Opinion Mining using Machine Learning Techniques

Nirmal Godara, Sanjeev Kumar

Abstract-Sentiment analysis or opinion mining has gained much attention in recent years.With the constantly evolving social networks and internet marketing sites, reviews and blogs have been obtained among them, they act as an significant source for future analysis and better decision making. These reviews are naturally unstructured and thus require pre processing and further classification to gain the significant information for future use. These reviews and blogs can be of different types such as positive, negative and neutral . Supervised machine learning techniquess help to classify these reviews. In this paper five machine learning algorithms (K-Nearest Neighbors (KNN), Decision Tree, Artificial neural networks (ANNs), Naïve bayes and Support Vector Machine (SVM))are used for classification of sentiments. These algorithms are analyzed usingTwitter dataset. Performance analysis of these algorithms are done by using various performance measures such as Accuracy, precision, recall and F-measure. The evaluation of these techniques on Twitter datasetshowed predictive ability of Machine Learning in opinion mining.

Keywords-Sentiment Analysis, KNN, Decision Tree , Artificial neural networks (ANNs), Naïve bayes and SVM

I.INTRODUCTION

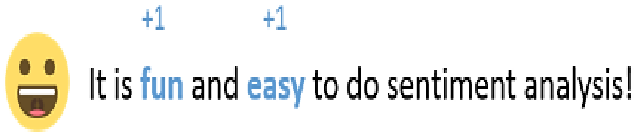
The incredible effect of web based life worldwide has prompted the disclosure of estimation investigation. The ongoing developments of keen innovations utilizing portable based correspondence has involved enormous measure of information creation. The web based life gives a capacity to share considerations, feelings, and feelings. The term assessment examination (SA) is prevalently known as sentiment mining. This is a procedure of feeling characterization generally passed on by a content that might be sure, negative or nonpartisan. The accessible information via web-based networking media has added to tremendous research utilizing estimation investigation. The twitter-based online networking speaks to a Gold-Dig method for investigating the presentation of the brand. Huge assessments of the general population are found over Twitter that are straightforward, educational, and easy-going when contrasted with the formal kind of information study examination utilizing magazines or reports. A large number of individuals offer and share their feelings over the media examining about the products whom they are associated with.

Revised Manuscript Received on December 30, 2019.
* Correspondence Author

***Nirmal Godara,** Ph.D Scholar, Department of Computer Science and Engineering, Guru Jambheshwar University of Science and Technology Hisar (Haryana) India. E-mail: nirmalgodara@gmail.com
Sanjeev Kumar, Associate Professor, Department of Computer Science and Engineering,Guru Jambheshwar University of Science and Technology Hisar (Haryana) India. E-mail: sanjukhambra@yahoo.co.in

At the point when such sort of opinions is distinguished over the social-media, at that point the data picked up from such estimations speaks to productive outcomes profiting huge organizations or associations. This information is exceptionally useful to screen execution of various brands and to find time span and viewpoints getting polar feelings.

The idea of opinion investigation is comprehended by joining the expressions "Senitiment" and "Examination". The word conclusion speaks to feeling that can be euphoric, confounding, bothering, diverting. The slants are the emotions dependent on specific mentalities and suppositions as opposed to actualities because of which feelings are of abstract nature [2] [14]Specifically, the method of online review-basedsentiment analysis (SA)has established hot field of research.Now days, the social networking sites (SNS) such as Facebook, Twitter, YouTube, and MySpacehaveachieved great popularity. These sites enablesan individual or a group to build connections and share information with others in a very simple and timely manner and it allows the users to use services like blogs, picture share, etc. Twitter has formed an exceptional collection of public opinions about each and every global entity generating interests known as Tweets. These are also known as the micro-blog due to its ability of having short text feature. Twitter presents an excellent platform for modelling an opinion and the way to present that opinion. Sentiment Analysis mainly intends to understand the public opinions and it distributes them into categories like negative, positive, or neutral. Fig1 shows process of sentiment analysis. With regards to a twitter supposition investigation, at its least difficult, estimation examination measures the state of mind of a tweet or remark by tallying the quantity of optimistic and destructive words. By subtracting the negative from the optimistic, the feeling score is produced [13]. For instance, this remark produces a general feeling score of 2, for having two positive words:



You can push this basic methodology somewhat further by searching for invalidations, or words which turn around the conclusion in an area of the content:

-1
-1
I don't like reading all of the negative Tweets!

