of tags of speech on movie review data set is told in the later part the introduced methodology. Kevin Anderey et al in their paper " Analysis of behaviour of customers in social networks using data mining techniques" showed the results using data mining techniques for scrutinizing the behaviour of customers of a fashion company on social networking site - instagram. The methodology used was CRISP-DM which was presented in terms of hierarchical process model containing set of tasks containing four levels of abstraction.

Table 1 Comparision Table For Previous Papers

AUTHOR	DOCUMENT REPRESENTATION	USED METHODS (FOR GENERALISATION)	PERFORMED TASKS (FOR GENERALISATION)
Craven et al[1]	-Ontology and Relational	-Inductive Logic Programming	-Classification of Hypertext
Crimmins et al[2]	- information of meta, URLs and Phrase	-Supervised and Unsupervised Classification Algorithms	-Clustering
Furnkranz at al[3]	-Bag of Hyperlink information	-Learning of Rule	- Classification of Hypertext
Mitchell et al[4]	-Bag of Hyperlink information	- Learning of Reinforcement	- Prediction of Hypertext
Muslea et al[5]	-Bag of word positions, Tags and words	- Learning of Rule	-Learning Extraction Rules
Shavlik and Elliasi et al	-Localised Word bags	-Reinforcement Learning	- Classification of Hypertext
Singh et al[6]	- Named entity and Concepts	-Classification Algorithm	-Pattern finding
Soderland[7]	- Named entity and Concepts	- Learning of Rule	- Extraction Rules Learning

III. SENTIMENT ANALYSIS (OPINION MINING)

Sentiment analysis also referred to as opining mining is an area of computational study that helps in analysing people's opinions, attitudes, sentiments. This is considered to be a bustling research area in the processing of natural language and is often studied in mining of Data, Text and Web. This study holds a great value in the field of business and society as a whole. The importance increase with the increasing number of online reviews. Opinion mining includes developing a system to collect and classify these

online reviews regarding a product. The industry depending on sentiment analysis has furnished due to the proliferation of commercial applications which provides a strong motivation for research in this field. Research in sentiment analysis not only holds importance in natural language processing but also has an importance in management sciences, political sciences, economics and social sciences as they are all altered by people's opinions. Though the research is of much importance, opinion mining has several challenges .Firstly , we cannot determine weather a word which is considered positive in one context is also to be considered positive in another context. For instance consider the word 'long'- If someone is referring a mobile's battery is long, that is positive, but if he said mobile's booting time is long, that is negative. Secondly, opinions are not always expressed by the people in the same way. People can also be contradictory in their statements. Therefore document level sentiment analysis recognises the opinions of the contents, mainly discuss the sentence level opinion mining, and treats the statements of features of the product for each view point as the analysis of object and therefore, we can find the customers inclination. It is the main focus of opinion minin. In order to avoid the over generalization comprehensive evaluation is done. Therefore specific opinion mining grabs attention of more people. An object is called an entity and is associated with types of components or set of attributes. To show both components and attributes a feature is defined and it is the subject of a review. The implication of the writer is found out in the comment by deciding the polarity of the word. Comments are expressed in either positive polarity or negative polarity which makes them positive comments or negative comments. The adjectives separated by 'and' have the same polarity whereas the adjectives separated by 'but' have opposite polarity. Sentiment Classification can be divided into several sub tasks : determining subjectivity, determining orientation and strength of orientation [17].

IV. CONVENTIONAL METHODS

A. Naive Bayes Classification

Naive bayes is used for sequencing of words with word frequencies as the features. Naive bayes are a type of classifiers, which are a family of simple probabilistic classifiers; having independent assumptions between features, known as 'independent feature model'. These are highly scalable classifiers, whose pre-requisites are a number of parameters linear in the number of variables in a problem; thus becoming one of the widely used and baseline method for the categorization of text. In this approach, the class

$$c^* = \text{argmaxc } P(c/d)$$ is assigned to a given document d;

