International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol. 9, Issue.9, September 2021 E-ISSN: 2347-2693

Movies Reviews Sentiment Analysis using Improved Random Forest Algorithm and ACO (Ant Colony Optimization) Approach

N.K. Deol[1*], V. Thapar[2], J. Singh[2]

Dept. of Information Technology, Guru Nanak Dev Engineering college, Ludhiana, India
Dept. of Computer Science and Engineering, Guru Nanak Dev Engineering college, Ludhiana, India

**Corresponding Author: navdeepdeol4@gmail.com*

DOI: https://doi.org/10.26438/ijcse/v9i9.2530 | Available online at: www.ijcseonline.org

Received: 05/Sept/2021, Accepted: 20/Sept/2021, Published: 30/Sept/2021

Abstract: Data mining, text mining and opinion mining have occurred in one form or another since modern record keeping began. As the number of online shopping users is increasing, access to social media sites produces vast quantities of information in the form of user feedback, comments, blogs and tweets tests. For this reason, Sentimental analysis is required, which classifies these reviews to gain insights into the data generated by the user. The main problem with the analysis of the feeling is the uncertain mood of the user, such that the interpretation of what the user has written and what he actually thought is somewhat different. The problem analysed in the existing work is that the decision-making trees, particularly when a tree is very large, are likely to parallelize. Random forest classification is used to eliminate both errors due to bias and variance. In the proposed research, the improved technology is implemented with Random forest and optimization of the Ant colony search is hybridised with the proposed classifier in order to accomplish the classification of film screens by studying the sentiments.

Keywords: Sentiment Analysis, Social Media, Movie Reviews, Data Mining

I. INTRODUCTION

Social media is the online exchange of ideas, information and business interests by various communities such as group of students, political parties, etc. Websites such as Twitter, Facebook and WhatsApp are free to connect with. This communication probably involves friendship, families, group relations and romantic relationships. Social networking allows people to establish personal connections and make new friends and thus helps the individual to have the sense of connectedness with their closed ones. Because of huge number of people linked to networking sites, there are growing numbers of connections. Social media functionality.

Incorporated into a single website include: user groups, the latest information regarding music groups, video and photo sites, forums, a personal profile etc. Social media sites also help people to retain and grow business connections. LinkedIn is a best example to discuss business and to meet professional people. Hence, it's very hard to interpret and analyse the huge volume of data from all these applications, which also consumes our time. To reduce the workload, Sentiment Analysis is used to interpret the views from different sources. The opinion of a human about any product is important because humans are subjective in nature and they show their level of satisfaction with any product or service, which helps in enhancing the quality of the product, improving marketing strategies and satisfying the customers according to their need. A movie review reflect the idea of the writer that may be positive, negative or neutral and based on that review the other people get the impression of the movie. So, it can be said that on the basis of its reviews a movie will be considered hit or flop.

Sentiment Analysis can be defined as, taking views and emotions of people on any subject of interest [1]. Sentiment research is carried out in natural language processing (NLP). The data mining, web mining and information processing are also studied extensively. In recent years, technologies for sentiment analysis have reopened to virtually all possible fields. Analysis of feelings and classifying feelings into categories based on feeling polarity. Polarity means that two

Opposing feelings, views and aspects are present. The polarity may be positive, negative or neutral. There may be a fourth form- a constructive one in which sentiment holders make recommendations for change.

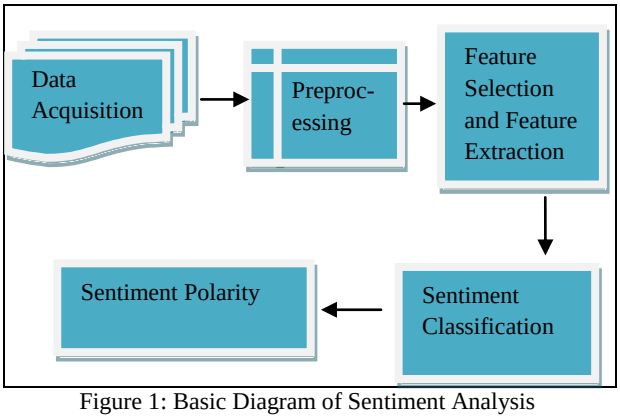

Figure 1: Basic Diagram of Sentiment Analysis

Sentiment Analysis consists of branches of engineering technological know-how like linguistic communiqué system, Machine Learning, Text Mining and Information Theory and Coding. By the usage of methods, strategies and fashions of described branches, we will categorize our data into fine, bad or neutral sentiment in step with the sentiment that is expressed in them.

