Sentiment Analysis Based Product Selection for Enhancing E-Commerce

P. Sudhakaran, M. Jaiganesh

Abstract: ***One of the fast growing, developing and highly used technology in various computing industries is data mining. Sentiment or opinion mining is a kind of data mining, where it follows the major processes of natural language processing. Nowadays, sentiment analysis meets a high demand. In this paper, it is aimed to consider the problems of sentiment analysis such as classification on opinion and attribute words, because it is the basic problem of sentiment analysis. This paper aimed to use one of the popular machine learning algorithms as Multi-Class Support Machine algorithm for classifying sentiment polarity with detailed description. The proposed method is implemented in Python software and experimented on online-product-reviews data taken from Amazon.com. Sentence level and opinion level classification is obtained with promised outcomes. From the results it is noted that the proposed method outperforms than the existing method such as Naïve Bayes and Random Forest algorithms.***

Keywords: sentiment analysis, opinion mining, e-commerce, polarity classification, machine learning algorithms, customer-review data, amazon data.

I. INTRODUCTION

Data Science is a computer science discipline which incorporates several technologies such as machine learning and data mining to analyse the unprocessed data. Data mining is a branch of data science where it mines a user requested pattern of data from a huge set of data. The process of data mining has several complexities due to Velocity, Volume, Value, Variety, and Veracity which degrades the performance in terms of mining accuracy. Data analytics provides a solution to overcome the above said problems in the field of data mining, which results in an increased performance of accuracy. There are many techniques in data mining such as Association mining, cluster analysis, sentiment analysis or opinion mining. There are various applications such as unsupervised learning, customer feedback data analysis and classifying web forum. In recent busy world, people buy and sell their products in online easy and quick. It improves the profit of customer-to-customer, business-to-customer and business-to-business structure. The opinion of each existing customer suggests the product or entities to any new customer. This paper is focused on sentiment analysis in the application of e-commerce to predict the performance accuracy of the data.

II. SENTIMENT ANALYSIS

In the day-to-day digital world, every user shares their experience about every single data over the internet. The data shared are either structured or unstructured. The structured data is obtained through a GUI of their respective e-commerce sites in the form of comments from which the data can be summarized. The unstructured data are the comments posted by the user which cannot be summarized as positive, negative or neutral. The user data shared are categorized using an interface, which is achieved by the process of sentiment analysis. This process extracts all the required data from the e-commerce site and categorizes the data.

Sentiment analysis measures the accuracy of the data by analysing the users experience. There are ten ideas to analyse the users experience. They are box office revenues [1], brand monitoring [2], computational history [3], customer feedback [4], dropout rates [5], monitoring of political sentiments [6], product reviews [7], stock market predictions [8], story arcs [9], sentiment analysis as a sub task [10].

(Pandey et al. 2016; Rain 2013) [11][12] stated that, in order to track the opinion of the customer reviews the dataset have to be classified and categorized based on common traits. (Liu et. al 2014) [13] mentioned that when huge number of datasets is used, the classification process can be automated. [14][15][16](Liu 2015; Pang et. al 2002; Turney& Littman 2003) described sentiment analysis in terms of NLP in which the user reviews were classified as positive and negative reviews. But (Singla et. al 2013) stated that this positive and negative classification might not work properly when certain words like “not” are used. Also Ye et. al (2009) [18] proved that the negative word “unpredictable” was considered as a positive word for the tourists. (Liu et al. 2007; Pang et al. 2009; Ye et al. 2009) [18][19] [14] implemented sentiment classification in various fields like travel, movie and product reviews.

III. LIMITATIONS AND MOTIVATION

To experiment the data, there is no preprocessing is applied on the data. Pre-processing is mainly used for making the data as perfect form before applied into the algorithms. Online product reviews data has more irrelevant and unnecessary information which cannot be used in the data process. Some of the preprocessing tasks are eliminating unwanted white spaces, removing stop-words, and arranging the data in a standard format, etc. The output of the data analytics methods with pre-processing gives high accuracy while applied on reviews summary. The reviews consist of huge amount of words where it leads to sparsity in bag of words. So, in classification the summary of the reviews is used for limiting the number of words. Another limitation is most of the customer defined methods and

Revised Manuscript Received on January 2, 2020.