Naive Bayes uses' Bayes rule' –

$$P(c/d) = [\, P(c)\, P(d/c)\,]/\, P(d) \qquad \text{- (1)}$$

For estimating the term P(c/d), Naive Bayes reduces it by considering fi's to be conditionally independent,

$$P_{NB}(c/d) = P(c)\, (\prod^{m}_{i=1} P(f_i/c)^{ni(d)})\, /\, P(d) \qquad \text{- (2)}$$

m = number of features

fi = feature vector

Though the conditional independence assumptions of the Naive Bayes doesn't hold in real world; still surprisingly it tend to perform well

B. Simple K Means Clustering

This is a method which is known for quantizing vector. The main aim of this method is the partitioning of n-observations into k-clusters. In simple k-means clustering , each cluster owns some observations and these observations belongs to the nearest mean. This is basically a method which aims at the grouping of the similar observations together into one group. This simple k-means clustering method is applied using two algorithms, both of these algorithms use cluster canters for the modelling of the data. Rocchio algorithm- we can use the 1-nearest neighbour classifier in order to classify new data. This classifier classifies new data into existing clusters on the cluster centres which are obtained by k- means. This is often referred as nearest centroid classifier. Lloyd's algorithm- An iterative refinement technique is used in this algorithm. The algorithm alternates between two steps: Assignment step: this is the first step. Here every observation is equated to the cluster whose mean holds the least within-cluster sum of squares (WCSS). Each is assigned to exactly one

C. Decision Tree

Decision tree is a technique which is used to display an algorithm. It is a tree- like model. It is a decision support tool which is based totally on the decisions, the possible results, outcomes, costs of resources and utilities. This is used for analyzing the decision and thereafter, helps in identifying which strategy when used, is more likely to reach the goal and give better results over the other

strategies. The best value, the worst value and expected values for different scenarios can easily be determined using decision tree and thus result in better outcomes It is linearized into what is called the decision rules. The outcome is the content that belongs to the leaf node and the conjunctions in the if clause are the conditions General form: - If condition1 and condition2 and condition3 then outcome. It can also be referred as a generative model of the rules of induction from the empirical data given.

V. RESEARCH PROBLEMS

Different and various challenges are faced at different levels or stages of web mining. As the data on web is immensely rising in amount every passing day we need effective techniques to mine Big Data. The feedbacks and reviews on e-commerce websites that we get are very large in number. This makes it very difficult for users to read and analyse and go through each and every comment present on the e-commerce site. Query processing techniques that are used in e-commerce web sites simply apply the blind keyword matching mechanism which ignores the relevance of queries with respect to documents. Deduction capabilities that are required are missing from e-commerce web sites. Various techniques that are available follow the hard rejection approach while finding out the accuracy and error rate in the comments while analysing the comments. This is not valid as relevance is a gradual property itself [16]. Since there are enormous comments or feedback for a single product on the e-commerce sites we need the ranking methodology in order to determine the quality of a product as we cannot scan through all the comments available on the e As the technology advances with time, various updates are made on the products, so the feedbacks which were provided to the earlier version of a product are now obsolete and have no value and should be removed from the site.

VI. PERFORMANCE PARAMETERS

True positives(TP), True negatives(TN), False positives(FP), False negatives(FN) are given to documents by a classifier. After this these performance parameters are used for comparison of the class labels; where True positive implies to truly classified as positive. True negative means truly classified as negative terms.

A. Accuracy

Accuracy gives us the classification of the performance. It can be calculated as ratio of correctly classified examples to the total number of examples.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

B. Precision and Recall

Precision and Recall are used as matrices for evaluation of the performance. Precision gives the exactness measure and recall gives the completeness measure. This can thus, be referred to as the extension of accuracy.

Precision can be calculated as the number of examples correctly labelled as positive divided by the total number classified as positive and recall is the number of examples correctly labelled as positive upon total number of examples that are truly positive.

C. F – Measure

The third parameter is F- measure. The harmonic mean of precision and recall is called as the F- measure. This F-measure optimizes the system to favour precision or recall, and determine which one has more positive influence on the final result. This is the evaluation metric for aspect identification and aspect sentiment classification.