In this paper, Improved Random Forest algorithm is hybridized with ACO (Ant Colony optimization) search based feature selection for improving the classification of movie reviews. As the problem analysed in the existing work is that the decision-making trees, particularly when a tree is very large, are likely to parallelize which results in error due to bias and variance. To reduce these both errors we have proposed an improved Random Forest classification algorithm [2].

II. LITERATURE REVIEW

Sentiment analysis is often called as opinion mining, a kind of data mining methodology that tracks and analyses people's humours [3]. In the field of following up people's opinions about a product, a service, an organisation, individual or an event, with different techniques and instruments, great research has been done.

Tripathi et al. (2016), examined sentiments on the Indian film review corpus using machine learning classification. This research used the Bayesian Classifier for the testing of mechanical feature quality. This classification focused on the terms / functions of the corpus that were extracted with the use of five algorithms for collection of functions (Chi-square, Info-gain, Gain-Ratio, One-R and relief). Relief F was noted to score higher than the other feature-selection algorithms examined in the study and to give the best 88.8 percent F-value accuracy due to other feature selection methods, again Relief F algorithms performed better with a 61.6 percent F-value for features that have been examined in the study [4].

Cagatay CATAL et al. (2017), explored the potential advantage of the idea of multi-classifying systems on the Turkish question of sentiment classification and suggested a new classification technique. The voting algorithm was used in tandem with the three classifiers Naive Bayes, SVM and Bagging. When used as a separate classifier, SVM parameters were optimized. Experimental findings have shown that the performance of individual classification systems in Turkish sentiment classification datasets is improved through the multiple classification systems and meta-classifiers helped to improve the power of these multiple classification systems [5].

Naik et al. (2017), aimed to automate the cycle of on-line collectability, end-user feedback and interpretation of those opinions expressed on specific features for any particular product or service. This involved the pre-processing of rare, inapplicable data, the classification of such data into different groups and finally the description of the data on the expressions of feelings. KNN was better for text processing than naïve Bayes [6].

Wankhede et al. (2017), concentrated sentimental analysis on the database for "Indian times." They examined the feelings of a polarized (positive / negative / neutral) grouping of film reviews in the article. The Random Forest classifier was also used for performance evaluation and accuracy assessment. They have reached the highest accuracy of 90% by using Random Forest Classification Technique [7].

Pandey et al. (2018), the paper suggested tests in sentiment analysis which improved updated methods of negative recognition than current methods. In the case of negation of the data supplied, which was stored in the document, it was fed into a vector. The use of dependency parses and prefix algorithms was handled by both syntactic and morphological negations. In this paper, 92% of the research has been successful [8].

Nanda et al. (2018), explored a Sentiment Analysis for film reviews in Hindi. The method suggested divided the exams in Hindi into two classes: positive and negative, and also assessed the results through different metrics of working on the algorithm. Machine learning, which is commonly used with many benefits over all other methodologies, represented a classification method used. 91.07 percent by current method (Random Forest) was the best accuracy accomplished [9].

Dholpuria et al. (2018), illustrated the research on deep learning model by using Convolutional Neural Networks (CNN) with respect to supervised machine learning classifiers (Naive Bayes, SVM, Logistic Regression, KNN and Ensemble Methods). The work proposed through deep learning model concluded that in terms of accuracy (99.330544 %) it's giving reliable performance with the large dataset. Since every day new data in the form of tweets, comments is being posted so it's very important to analyse the classifier on a large dataset. The improvement in classification model accuracy through CNN classifier is presented as comparative analysis of their performance [10].