The nearness of the word doesn't before like delivers a negative score instead of a positive one, giving a general notion score of - 2.

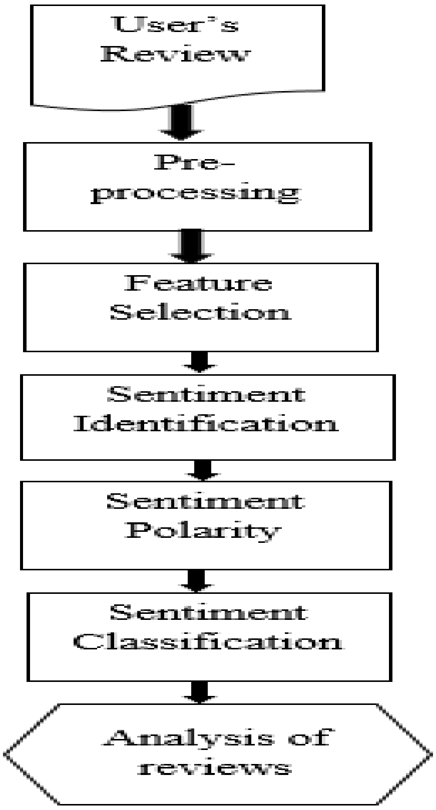


Figure 1: Sentiment Analysis process

There are various levels of sentiment analysis

Has been discussed are as follows:

A. Aspect level or Feature level*:*Aspect level is the opinion mining and feature-based summary. The classification concerns the identification and extraction of product characteristics from source data. This type is used when we need sentiments about the desired appearance.

- Task1: identify and extract the features of objects that were commented on by a commentator (for example, a reviewer)
- Task2: determine if the views of functions are negative, positive or neutral.

B. Document level: Sentimental analysis at the document level classifies the entire opinion of the document based on different feelings about the product or service. This level classifies the opinion document as positive, negative or neutral.

- Task: Opinion Mining of whole document
- Classes: Positive, Negative and Neutral
- Assumptions: Each document emphases on a single entity and contains opinion from a single opinion holder.

C. Sentence or phrase level: The sentence level Opinion Mining identify whether each sentence having a positive, negative or neutral opinion for a brand or service. This type of sentence is used for reviews and comments that contain one sentence and written by the user.Various activities related to this:

- Task1: Identifying subjective/objective sentences

Classes: Objective and Subjective

- Task2: Opinion Mining of Sentences.

Classes: Positive or Negative

Assumption: A sentence holds only one opinion which may not always be true

D. Word Level*:* In the most recent works, the previous polarity of words and phrases was used to classify sentiment at the level of sentences and documents. Classification of words by sentiment mainly uses adjectives as characteristics, but adverbs. Two methods for automatically annotating sentiment at the word level:

- Dictionary-Based Approaches.
- Corpus-Based Approaches.

II.RELATED WORK

Shidaganti, *etal.* [2] tweeter data was analyzed to collect the opinions of the users . Tweeter is mostly used for voicing their views bysmall note in reply to many types of merchandises, brands and celebrities along with political criticisms. Clustering and TF-IDF process was used.

Mumtaz, *etal.* [3] proposed machine learning based approach,which gave more accuracy than the existing lexical method .