$F_1 = (2*\text{percision}*\text{recall}) / (\text{precision} + \text{recall})$

VII. EVALUATION STEPS

A. Text Pre-Processing

Text pre-processing is considered to be an important process in sentiment analysis. The data which has to be analysed can be in any format and any form. Pre-processing can be either supervised or unsupervised. Further it is divided into two subcategories

1) *Tokenisation :* Comments are composed of sets of words that are known as tokens. The entire comment is first broken down into individual words or tokens and then passed on for further processing.
2) *Filtering :* Each language comprises of its own stop words. These stop words are to be removed before further processing. Few often used stop words of English language are ‘a’ , ’of’ , ‘the’ , ‘i’ , ‘it’ , ‘and’ these are commonly called as functional words which are meaningless with regard to the document. Hence it is practical and required not to take such words into account.

B. Text Transformation

Each sentence has a score and each sentence's score in the comment is calculated by summing weights of each token in that comment. Each token's weight is calculated by multiplying TF and IDF of that word based on adjective word extracted. The TF and IDE are defined as

C. Feature Selection

Most of the feature selection methods used for document level categorization can also be applied in Opinion mining. Easiest approach being choosing the most frequently occurring words of same polarity and keeping them together, this way we can increase the chances of Comments falling into one of the two polar categories.

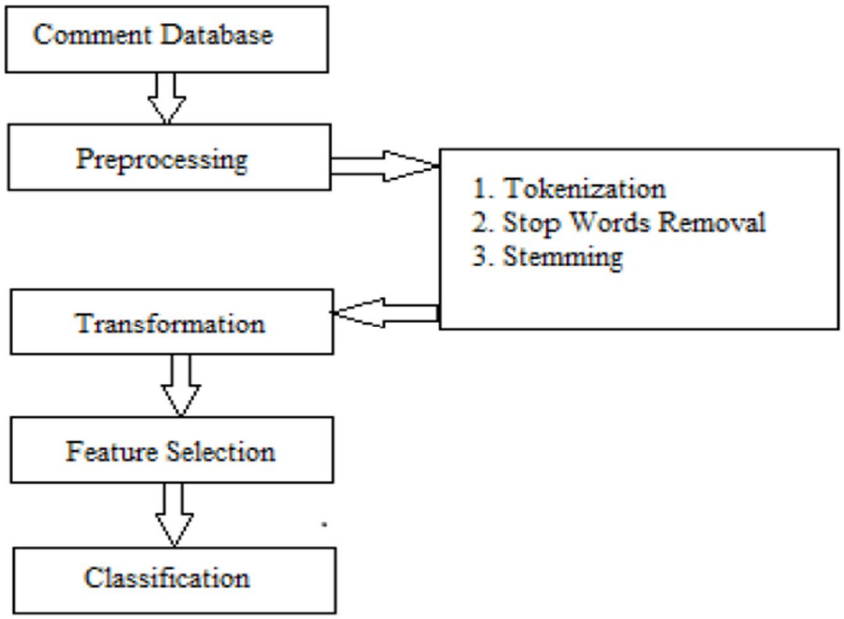


Fig. 2 Evaluation Steps

VIII. ALGORITHM USED

A. Download comments
B. Form keywords by tokenising comments
C. Filter the functional words
D. If the word is not in dictionary, correct the spelling
E. Apply Simple K Means Clustering
F. Then Apply Naive Bayes Classifier
G. Calculate Accuracy, Precision, Recall and F measure
H. Apply Decision Tree Algorithm
I. Compute Sentiment
J. Calculate the Average Rating
K. Return Sentiment and Sentiment Score of comments