Yin et al. (2018), meanwhile, the vector analysis approach for word vector is applicable to several different languages, for instance Chinese and English. In addition, this paper addressed the effect of the word vector dimensions on the accuracy of the feeling analysis and the implementation of the method on phrases of different lengths. The test result has shown that the word vector-based analysis technique is an efficient and simple way not only to evaluate emotional speech but also to be extensible and accessible in various lengths and multiple languages for comment. Thus the precision of classification of 86.18 percent using this technique was achieved [11].

Hourrane et al. (2019), reviewed the state of the art to determine how the previous researches have addressed this

task. Here, they also introduced an empirical study on two annotated datasets; 25,000 IMDB movie reviews and 25,000 tweets, where used nine supervised learning models, the next step was to implement a voting ensemble classifier, got using the top four models from the previous steps. The experimental result and the ROC curve gave 90.76% accuracy on IMDB movie review data, 76.88% accuracy on the Twitter dataset [12].

Yasen et al. (2019), the purpose of this research was to deal with SA by developing a methodology that can classify film reviews and then compare the results in an extensive analysis of eight well-known classifiers. The IMDB evaluations of the actual dataset were used to test the proposed model. The tokenization of the dataset was applied to move strings in word vector, and stemming was used to remove the root of the words, and then the data set gain ratio has been used as a selection algorithm for the attributes. The data was subsequently divided into training and research data sets by 66%, 34% respectively. In contrast to all other classifiers, Random forest got the highest accuracy (96.01 percent). Moreover, it got the highest precision (0.93), f-measure (0.96) and AUC (0.96) [13].

Tiwari et al. (2020), defined the idea of a detailed analysis of feelings through a widely popular Twitter platform, which is recognised and used worldwide. All reached the conclusion after knowing the people's opinions on any matter. Latest research has taken Natural Language Processing (NLP) into consideration to do this form of analysis. The precision of Random Forest (99.4%) and Decision Tree (99.3%) was more precise than the SVM analytical method [14].

III. PROPOSED METHODOLOGY

The objective of our work is sentiment analysis by fusing Improved Random forest algorithm and search feature selection based on ACO (Ant Colony Optimization), to improve the film reviews classification.

A. Collection of datasets: The data has been collected from the **IMDb** website with positive, negative and neutral sentiments. Movie reviews are collected online from the website with their sentiment scale. Then according to the scale, reviews are divided into positive, negative and neutral.

B. Pre-Processing and Filtering: Some strategies for filtering the raw materials in a standardized format will be implemented for pre-processing the data. Pre-processing involves several phases, including tokenization, stop-word deletion and case normalization.
Pre-processing includes 3 stages:

a. Data Cleaning: Cleaning of data involves handling of missing values by ignoring that particular tuple. If any tuple or cell is empty then that will be filled with some specific value. Inconsistency of data may be handled manually. It also handles noisy data by implementing machine inspection, clustering, binning methods and regression. All the quotes ("") from the sentences are removed, URL's are removed and other characters that are not considered to be in the category of texts are removed.

b. Data integration: Data is always collected from various sources like data warehouse, internet etc. so, the collected data in particular is of no use and it has to be added altogether for further analysis. So, this step will integrate the data collected from various sources.

c. Data transformation: Transformation of data means to change the data from one form to another. For this purpose, various methods like smoothing, normalization, aggregation and generalization are available for the transformation. Transformation steps are as follows:

- **Sentence Splitting:** The first step involved is sentence splitting i.e. the splitting of string into words. Identifying sentence boundaries in a document is not a smaller task.
- **Tokenization:** Tokenization of words means to split a sentence into tokens or smallest unit of a sentence. Tokenization is an important task because many succeeding components need tokens clearly identified for analysis.
- **Stop Word Filtering:** There are a lot of words that do not have any meaning and can be removed from the input file. Words like "the", "and", "for", "or", "if", "that"; are referred to as stop words because they don't signify any meaning or sentiment. Therefore, removal of such words means stop word filtering and it also improves the performance of the system.
- **Stemming:** A stemming calculation is a procedure of semantic normalization. In this process, the variations of a word are lessened to a typical frame. For instance, consider a simple example below:

  Communication
  Communications
  Communicative ——> Communicate
  Communicated
  Communicating

C. Ant colony search-based Feature selection

Initialization: Initially, population of ants and intensity of pheromone trail associated with any feature is determined. Moreover, maximum number of allowed iterations is defined.