Onam Bharti, et.al [4] given a prologue to this interesting issue and to display a system which will perform slant examination on online cell phone audits by partner changed K implies calculation with Naïve Bayes characterization and KNN. We got a general characterization precision of 91% on the test set of 500 portable audits. The running time of the calculation was O (n + V log V) for preparing and O (n) for testing, where n is the quantity of words in the records (direct) and V the size of the diminished vocabulary. It is a lot quicker than other AI calculations like Naïve Bayes characterization or Support Vector Machines which set aside a long effort to meet to the ideal arrangement of loads. The precision was similar to that of the present cutting edge calculations utilized for supposition arrangement on portable surveys.

Bo, Pang, et.al [5]considered the issue of collection reports not by topic, anyway by as a rule estimation, e.g., choosing if a review is sure or negative. Using film studies as data, we find that standard AI procedures convincingly beat human-conveyed baselines. Regardless, the three AI procedures we used Naive Bayes, most outrageous entropy gathering, and support vector machines) don't execute too on assessment portrayal as on standard subject based characterization. The experts shut by taking a assumption at parts that make the presumption gathering issue all the all the more testing.

Ellen Spertus, et.al [6] portrayed a few ways to deal with fire acknowledgment, including a model framework, Smokey. Smokey assembles a 47-elementfeature vector dependent on the linguistic structure and semantics of each sentence, joining the vectors for the sentences inside each message. A preparation set of 720 messages was utilized by Quinlan's C4.5 choice tree generator to decide highlight based guidelines that had the option to accurately order 64% of the blazes and 98% of the non-flares in a different test set of460 messages. Extra strategies for more noteworthy exactness and client customization are likewise examined.

Agarwal, *et al.*[7]did work on semi-structured, structured and unstructured data. Domain-based clustering using cosine and Jaccardsimilarity indexwas used.

Cosine index gave fast results due to steady form in comparison .Jaccardindex was used in composite form of scheme to find the similarity.

Zahrotun, *et al.*[8]examined several clustering techniques.Clustering used for classificationof datawhich usually fits to same class . Cosine, Jaccard similarity and their combination was used for value-based similarity.

Basari,*et al.* [9] discussed the concept of opinions-based mining which referred to the application of computational linguistics, natural language processing (NLP)and mining of the text in order to classify the good or bad movie on the basis of message opinion. SVM mainly presents a supervised method of learning which helps in analyzing the data and recognizes the patterns used for the purpose of classification. Abdul-Mageed, *et al.*[10] presented work on standardized version of Arabic data for the case of sentiment analysis. Here, a set of data was collected and further it performs an automatic classification of step where the process of tokenization was performed over the data. The process of two-stage classification was performed. The results obtained shows that the used approach works in an efficient and effective manner. A. M. Popescu, *et al.*[11] presented a method of unsupervised extraction of information that was used in extracting the review-based opinions. This kind of work was done using the following steps. In the first case, the product-based features were identified and secondly, the opinions associated with the product were identified and in the third case, the opinion’s polarity was identified. Finally, the method proposed was rank and the opinions were based on analysis of strength. The semantically built orientation was obtained using an approach named relaxation-labelling approach. The results based on recall and precision method of the proposed approach effectively shows the effectiveness in identification of sentiments of an individual. X. Song, *et al.*[12]analysed the social-media based usage of the platform through the process of micro blogging and it extracts the informational data from them. The experts have analysed that the users of the system mainly use the platform of social media for updating their routine-based tasks and to know what is used by them on daily basis. Such type of analysis was based on distinct type of micro blogs platforms in distinct kind of geographical area based on time.

[15] proposed a deep learning based approach for Medical Diagnosis. In this paper Function Approximation (FA) is modified by Convolutional Neural Network (CNN).[16] proposed A Novel Weighted Class Based Clustering for Medical Diagnostic Interface.[17] proposed Gaussian Kernel Approximation for Medical Diagnostic Interface.

III.MACHINE LEARNING TECHNIQUES FORTWITTER SENTIMENT ANALYSIS

A. K-Nearest Neighbors (KNN)[18]

K-Nearest NeighborTechnique (KNN) is used for classification. This classification of data is performed on basis of the majority voting of the Neighbors[18].