Dr. P. Sudhakaran, Professor, Department of CSE, SRM TRP Engineering College, Trichy, Tamilnadu – India.

M. Jaiganesh, Assistant Professor, Department of CSE, SRM TRP Engineering College, Trichy, Tamilnadu – India

some of the machine learning methods require preprocessing to improve the accuracy of classification. Haddi et al. (2013) [20] insisted that preprocessing improves the percentage of accuracy. Another limitation is, the bag of words never cares about the word location in the text where it has a negative effect on the semantic of a review. Pang et al. (2002)[15] described that a machine classifies a thing as positive or negative based on the number of positive or negative reviews respectively. Another challenge addressed in this paper is to identifying the negative and positive words by comparing the semantics of the sentences. Hence in this paper, it is motivated to design and implement MSVM method and compared with NB and DT classification methods.

IV. PROPOSED METHOD

Here the method and the environments are explained in detailed, that is how the data is obtained, prepared and experimented. In order to experiment and evaluate the proposed approach this paper used Phyton software. Python has a big set of libraries where it has methods for solving the machine learning algorithms. The libraries can be included easily as in C programming. One of the best libraries to feature out the various supervised machine learning algorithm is Sci-kit learn. Python also includes various libraries that belongs to SVM and Naive Bayes which are used in this paper. It also supports file extraction methods.

Naïve Bayes Method

Naïve Bayes (NB) method is one of the classification methods where it classified the data D into m number of classes C1 to Cm. Initially it trains the data and create a trained model for testing purpose. Each data is represented in the form of tuples X where the classifier predicts the class Ci on X, if and only if: P(Ci|X) > P(Cj|X), where, i and j belongs to [1 , m], and both are not equal. X is represented as x1, x2, …,xn, where n is the number of data tuples collected from n number of attributes (features). The P(Ci|X) is calculated as:

$$\boldsymbol{P(C_i|X) = \prod_{k=1}^{n} P(x_k|C_i)}$$

Where it is used to predict the classes from X.

Random Forest Method

Another important classification model is random forest classifier which has provided superior performance than decision tree algorithm regarding accuracy. One of the popular and superior than decision tree in terms of performance is random tree classifier. Random forest classifier is ensemble method due to bagging. Initially it creates k number of bootstrap samples on the data D =d1,d2,....dn, where each sample is represented as d1,d2...dk. Each sample Di comprises of equal number of tuples as d. Which were samples d with replacement from d. It means the original tuples of d have not been included in di but the remaining tuples may occur more than once. Then a decision tree is constructed over each Di and makes a forest (the forest comprises of k number decision trees). In order to predict the unknown tuple x every decision tree delivers a class predicted and counted as 1 vote. The final predicted class is considered based on the highest votes. Basically, the DT algorithm performs based on the SCI it learns algorithm is also called as CART (classification and regression trees). An index is used to represent the trees induction is called as Gini index (Gi). Now Gi for D is calculated as

$$\boldsymbol{GI(D) = 1 - \sum_{1}^{m} i\, P_i^2}$$

Where Pi is represented as the probability of tuples under the same class Di. The Gi measures the impurity of D. The measurement Gi always provides the impurity of D.

Support Vector Machine

SVM is a popular classification method that works on linear and nonlinear data. A linear hyper plane is used to separate the linear data from one class to another class. The linear hyperplane is an optimal hyperplane considered as the decision boundary for separating the classes 1 and 2 available in the data. SVM can classify only 2 classes (1 or 0, True, false). The hyper plane is represented mathematically as

$$\boldsymbol{W.X + b = 0}$$

Where the weight of the vector id represented as W. That is,

$$\boldsymbol{W = w1, w2 \ldots\ldots wn}$$

And the training tuple is denoted as X, scalar is represented as b. The hyper plane optimises the data by minimising the problem transformation ||W|| is mathematically calculated as

$$\boldsymbol{||W|| = \sum_{i=1}^{n} i\, \alpha_i y_i x_i}$$

Where α_irepresents the numeric parameters, y_irepresents the labels of the support vector Xi.