Heuristic Desirability: In ACO algorithm, constructive heuristic is a basic requirement for probabilistically constructing solutions. A solution construction is empty in the beginning and solutions are assembled as sequences of elements from finite set of solution components. After that, a feasible solution component is added to the current partial solution at each construction step. Heuristic desirability of choosing between features could be any subset evaluation function. In the proposed algorithm, CFS (Co-Relation based Feature Selection) subset evaluation is used as heuristic desirability.

Update Pheromone: Unlike, Ant Colony system, in this approach only one single ant which is best-so-far ant is

allowed to deposit pheromone and update pheromone trails.

Solution Construction: The overall process of feature selection begins by generating a number of ants which are then placed randomly on the graph i.e. each ant starts with one random feature. From these initial positions, they traverse nodes probabilistically until a traversal stopping criterion is satisfied. The resulting subsets are gathered and then evaluated. If an optimal subset has been found or the algorithm has executed a certain number of times, then the process halts and outputs the best feature subset encountered. If none of these conditions hold, then the pheromone is updated, a new set of ants are created and the process iterates once more.

D. Classification using Improved Random Forest

The classification process is the most important process in sentiment analysis. Random forest algorithm is an ensemble-based learning algorithm which is capable of performing both regression and classification tasks. It is combined with improved ensemble technique which works on the principle of error rate of model and misclassified instances are more focused in each iteration to make them correctly classified based on the weightage assigned to each model. In improved technique, random forest is used as a weak learner technique. Random forest contains number of decision trees constructed by this method at training time and return the output of class which is the mode or mean prediction of individual trees. It also carries out dimensional reduction method and treats the missing values and outliers. The basic principle behind this random forest algorithm is that a 'strong learner' can be formed by the group of 'weak learners'. Accurate classifiers and regressors are also generated by introducing right kind of randomness. Mostly simple decision trees have high bias and high variance but random forests resolve the problems of high bias and variance by finding average between two extremes. Random forests algorithms are easy to learn and also more accurate predictions can be made without making basic mistakes common to other methods.