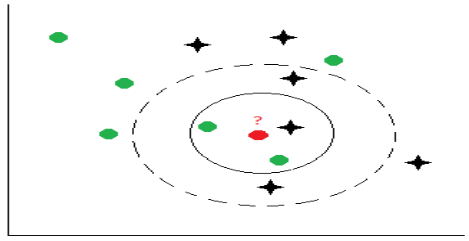


Fig. 2: KNN Used for Classification

Process of KNN explained in Fig2 wherea red input point wants to be classified. Then according to KNN, if value of K is1 then this point will be allocated to class of star which is closest one. If the value of k is 3 then the class of data pointwill be green circle.

B. Decision Tree

Decision tree algorithm is used forpartitioningthe training set is into smaller subsets. These subsets further partitioned recursively as thetree is being built [18]. Decision tree algorithm uses information gain as its traitfor partition. This selection method is based on Claude Shannon information theory. Suppose N represents the records of partition. The elementhavingmaximum information gain in this partition is chosen as the splitting attribute. This elementreduces the information needed to categorize the records in the partition. This attribute givestheminimum randomness in the partition.This process of attribute selection and randomness minimization is repeatedrecursively. The expected information needed to classify a record D is given by:

$$Info\ (D) = -\sum p_i \log_2 (p_i) Where\ i = 1, 2..................n$$

Where p_i is the probability with which recordbelongs toD with class C_i. Info (D) is the average amount of information required to recognize the class label of a record in D.

C. Artificial Neural Network

ANNs architecture is also called multi-layer perceptron (MLP) with back-propagation . Fig 3 [18] shows MLP feed forward Neural Network where Input layer used for providing the input .Input layerfurther provides data to the next layeri.e hidden layer. Hidden layer accepts and process the data received from the first layer (Input layer). There may many hidden layers and many neurons in each hidden layer for resolving the particular task. From hidden layer, recordsisgiven to the output layer. After this, data obtained from output layer and target values are matched. Ifthere is a difference between these two,then the weights areattuned at each processing element. This process is reiterated until target and output values are matched or error reduces to the required limit [18].

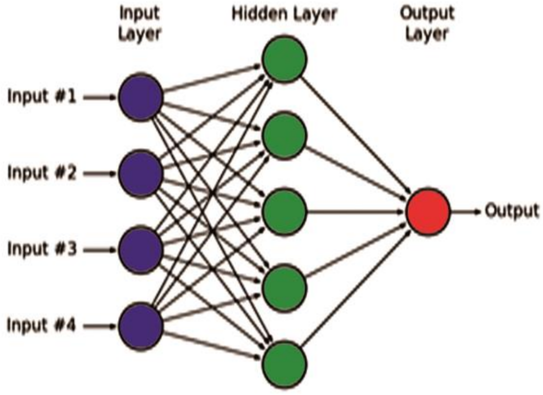


Fig. 3: Multi-Layer Perceptron

D. Support vector machine

Support Vector Machine (SVM) is a technique used for classification of data. A mapping technique is used to convert the data into a higher dimensions[18]. N-dimensional hyper plane used to separate the data into two categories by SVM.The optimal hyper plane separates the data in such a way that data with one class of the target variable is on one side of the plane and data with the other category is on the other side of the plane. The data points close to the hyper plane are called support vectors. The Fig 4 [18] gives the SVM process.