$$\textbf{Means if } \boldsymbol{y_i = 1} \textbf{ then } \boldsymbol{\sum_1^n i\, w_i x_i \geq 1}$$
$$\textbf{if} \boldsymbol{y_i = -1} \textbf{ then } \boldsymbol{\sum_1^n i\, w_i x_i \geq -1}$$

In case of high dimension data SVM uses nonlinear hyperplane separable method. From the above discussion it is understanding that SVM can classify only 2 class such as 1 and -1. But the customer review data has 3 classes such as positive, negative and neutral. The classification process of SVM is shown in Figure-1.

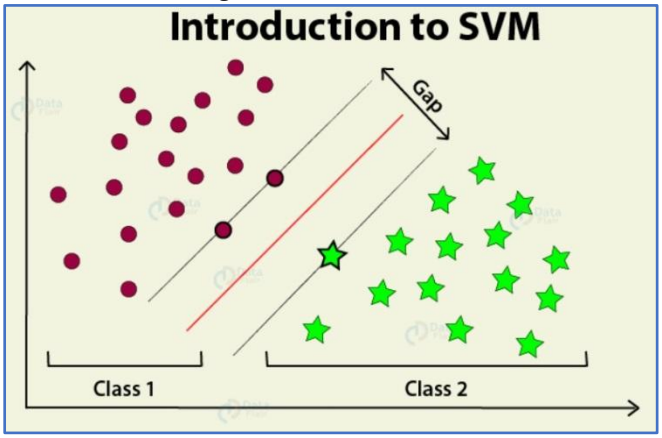

Figure-1. Classification using SVM

Multi Class SVM

Comparing with the SVM, the multiclass SVM can predict more classes. In accordance the objective of this paper, to classify the three different classes in the online customer review data this paper aimed to use multiclass SVM algorithm. The hyperplaneoptimises the data by minimising the problem transformation ||W|| is mathematically calculated as

$$||W|| = \sum_{1}^{n} i\,\alpha_i y_i x_i$$

Where,$\alpha_i$represents the numeric parameters, $y_i$represents the labels of the support vector **Xi**.

Means **if** $y_i = 1$ **then** $\sum_1^n i\, w_i x_i \geq 1$ **// Positive class**
if$y_i = -1$ **then** $\sum_1^n i\, w_i x_i \geq -1$ **// negative class**
if$y_i = -1$ **then** $\sum_1^n i\, w_i x_i == 0$ **// neutral class**

Based on the above equation, the MSVM can predict the customer reviews as positive, negative or neutral. To understand the MSVM method very clearly, it is given in the form of algorithm and it is given in Figure-2.

Algorithm MSVM **(*input data D*)**
{
X *is the training dataset*
X1 *is the test data*
I *is the infinite number*
Fn *is the final number of clusters*
D *is the reduction parameter*
While (***I*** $>$ $-Fn$) *do*
{
I $=$ *cluster the data and get* ***X1*** *as* ***X1, X2, …, Xn*** */*
/ using ensemble clustering method
For each cluster ***j***
$=$ **1** ***to I*** *compute the score*$(X(X1_i), f, r)$
$X1_i$ **= SA (*score*$(X(X1_i), X, X1)$)**
Eliminate ***D****% of clusters having lower scores*
End for
Merge all $X1_i$*into one pool* ***X1***
}