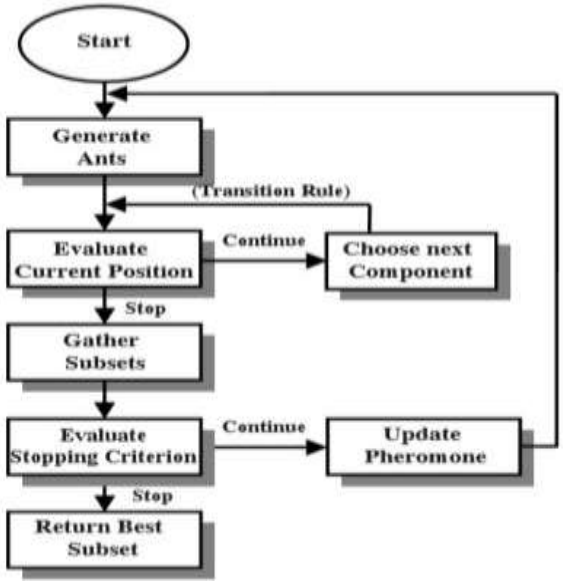

Figure 2: ACO based Feature Selection

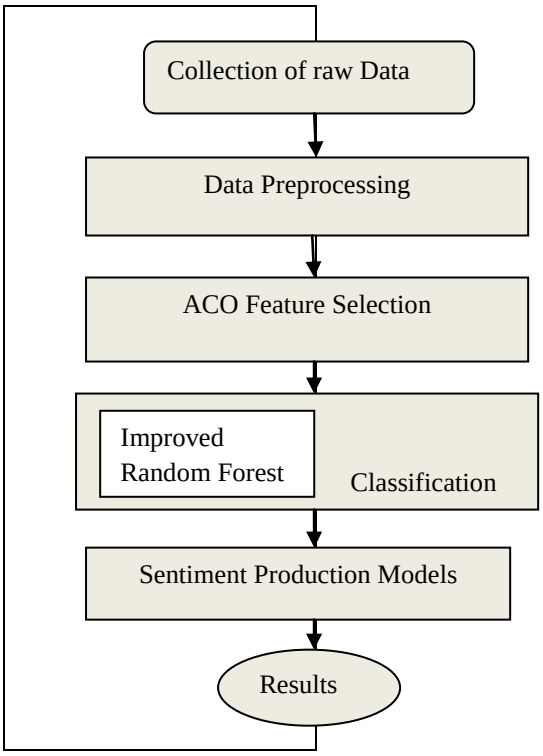

Figure 3: Flowchart of proposed methodology

IV. EXPERIMENTAL RESULTS

There are various words in the definitions of the evaluation. The words are real, real negative, false and false positive. These are the concepts used to distinguish class marks chosen with documents from the classes with which things actually have a place. True positive terms are truly positive terms delegated. False positives are not called a positive class, but should have been identified by the classifier. Genuine negative words are generally referred to by the classifier as in the negative class. False negative terms are those concepts that are not labelled as having a negative class place, but should still be ordered by the classifier. Disarray Matrix contains the words used for evaluation. Evaluation parameters are:

Accuracy: Accuracy is the basic criterion for the implementation of arrangements. It is the closeness of a calculated value to the standard or true value.

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative}$$

Table 1: Accuracy Comparison

S.No	Techniques	Accuracy
1.	KNN	62.64%
2.	Decision Tree	76.60%
3.	Naïve Bayes	83.40%
4.	Random Forest	85.91%
5.	IRF with ACO FS	96.60%

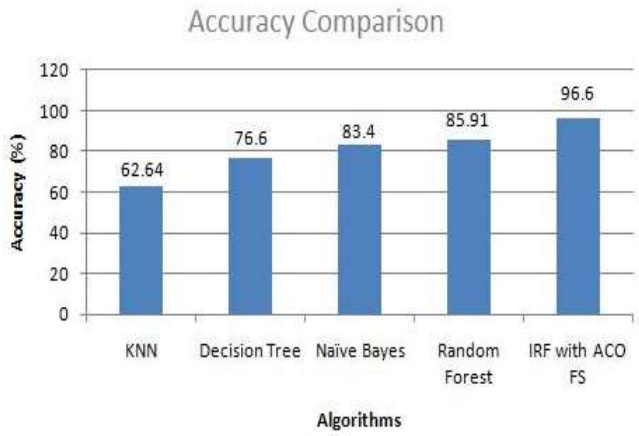

Figure 4: Accuracy comparison

When we compared the Accuracy of all the algorithms, then our proposed Improved Random Forest algorithm achieved the maximum accuracy of 96.6038% whereas the KNN got the minimum.

Precision: Precision is the fragment of correct positives classifications (true positives) from cases that are predicted as positive.

$$Precision = \frac{True\ Positive}{True\ Positive\ +\ False\ Negative}$$

Table 2: Precision Comparison

S.No	Techniques	Precision
1.	KNN	0.601
2.	Decision Tree	0.749
3.	Naïve bayes	0.863
4.	Random Forest	0.842
5.	IRF with ACO FS	0.966

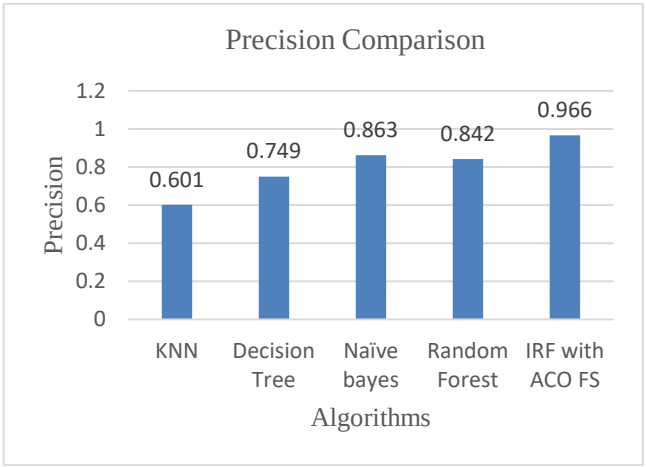

Figure 5: Precision comparison

Recall:
It is the fraction of total number of relevant instances that were actually retrieved.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Table 3: Recall Comparison