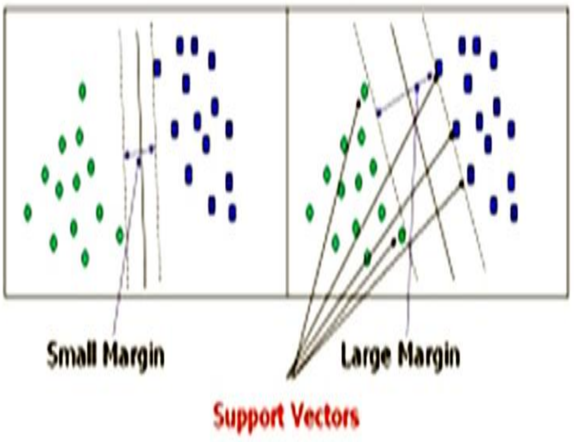


Fig. 4: SVM Topology

E. Naïve Bayes method

Naïve Bayes is used for classification. Itshows the probabilistic relationship among a set of random variables and their conditional dependences. It also gives joint probability of the distribution. Fig 5 network illustrates the relationships among the random variable (season) of the given year (*X1*), If it is raining then its value is *X2*, if the sprinkler then its value is *X3*, if pavement is wet then its value is *X4*, and if the pavement is slippery its value is *X5*. There is no direct link between *X1* and *X5*.It means that there is no direct influence of season on slipperiness. The influence is intervened by the wetness of the pavement.

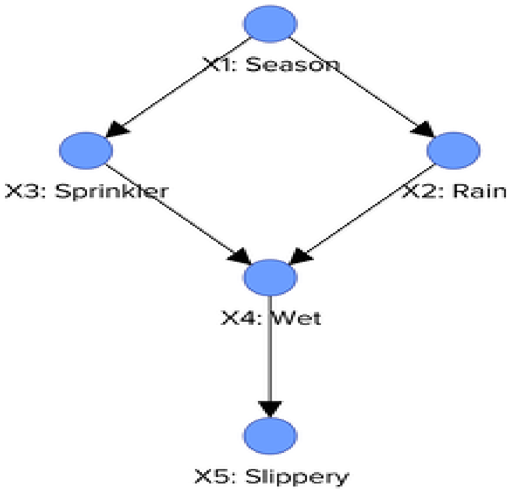


Fig. 5: Naïve bayes method

IV.RESULTS AND ANALYSIS

A. Twitter Datasets

The word "micro" in microblogging. This defines the limitation of the content of the view expressed on it. A Twitter user can make up their message up to 140 characters on a single tweet. Tweetis a combination of text data and metadata. It is not a simple text message. These attributes are the characteristics of tweets and used to express tweet content . The metadata can be used to discover the tweet domain. Tweet metadata is some entities and places. These entities contain mention of users, hashtags, URLs and multimedia users, user ID on Twitter. RT means retweets, "@" trailed by a user ID, signals the user, and "#" trailed by a word, shows a hashtag.

B. Pre-processing of the datasets

This step executes the mechanism of text preprocessing.Text is divided into tokens and then it removes the stopwords like .,?,# etc. and later on itutilizes stemming process for the reduction of noise and every word is defined as a root word.

C. Performance Measures and Results

Confusion matrix is used to explainclassification results. In this paper confusion matrixis shown below in Table 1 The upper left cornershows thequantity of data points classified as sensitiveas they called True Positive (TP), and the lower right cell shows the quantity of data points classified as non-sensitiveas they called True Negative(TN). The remaining columns show the quantity of misclassified data points.

Table 1: Confusion Matrix

	Classified as sensitive	Classified as not sensitive
Actuallysensitive	TP	FN
Actual not sensitive	FP	TN

Below formulae were used to calculate accuracy, precision and recall [18]:

Sensitivity = TP / (TP + FN)

Specificity = TN / (TN + FP)

Accuracy = (TP + TN) / (TP + FP + TN + FN)

Precision= TP / (TP + FP)

Recall= TP / (TP + FN)

A measure that is used to combine precision and recall is called F Measure and it is the Harmonic mean of precision and recall.

F Measure= 2(Precision* Recall)/(Precision+ Recall)

In Table 2 five machine learning techniques(K Nearest Neighbors (KNN), Decision Tree, Artificial Neural Networks (ANNs), Naïve Bayes and Support Vector Machine (SVM))are evaluated on basis of Accuracy, Precision, Recall and F-score. As results shown in Fig 6,Accuracy of SVM is 4% to 32% higher than others.Consequently models have good predictive capabilities then the others. SVM has attained precision value 94.34% which is 5% to 30% higheramong all techniques.SVM has attained recall value 92.34% which is 8% to 33% higheramong all techniques which implies there is agreater degree of agreement betweenactual and predicted class and model is stable.