}

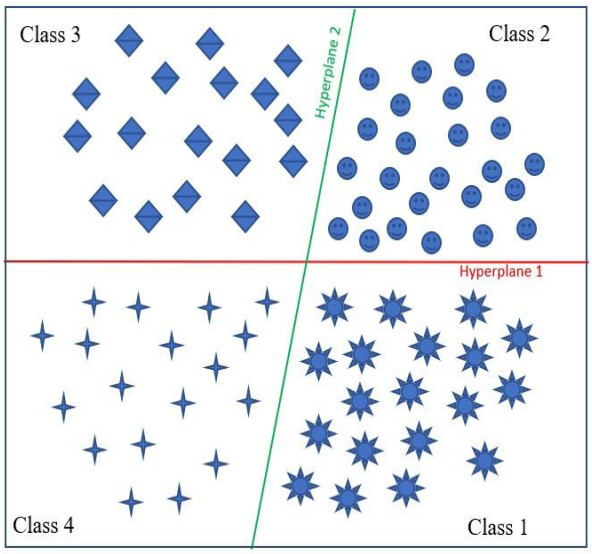

Figure.2. Multi-Class Support Vector Machine

Data Set

Data is collected from SNAP set of Amazon.com. The format of this data is a collection of sentences and paragraphs. Also, there is no facility to extract data from amazon. So JSON is used to download data in the form of one review per line and are converted into a famous CSV format, which can be easily handled by python. The entire data set comprises of comprises of 252000 reviews over beauty products. To understand the dataset and the experimental evaluation clearly, a set of important features selected for analysing the data is given in table.1. Also, a sample customer review in JSON file format is given in Figure.3. The has all the fields given in Table-1. Among the fields the review text is a more than one sentence and it has a greater number of words, having collection of stop words, delimiters and keywords.

Table.1. Features of Data and Description

Feature	Description
ReviewerId	Id of the user
"asin"	Product Id
ReviewerName	Name of the user
Helpful	Fraction of users who found the review helpful
ReviewText	Text of the review
Overall	Rating of the product
Summary	Review summary
UnixReviewTime	Time of the review in unix time
Reviewtime	Date of the review

{"reviewerID": "A4DIVP0NEQDA3",
"asin": "B000052WYD",
"reviewerName": "Amazon Customer",
"helpful": [1, 1],
"reviewText": "This product met my needs for Upper Lip shadow. I found it to be non oily, easy to smooth over skin and light on skin. You will need to use a press power for your skin tone after applying the concealer. Very please with the results.",
"overall": 5.0,
"summary": "Magic Stick for Upper Lip Shadow",
"unixReviewTime": 1386806400,
"reviewTime": "12 12, 2013"}

Figure-3. Customer Review Data in JSON Format

Data Preparation

To improve the accuracy of data mining the data is written in simple format to remove the unnecessary features. Initially it removes the unnecessary features and summarizes the important features to expose the reviews. Examples are review, text of the review and productId. The review score is given in the form of number of stars from 1 to 5. The reviews with stars less than 3 is treated as negative review and if the score is greater than or equal to 3 is treated as positive reviews.

The reviews which have mixed score are referred to as neutral reviews.

During the implementation and experiment it is found that the proposed approach is not only suitable for analysing the product reviewed data. It is also used as a common recommendation system to find highest recommended entities such as hotels, restaurants and products. It finds the entities in accordance to the structured attributes and unstructured opinions given by the web users. The entire architecture of the proposed approach suitable for finding the highest score-based entities are illustrated in Figure-4.

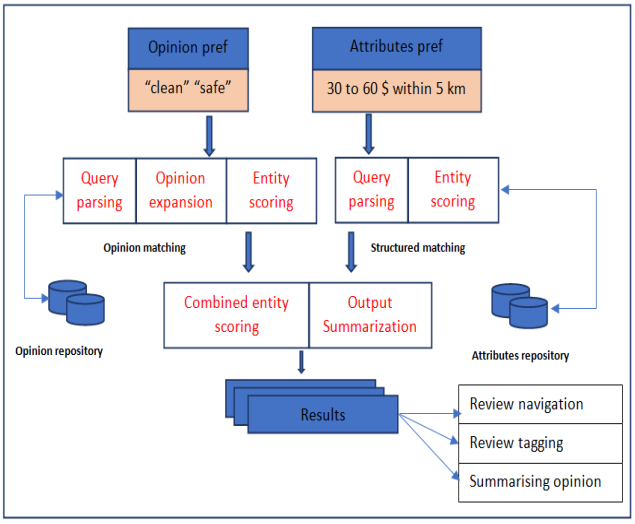

Figure-4. Proposed Methodology

From the figure-4 it is understood that the user query is categorized into structures attributes and unstructured opinions. The unstructured opinion is parsed using the query parser and the opinion is given a score ranging from 1 to 5. Similarly, the structured attributes are parsed and a score is given ranging from 1 to 5. Both the scores are combined to generate an entity score and the results are summarized. Based on the summarized result and the entity scores the final results are generated as whether the product is recommended or not. Table 3 shows the comparison of the performance accuracy of three different kinds of algorithms based on reviews and summaries.