S.No	Techniques	Recall
1.	KNN	0.626
2.	Decision Tree	0.766
3.	Naïve bayes	0.834
4.	Random Forest	0.859
5.	IRF with ACO FS	0.966

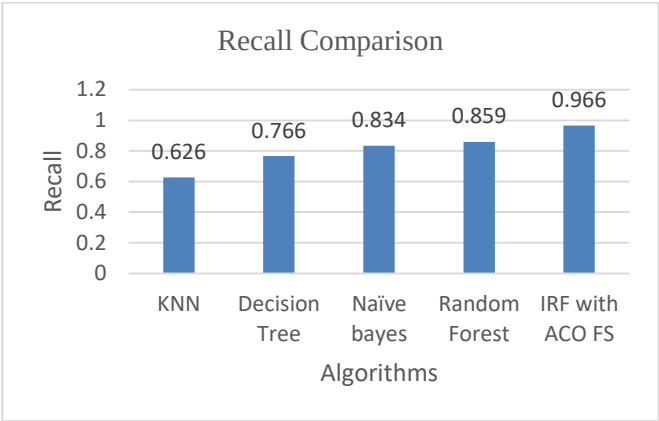

Figure 6: Recall comparison

F- Measure: It is the measure of a test's accuracy and is defined as the weighted harmonic mean of the precision and recall of the test.

$$F - Measure = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Table 5: F Measure Comparison

S.No	Techniques	F Measure
1.	KNN	0.585
2.	Decision Tree	0.753
3.	Naïve bayes	0.846
4.	Random Forest	0.823
5.	IRF with ACO FS	0.94

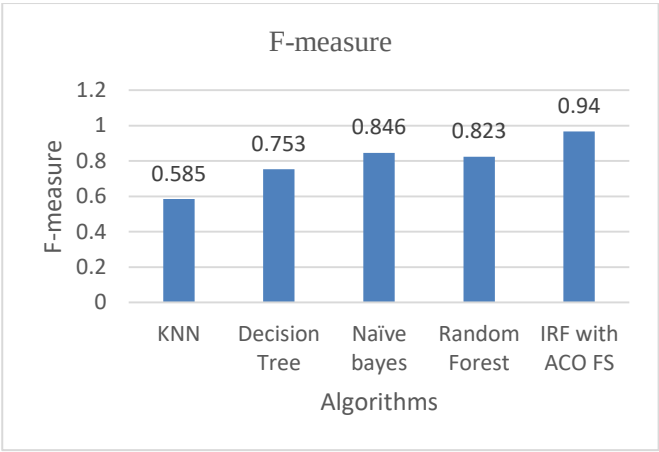

Figure 7: F-Measure comparison

The above findings demonstrate that with regard to the above mentioned parameters, the suggested work has better Accuracy, Precision, Recall and F-Measure. The comparison has been carried out in several instances and the results show that the proposed work is better than the Random forest technique (finest so far).

V. CONCLUSION AND FUTURE SCOPE

The social media companies that should be trained to analyse the viewpoint of individuals in the fields of goods, objects, film evaluation etc. produces a number of types of knowledge. On social media platforms like Twitter and Facebook there are millions of users. In addition to e-commerce websites in social media, have numerous users. Sentiment analysis using various types of feedback will provide new insights into the business model that the individual businesses adopt and will improve the competitiveness of the organization. The biggest issue with the analysis of the sentiment is the uncertain mood of the user, such that the interpretation of what the user has written and what the user actually thought is somewhat different. The information problem is the numerous qualities of the attributes. It is unilateral to choose properties with a wide range of qualities. Individual classification algorithms were used in the existing paper to assume that random forest is the finest of all other algorithms. In the proposed study, improved methodology is used with the random forest, rather than the use of the random forest, and the optimisation of the Ant colony search is hybridized with the proposed classifier in order to achieve the classification of the sentiment of research by film reviews. The results indicate that on the basis of accuracy and class parameters, the proposed technique is superior to the current techniques.