F Measure values of Decision Tree, MLP and SVM are78.45%, 85.78% and 92.34% respectively.F Measure for SVM nearly equal to 100%. Itshows that there is excellentrelation between actual and predicted class values. .

Table.2 Classification methods based on Accuracy, Precision, Recall, and F-Measure

Classification Method	Accuracy %	Precision %	Recall %	F-score %
KNN	60.34	60.25	61.34	60.75
Naïve Bayes	69.16	65.56	65.45	65.45
Decision Tree	78.34	78.45	79.34	78.45
MLP	88.34	85.45	86.34	85.78
SVM	92.34	90.12	94.34	92.34

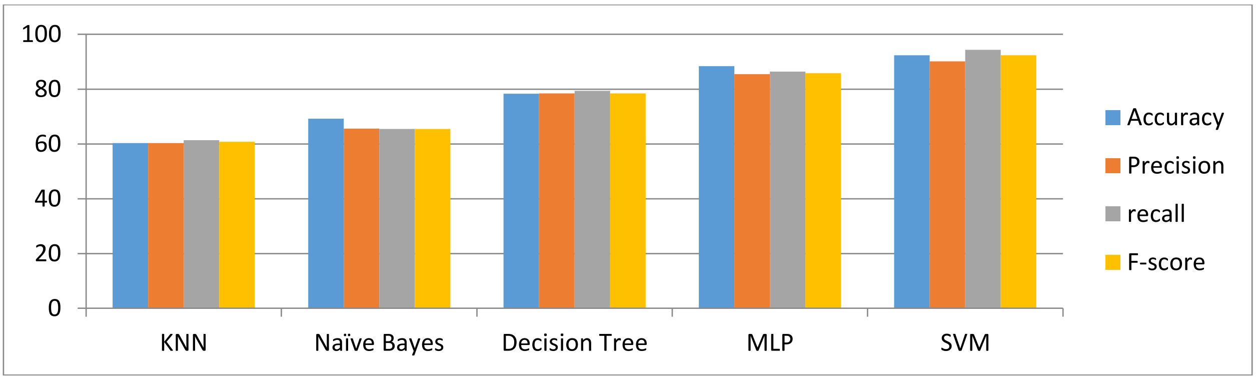


Fig. 6: Comparison of various performance measures

V.CONCLUSION

In this paper five machine learning techniques, K-Nearest Neighbors (KNN), Decision Tree, Artificial neural networks (ANNs), Naïve bayes and Support Vector Machine (SVM) are used for Opinion Mining. These Techniques are analyzed on Twitter dataset. Performances of these methods are compared through various performance metrics such as accuracy, precision, recall and F measure. Results showthat SVM outperformsothers.Consequently models have good predictive capabilities then the others. All the metricsAccuracy, Precision, Recall and F-scoreare high in case of SVM. Our Analysis shows that SVM techniquecomes out to be most outstanding classifier for Opinion Mining.

REFRENCES

1. M. M. Fouad, T. F. Gharib, and A. S. Mashat, “Efficient Twitter Sentiment Analysis System with Feature Selection and Classifier Ensemble,” in *The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2018)*, vol. 723, pp. 516–527, 2018.
2. G. Shidaganti, R. G. Hulkund, and S. Prakash, “Analysis and Exploitation of Twitter Data Using Machine Learning Techniques,” in *International Proceedings on Advances in Soft Computing, Intelligent Systems and Applications*, vol. 628, pp. 135–146, Springer Singapore, 2018.