V. EXPERIMENTAL RESULTS AND DISCUSSION

Two experiments were conducted where the first experiments used the entire data set and the second experiment used only a small set of datasets. In the first experiment the reviews with 3 stars were removed as there are enough reviews and in the second experiment the reviews with 3 stars are considered as negative. The positive reviews were scored as "1" and the others as "0".

Machine Learning Classifiers

The experiment is carried out in two phases namely training the dataset and testing the dataset. In both the experiments mentioned above, the reviews and the review summaries were trained and tested separately. In the first experiment the classifier trained 150000 datasets and tested 48500 datasets for performance accuracy. Using Bag of words model the review texts are converted into numerical features. The performance is measured by training and applying the classifiers on the test data.

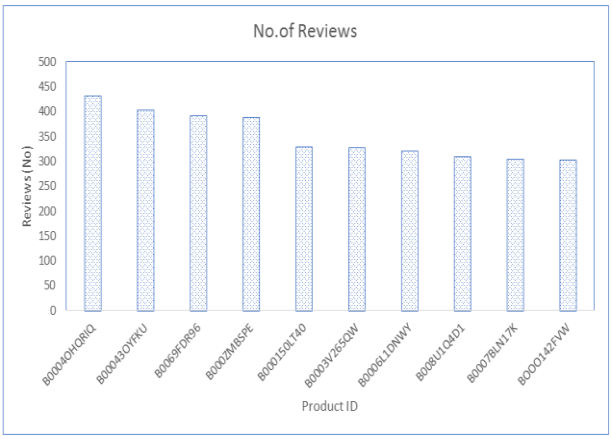

Figure-5. Product ID versus Number of Reviews

In the second experiment with limited number of datasets the first 10products with most reviews were grouped based on the productId. It is depicted in Figure-5. 300 samples were taken as training dataset from each product and then they are tested for performance accuracy. Figure-6 shows the comparison of the performance accuracy of three different kinds of algorithms based on reviews and summaries. From the graph it is depicted that the Naive Bayes classifier produced 90.16% -92.72% of accuracy and the Random Forest algorithm produced 93.02%-93.20% of accuracy. The proposed method MSVM obtained 96.03%-96.31% of accuracy.

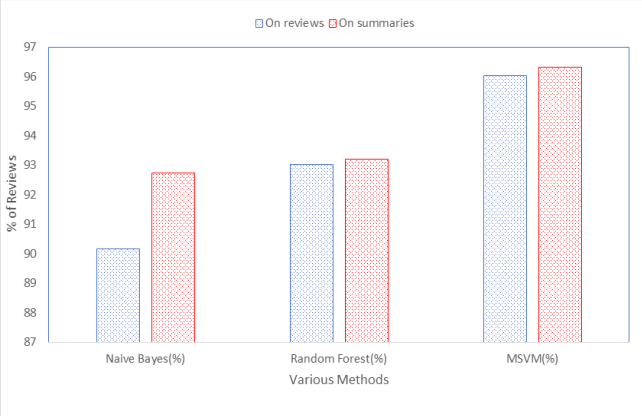

Figure-6. Performance accuracy based on reviews and summaries

Figure-7 shows the comparison of the performance accuracy of three different kinds of algorithms based on the productId for the first 10 datasets using the reviews. From the graph it is depicted that the performance accuracy of Naive Bayes classifier was ranging from 89.97%-89.98% and that of the Random Forest algorithm was ranging from 89.43%-89.69% of accuracy. The proposed method MSVM obtained 91.78%-93.8% of accuracy.