In future, for Movies Reviews we can work with number of clustering techniques to find out more accurate results and we can also try for different languages of Movies Reviews apart from English and can work with some other Data Mining applications like News, Product Reviews and so on.

REFERENCES

[1] C. Ouyang, L. Yongbin, Z. Shuqing, and Y. Xiaohua, "Features-level Sentiment Analysis of Movie reviews", Advance Science and Technology Letters, pp. **110-113, 2016**.

[2] S. Kumar Yadav, "*Sentiment Analysis and Classification: A Survey*", International Journal of Advance Research in Computer Science and Management Studies, Vol. **3**, Issue. **3**, **2015**.

[3] R. Baldania, "Sentiment analysis approaches for movie reviews forecasting: A survey", International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), **2017**.

[4] A. Tripathi, S. K. Trivedi, "Sentiment analysis of Indian movie review with various feature selection techniques", IEEE International Conference on Advances in Computer Applications (ICACA), **2016**.

[5] C. Catal, M. Nangir, "*A Sentiment Classification Model Based On Multiple Classifiers*", Applied Soft Computing Elsevier, vol. **50**, pp. 135–141, **2017**.

[6] K. Naik, A. Joshi, P. Khanna, N. Shekokar, "A Model to Analyse Social Media Data to Gain a Competitive Edge", International Conference on Computing, Communication, Control and Automation (ICCUBEA), **2017.**

[7] R. Wankhede, A. N. Thakare, "Design approach for accuracy in movies reviews using sentiment analysis", International Conference of Electronics, Communication and Aerospace Technology (ICECA), **2017**.

[8] S. Pandey, S. Sagnika, B. S. P Mishra, "A Technique to Handle Negation in Sentiment Analysis on Movie Reviews", International Conference on Communication and Signal Processing (ICCSP), **2018**.

[9] C. Nanda, M. Dua, G. Nanda, "Sentiment Analysis of Movie Reviews in Hindi Language Using Machine Learning", International Conference on Communication and Signal Processing (ICCSP)", **2018**.

[10] T. Dholpuria, Y. Rana, C. Agrawal, "A Sentiment analysis approach through deep learning for a movie review*",* 8th International Conference on Communication Systems and Network Technologies (CSNT), **2018**.

[11] F. Yin, Y. Wang, X. Pan, P. Su, "A Word Vector Based Review Vector method for Sentiment Analysis of Movie Reviews Exploring the Applicability of the Movie Reviews", 3rd International Conference on Computational Intelligence and Applications (ICCIA), **2018**.

[12] O. Hourrane, N. Idrissi, E. H Benlahmar, "Sentiment Classification on Movie Reviews and Twitter: An Experimental Study of Supervised Learning Models", 1st International Conference on Smart Systems and Data Science (ICSSD), **2019**.

[13] M. Yasen, S. Tedmori, "Movies Reviews Sentiment Analysis and Classification", IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), **2019**.

[14] S. Tiwari, A. Verma, P. Garg, D. Bansal, "Social Media Sentiment Analysis on Twitter Datasets", 6th International Conference on Advanced Computing and Communication Systems (ICACCS), **2020**.

AUTHORS PROFILE

Ms. N.K. Deol is currently working as an Assistant Professor in the Department of Information Technology at Guru Nanak Dev Engineering College, Ludhiana. Her area of interest is Data Mining.

Dr. T. Vivek is currently working as an Assistant Professor in the Department of Computer Science and Engineering at Guru Nanak Dev Engineering College, Ludhiana. His Research interests include Network Security and Web Technologies.

Mr J. Singh is currently working as an Assistant Professor in the Department of Computer Science and Engineering at Guru Nanak Dev Engineering College, Ludhiana. His Research interests include Web Technologies, User Interface Design and Cloud computing.