IX. CONCLUSION

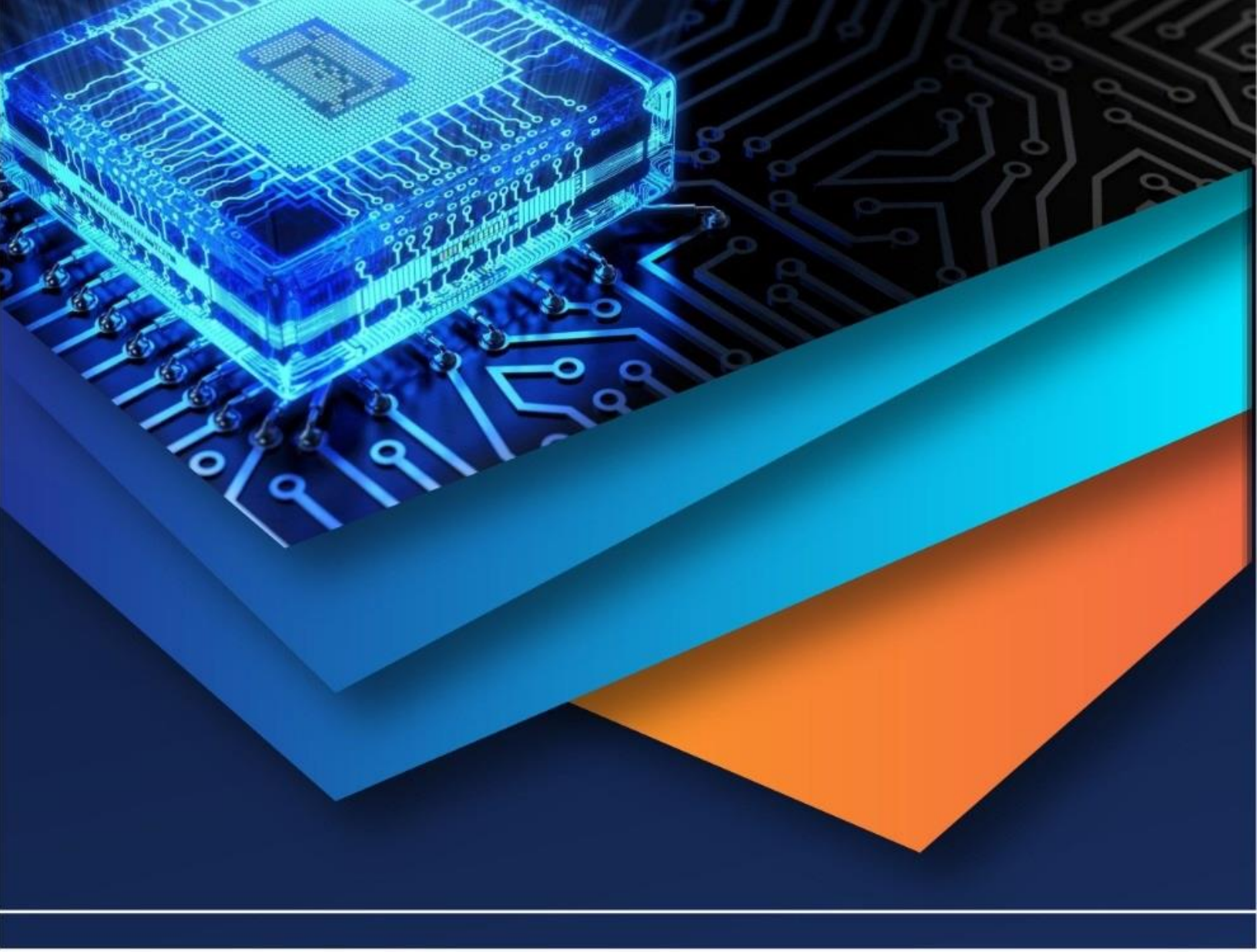
Even though e-commerce sites have gained huge popularity because of its various advantages like cost, variety, faster delivery, but trusting these sites blindly is still an issue. Users have to choose not from hundreds but from millions of products in an online store. The quality of products being uneven the opinion of users expressed via their comments becomes an important source of information to judge the product's quality. In this paper we have used a technique called semantic analysis of e-commerce sites, which makes it possible to analyze large amount of data and thereby extracts the actual opinions expressed through them that help the customers. First, the downloaded comments are pre-processed where the comments are divided into tokens called as

tokenization. Followed by this removal of stop words is done and a spell check of the words also take place. After this, out of many possible permutations of pre-processing and classifying algorithms that exists, we have applied a combination consisting of NAIVE BAYES as a classifier, K-means clustering as a clustering technique and decision tree for calculating the overall polarity of words. This gave us better accuracy, better precision and an improved error rate. Sentiments are computed after which average ratings are calculated and finally the sentiments are returned along with the sentiment score of comments. In future the performance can be improved by applying other techniques. The proposed work can be enhanced by using some other combinations of classifiers and clustering algorithms. We can also extend our study to analyzing the comments that are in languages other than English.