3. D. Mumtaz and B. Ahuja, “A Lexical and Machine Learning-Based Hybrid System for Sentiment Analysis”, *Innovations in Computational Intelligence*,vol. 713, pp. 165–175, Springer Singapore, 2018.
4. Bharti, O., &Malhotra, M. M. ,” SENTIMENT ANALYSIS ON TWITTER DATA”,*International Journal of Computer Science and Mobile Computing*, Vol.5 Issue.6, June- 2016, pg. 601-609
5. Pang, B., Lee, L., &S.Vaithyanathan, “Thumbs up? : Sentiment classification using machine learning techniques”, *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*(pp. 79-86). Association for Computational Linguistics.
6. Spertus, “ Smokey: Automatic recognition of hostile messages”, In *Aaai/iaai* (pp. 1058-1065),1997.
7. N. Agarwal, M. Rawat, and M. Vijay, “Comparative Analysis Of Jaccard Coefficient and Cosine Similarity for Web Document Similarity Measure,” *Int. J. Adv. Res. Eng. Technol.*, vol. 2, no. 5, pp. 18–21, 2014
8. L. Zahrotun, “Comparison Jaccard similarity, Cosine Similarity and Combined Both of the Data Clustering With Shared Nearest Neighbor Method,” *Computer Engineering and Applications Journal*, vol. 5, no. 11, pp. 2252–4274, 2016.
9. A. S. H. Basari, B. Hussin, I. G. P. Ananta, and J. Zeniarja, “Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization,” *Procedia Engineering*, vol. 53, pp. 453–462, 2013.
10. M. Abdul-Mageed, M. T. Diab, and M. Korayem, “Subjectivity and Sentiment Analysis of Modern Standard Arabic,” *Proc. 49th Annu. Meet. Assoc. Comput. Linguist. Hum. Lang. Technol.*, vol. 27, no. 1, pp. 587–591, 2011.
11. Yang, Y., & Eisenstein, J. ,”Overcoming language variation in sentiment analysis with social attention”,*Transactions of the Association for Computational Linguistics*, *5*, 295-307,2017.
12. A. Java, X. Song, T. Finin, and B. Tseng, “Why We Twitter: Understanding Microblogging Usage and Communities,” in *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis - WebKDD/SNA-KDD ’07*, San Jose, California, pp. 56–65, 2007.
13. Madhoushi, Z., Hamdan, A. R., &Zainudin, “Sentiment analysis techniques in recent works”, In *2015 Science and Information Conference (SAI)* (pp. 288-291). IEEE,2015.
14. Yujiao, L., &Fleyeh, H. ,” Twitter Sentiment Analysis of New IKEA Stores Using Machine Learning”, In *International Conference on Computer and Applications*. (pp. 4-11). IEEE,2018.
15. Godara S., Singh R and Kumar Sanjeev ,” Function Approximation with Kernel Approximation by Convolutional Neural Network (CNN) for Medical Diagnosis”,*Ciência e TécnicaVitivinícola, Vol. 34 ,21-3,2019.*
16. Godara S., Singh R and Kumar Sanjeev ,” A Novel Weighted Class Based Clustering for Medical Diagnostic Interface”,*Indian Journal of Science and Technology*,Vol9(44),2016.
17. Godara S., Singh R and Kumar Sanjeev,” Gaussian Kernel Approximation for Medical Diagnostic Interface”,*Jour of Adv Research in Dynamical & Control Systems*, Vol. 10(44),2018.
18. SunilaGodara , Rishipal Singh, “ Evaluation of Predictive Machine Learning Techniques as Expert Systems in Medical Diagnosis”, *Indian Journal of Science and Technology*,Vol 9, Issue 10, March 2016.

AUTHORS PROFILE

Dr. Sanjeev Kumar, did his M.Tech and Ph.D in Comp. Sc. &Engg from Guru Jambheshwar University of Science and Technology Hisar(Haryana).He is working as Associate Professor in CSE deptt. GJUS&T,Hisar.He had more than 50 papers in his account.His area of interest are machine learning, computer networks and cloud computing.

Nirmal Godara, is doing her Ph.D D in Comp. Sc. &Engg from Guru Jambheshwar University of Science and Technology Hisar(Haryana). Her area of interest are machine learning and cloud computing.