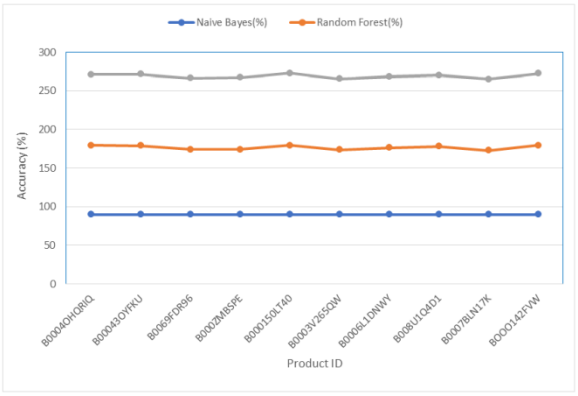

Figure-7. Performance accuracy based on productId using reviews

Figure-8 shows the comparison of the performance accuracy of three different kinds of algorithms based on the productId for the first 10 datasets using the summaries. From the graph it is depicted that the performance accuracy of Naive Bayes classifier was ranging from 88.55%-90.11% and that of the Random Forest algorithm was ranging from 81.54%-88.73% of accuracy. The proposed method MSVM obtained 91.51%-93.59% of accuracy.

From all the experiments it is obtained that MSVM outperforms than the other two algorithms and the performance accuracy of the proposed method is obtained to a maximum range.

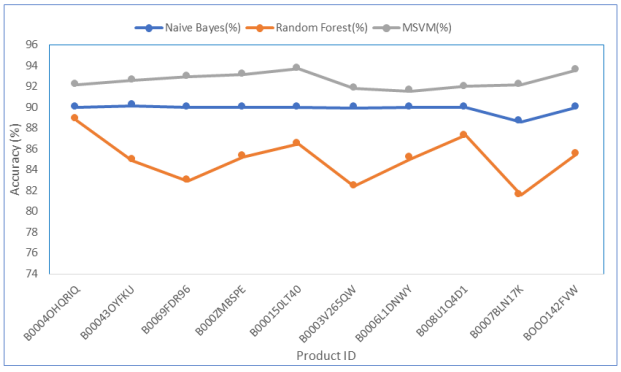

Figure-8. Performance accuracy based on productId using Summaries

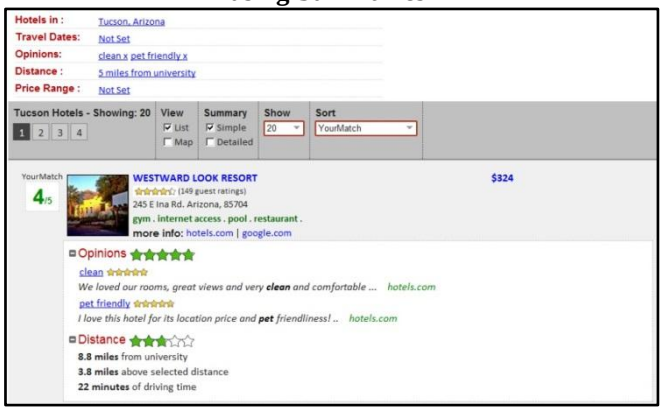

Figure-9. Hotels within 5 miles of University of Arizona – pet friendly and clean

Figure-10. Buzzwords for the Radisson Suites Tucson

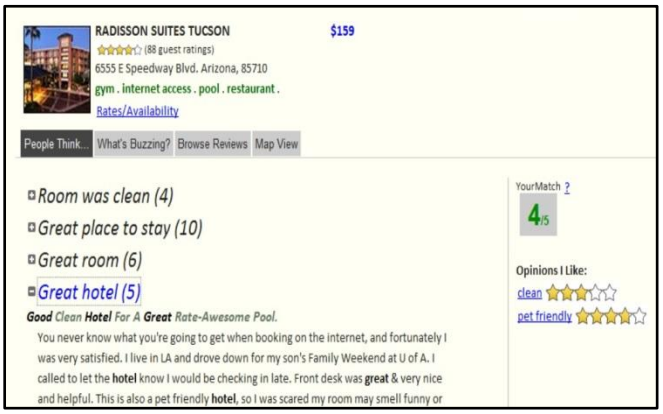

Figure-11. Opinion summary for the Radisson Suites Tucson

The experiment was tested through a web GUI and the following results are obtained. Figure-9. Shows the test result based on both structured attributes and the unstructured opinion. Figure-10 shows the result of MSVM based on the structured attributes using reviews and Figure-11. Shows the result of MSVM based on unstructured opinion using summaries.

VI. CONCLUSION

The main objective of this paper is to design and implement a ML algorithm for analysing the online product reviews data. To do that a multiclass SVM classifier is designed, implemented and experimented on product reviews data taken on amazon.com. The main advantage of this paper is to analyse the data based on the structured attributes and unstructured opinions. Hence it is highly suitable for any kind of sentiment or opinion based mining applications. The MSVM is experimented and the results are verified in both scenarios such based on the reviews and integration of structured attributes. From the obtained results it is found that the MSVM outperforms than the Naive Bayes and Random Forest. Hence it is proved that MSVM can be applied for high dimensional data also.

In future work, a deep learning algorithm is applied for sentiment analysis and the performance is compared.

REFERENCES

1. Asur, S. and Huberman, B.A., 2010, August. Predicting the future with social media. In Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01 (pp. 492–499). IEEE Computer Society.
2. Ghiassi, M., Skinner, J. and Zimbra, D., 2013. Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. Expert Systems with applications, 40(16), pp.6266–6282.
3. Acerbi, A., Lampos, V., Garnett, P. and Bentley, R.A., 2013. The expression of emotions in 20th century books. PloS one, 8(3), p.e59030.
4. Gamon, M., 2004, August. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In Proceedings of the 20th international conference on Computational Linguistics (p. 841). Association for Computational Linguistics.
5. Wen, M., Yang, D. and Rose, C., 2014, July. Sentiment Analysis in MOOC Discussion Forums: What does it tell us?. In Educational data mining 2014.
6. Abbasi, A., Chen, H. and Salem, A., 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. ACM Transactions on Information Systems (TOIS), 26(3), p.12.
7. Turney, P.D., 2002, July. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th annual meeting on association for computational linguistics (pp. 417–424). Association for Computational Linguistics.
8. Bollen, J., Mao, H. and Zeng, X., 2011. Twitter mood predicts the stock market. Journal of computational science, 2(1), pp.1–8.
9. Reagan, A.J., Mitchell, L., Kiley, D., Danforth, C.M. and Dodds, P.S., 2016. The emotional arcs of stories are dominated by six basic shapes. EPJ Data Science, 5(1), p.31.
10. Pang, B. and Lee, L., 2008. Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval, 2(1–2), pp.1–135.
11. Kotzias, D., Denil, M., De Freitas, N. and Smyth, P., 2015, August. From group to individual labels using deep features. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 597–606). ACM.
12. PriyankPandey, Manoj Kumar, and PrakharSrivastava. Classification techniques for big data: A survey. In Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on, pages 3625–3629. IEEE, 2016.
13. Callen Rain. Sentiment analysis in amazon reviews using probabilistic machine learning. Swarthmore College, 2013.
14. Bing Liu. Sentiment analysis: Mining opinions, sentiments, and emotions. Cambridge University Press, 2015.
15. Bo Pang, Lillian Lee, and ShivakumarVaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processingVolume 10, pages 79–86. Association for Computational Linguistics, 2002.
16. Peter D Turney and Michael L Littman. Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems (TOIS), 21(4):315–346, 2003.
17. Qiang Ye, Ziqiong Zhang, and Rob Law. Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. Expert systems with applications, 36(3):6527–6535, 2009.
18. Jingjing Liu, Yunbo Cao, Chin-Yew Lin, Yalou Huang, and Ming Zhou. Low-quality product review detection in opinion summarization. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), 2007.
19. Emma Haddi, Xiaohui Liu, and Yong Shi. The role of text pre-processing in sentiment analysis. Procedia Computer Science, 17:26–32, 2013.

AUTHORS PROFILE

Dr. P. Sudhakaran, Professor, Department of CSE, SRM TRP Engineering College, Trichy, Tamilnadu – India.

M. Jaiganesh, Assistant Professor, Department of CSE, SRM TRP Engineering College, Trichy, Tamilnadu – India