fig. 3 conclusion

REFERENCES

[1] M. Craven and J. Shavlik, "Using neural networks for data mining,"Future Generation Comput. Syst. (Special Issue on Data Mining), vol.13, pp. 211–229, 1998. [1]
[2] F. Crimmins, A. Smeaton, T. Dkaki, and J. Mothe, "Information dis-covery on the internet," IEEE Intell. Syst., vol. 14, pp. 55–62, 1999. [2]
[3] J. Furnkranz, "Exploiting structural information for text classification on the WWW," in Proc. Advances Intell. Data Anal. 3rd Int. Symp., IDA99,1999, pp. 487–498. [3]
[4] T. Mitchell, D. Freitag, and T. Joachims, "Webwatcher: A tour guide for the world wide web," in Proc. Int. Joint Conf. AIIJCA197, 1997, pp.770–777. [4]
[5] Muslea, S. Minton, and C. Knoblock, "Hiererchical wrapper induction for semistructured information sources," J. Autonomous Agents Multia-gent Syst., vol. 4, pp. 93–114, 2001. [5]
[6] L. Singh, B. Chen, R. Haight, P. Scheu, I. Muslea, S. Minton, and C.Knoblock, "Wrapper induction for semistructured web based information sources," in Proc. 2nd Int. Conf. KDD Data Mining, 1998, pp.329–333. [6]
[7] S. Soderland, "Learning information extraction rules for semistructured and free text," Machine Learning (Special Issue Natural Language Learning), vol. 34, no. 1/3, pp. 233–272, 1999. [7]
[8] Mladenic and M. Grobelnik. Efficient text categorization. presented at Proc. Text Mining Workshop 10th European Conf. Machine Learning ECML98. [Online]http://www=ai.ijs.si/DunjaMladenic/papers/PWW/pwwWsheEMCL99.ps.gz [8]
[9] R. Ghani, R. Jones, D. Mladenic, K. Nigam, and S. Slattery, "Data mining on symbolic knowledge extracted from the web," in Proc. 6th Int. Conf. Knowledge Discovery Data Mining (KDD-2000) Workshop on Text Mining, Boston, MA, Aug. 2000, pp. 29–36. [9]
[10] S. Loh, L. K. Wives, and J. P. M. de, "Concept based knowledge discovery from texts extracted from the web," ACM SIGKDD Explorations,vol. 2, pp. 29–40, July 2000. [10]
[11] S. Chakrabarti, "Data mining for hypertext," ACM SIGKDD Explorations, vol. 1, no. 2, pp. 1–11, 2000. [11]
[12] R. Cooley, B. Mobasher, and J. Srivastava, "Web mining: Information and pattern discovery on the world wide web," Proc. 9th IEEE Int. Conf.Tools with Artificial Intelligence, pp. 558–567, Nov. 1997. [12]
[13] Levy and D. Weld, "Intelligent internet systems," Artificial Intell.vol. 118, no. 1–2, 2000. [13]
[14] W. Y. Lin, S. A. Alvarez, and C. Ruiz. Collaborative recommendation via adaptive association rule mining. presented at Int. Workshop Web Mining for E-Commerce (WEBKDD'00). [14]
[15] Mobasher, H. Dai, T. Luo, M. Nakagawa, Y. Sun, and J. Wiltshire,"Discovery of aggregate usage profiles for web personalization," presented at the Proc. KDD-2000 Workshop Web Mining E-Commerce,Boston, MA, Aug. 2000. [15]
[16] S. Mitra, S. K. Pal, and P. Mitra, "Data mining in soft computing framework: A survey," IEEE Trans. Neural Networks, vol. 13, pp. 3–14, Jan.2001. [16]
[17] Yu, C., & Ying, X. (2009, December). Application of Data Mining Technology in E-Commerce. In Computer Science-Technology and Applications, 2009. IFCSTA'09. International Forum on (Vol. 1, pp.291-293). IEEE. [17]
[18] Mouthami, K., Devi, K. N., & Bhaskaran, V. M. (2013, February). Sentiment analysis and classification based on textual reviews. In Information Communication and Embedded Systems (ICICES), 2013 International Conference on (pp. 271-276). IEEE. [18]



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY