IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol. 10, Issue.1, January 2022 E-ISSN: 2347-2693

IMDb Movie Data Classification using Voting Classifier for Sentiment Analysis

Karishma Kaushik[1*], Mahesh Parmar[2]

[1,2]Department of Computer Science, Madhav Institute of Technology and Science, Gwalior India

[*]*Corresponding Author: karismasagargwl2019@gmail.com,*

DOI: https://doi.org/10.26438/ijcse/v10i1.1823 | Available online at: www.ijcseonline.org

Received: 29/Jan/2022, Accepted: 24/Jan/2022, Published: 31/Jan/2022

***Abstract*—** Social networking sites have become popular and common places in which short texts share emotional diversity. These emotions are sadness, happiness, fear, anxiety, and so on. In order to identify sentiments expressed by the crowd, it helps in analyzing short texts. On IMDb movie reviews, sentiment analysis identifies a reviewer's overall sentiment or opinion on a movie. We worked on the IMDb movie dataset in this paper. which was retrieved from Kaggle which was crawled and labelled positive/negative. The available dataset consists of emoticons, Id, Data, Query, username and converted into a standard from. We get these results by utilizing a Voting Classifier with Logistic Regression & Random Forest, which is a traditional machine learning algorithm. Furthermore, the results of these algorithms were compared using five evaluation criteria. metrics – accuracy(89.34),precision(88.71),recall(90.35), F1 measure(89.52),and Area under Curve (89.33).

***Keywords*—** Sentiment Analysis, Feature Extraction, Voting classifier, Machine Learning, IMDb data**.**

I. INTRODUCTION

The internet has been converted into a massive database in the digital era. Internet users voice their thoughts on a variety of platforms, including, Facebook, Twitter, and IMDb. These opinions are quite useful in determining the social community's general preferences. Every day, a large number of films are released. The choice to see a movie is based on its reviews. A substantial portion of movie reviews are polarity and contain a lot of sentiment-based language, whether positive or negative. In Sentiment Analysis, these terms might be quite useful. sentiment Analysis is "the mechanism behind sequence of terms that determine emotional tone used for interpreting attitudes, thoughts & emotions expressed in an online mention"[1].

When movie review data is exposed to Sentiment Analysis, it is possible to determine whether the movie has a higher number of negative and positive reviews. Tajindersingh et al [2] Used the SVM classifier to analyze the sentiment of tweets and discovered that effective text data pre-processing enhanced accuracy. Isha Gandhi et al. [3] introduced a hybrid ensemble classifier that integrates instance-based learner representative algorithms, Naive Bayes, and Decision Tree Algorithms with voting techniques, concluding that ensemble provided the best accuracy. Bin Lu et al. [4] showed that when they integrated a sentiment lexicon with an SVM classifier for opinion analysis, the lexicon classification outperformed the individual machine learning approaches using SVM. Zamahsyari et al[5]. To analyze the economic news in Bahama, a majority vote ensemble was built using Random Forest, Decision Tree, and Naive Bayes. A number of learning-based classifiers, including Decision tree, Naive Bayes, and SVM, were used to evaluate the big movie review dataset [6] from IMDb in this study. Finally, a majority vote classifier is used to merge the learning-based classifiers into an ensemble. Ensemble is an efficient method combining different learning algorithms to increase overall accuracy of predictions. The accuracy of each classifier's output is compared. This research discovered that the ensemble classifier outperforms individual classifiers and outperforms the ensemble that includes the random forest as one of the classifiers.

The following is the structure of a paper: The article's second section discusses related work in the SA field. The proposed and deployed approaches are shown in the third section. The fourth section shows the results of the proposed approach, and the fifth section concludes the research and discusses future work.

II. LITERATURE REVIEW

Kumar, V., et al. [1] The proposed framework using a new methodology to tokenize text documents which exclude stopping terms, specific characters, and emoticons in text documents. Moreover, terms with common definitions & annotation are grouped into 1 type using the stemming concept. A tokenized documentation preprocessed would then be vectorized into n-gram integral vectors using 'TfidfVectorizer' as an input to the SVM-based Machine learning classifier model.

Mahmud, Q. I., et al. [2] It differences between positive & negative feedback using the SVM to predict movie performance using statistical reasons. For our sentiment classifier, they used a non-linear RBF kernel, It outperformed classifiers using the linear kernel in the popularly IMDB Movie Review data set (89.51 percent accuracy) as well as the Pang & Lee Movie Review Dataset (89.51 percent accuracy) (86.86 percent)accuracy.

Xu. G. et al. [3] In this work extended dictionary of sentiments is created . The expanded dictionary of sentiments includes essential terms of sentiments, the words for the field sentiments & polysemic words for sentimental which increased the accuracy of the sentimental analysis. In a naive Bayesian classifier, the area in which the polysemic word of sentiment is described. Thus, in the area is obtained the sentiment value of term polysemic sentiment. The sentiment of the text is obtained by the use of the expanded sentiment dictionary & built sentiment score rules.

Sahu, T. P., et al. [4] Various classification methods were carried out to evaluate the most appropriate classifier for our problem area. They conclude that our suggested technique for the classification of sentiments supplements current web-based rating systems for film rating & is the basis for future research in this field. "The better accuracy of our strategy is 88.95 % using classification methods."

Hourrane, O., et al. [5] Based on the IMDB movie review data, we proposed a comparative analysis, compare word embedding approaches & more deep learning models in sentimental analyses & offer broad empirical results for those who want to use deep learning to analyze sentiments in real-world environments.

Manjunath, D. R et al. [6] The proposed data pre-processing methods first clean the dataset, then apply hierarchical clustering to selected features, followed by regression-based attribute classification to clusters based on negative and positive user evaluations. Finally, the accuracy of the proposed hierarchical clustering and regression technique is compared with that of the existing decision tree approach, and the accuracy of the suggested approach is determined.

III. PROPOSED WORK

A. Problem Identification

A fact that several demanding & fascinating research issues remain to be resolved in this area makes this sentimental classification area more difficult. Sentimental analysis of documents in comparison with topic-based text classification is very difficult to do. Situations are often different in words & sentiments. Consequently, the term 'opinion' may in one circumstance be seen as positive, but in another circumstance may become negative.

A Large Movie Review Dataset (often called the IMDB dataset) has 25,000 intensely polar moving reviews (good or bad) training & the same number for testing again. An issue is determined whether there is positive or negative sentiment in particular moving review.

B. Methodology

In this proposed work, first of all, the movie text data has been collected from the gaggles. After that preprocessed the data then done the feature extraction using the bigram method, bigram method is used for counting. After that used voting classifier for classifying the sentiments into positive or negative than in voting classifier Logistic Regression (LR) and Random Forest (RF) is used.

1) Pre Processing

Data in IMDb must be normalized to construct a dataset that different classifiers can benefit from. Data pre-processing is performed to standardize & reduce the size of the dataset. Training & testing was performed using pre-processed dataset. Tweets have some unique features like rewets, emoticons, user mentions, etc.

2) CountVectorizer

The Countvectorizer of Scikit-learn is used to convert a corpora text into a vector of words. It can also preprocess text data before converting it to a vector representation, making it a very scalable text representation module.

In this paper, we will examine in depth how you can use CountVectorizer so that you not only compute the count but also preprocess the text data properly & extract additional features from your test dataset.

3) Bigram

We have used the bigram method in CountVectorizer. The bigram is a sequence of 2 adjacent elements, usually letters, syllables, or words, from a string of tokens. The bigram for n = 2 is an n-gram. Many applications, such as voice recognition, computational linguistics, cryptography, and so on, depend on the frequency distribution of each bigram. for the basic statistical analysis of the text.

Gappy Bigrams are word pairs that make gaps or bigrams (avoiding linking terms or enabling some simulation of dependency, as in a dependency grammar).

4) Voting Classifier

the voting classifier is a technique that uses many prediction classifiers. It varies whether a data scientist or machine learning engineer is unsure about the classification technique to use. Therefore, the voting classifier makes predictions based on the most common one using predictions from other classifiers. For this Voting classifier, we have to use Logistic Regression (LR) and Random Forest (RF) classifier that is described below.

A. Logistic Regression (LR)

The Logistic Regression model is a set of classification rules that are commonly used to handle binary classification, however it has to be modified to handle multi-class classification. This logistic classifier program takes a set of weighted features from input and determines the correlation between the event and the generated features. Allison [13] states that the logistic regression adapts to the suitable fit data by maximizing the log-likelihood function. In general, it leads to weak prediction for all predictors in one model. A correct selection of variables makes the model more precise & more general.

Logistic regression may be used to calculate the probability of feature vector I having a positive class:

$$P\ (c = 1|\ f) = \quad l(f) = \frac{1}{1+e^{wT}} \qquad (1)$$

The probability of document ‘f' of class ‘c' is indicated by p (c = 1|f). The letter ‘w' stands for feature-weight parameters that need to be approximated.

B. Random Forest (RF)

Random forest (RF) was first introduced by Leo Breiman from the University of California in2001. It is composed of several simple classifiers (decision trees) which are independent of each other. A sample will be included in the new classifier and the class label of this sample depending on the voting outcomes from every single classification. [14].

The main steps for the RF classifier are as follows:

1. Set the proper & “M” value, which is the number of components of each sub-set of features.
2. Choose a new subset of feature hk from the entire feature set randomly based on the value of M. hk is free from another subset in h1; ...; hk sequence.
3. Training the data set for each training category with the feature sub-set to construct a decision tree. Every single category can be represented as h(X, hk) (where X specifies thein puts).
4. Select a new hk and repeat it until all the feature subsets are moving. An RF classifier has been achieved.
5. Input the test set. Decide based on voting outcomes for each classification of this sample.

Proposed Algorithm

Step 1. Input the IMDb dataset.

Step 2. Pre – processing the Dataset. The noisy data or special characters are removed by stemming, lemmatization, stop words removal, etc.

Step 3. Convert the pre-process dataset into 70% Training set and 30% Testing set.

Step 4. Apply Bigram Feature Extraction method to calculate count of the word.

Step 5. Voting classification has been applied on the Extraction feature to perform classification techniques for classification

Step 6. Stop.

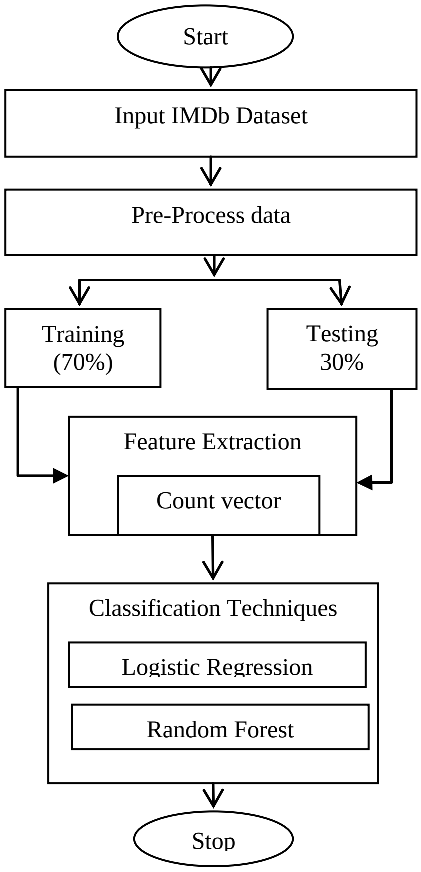


Figure 1. Proposed algorithm flowchart

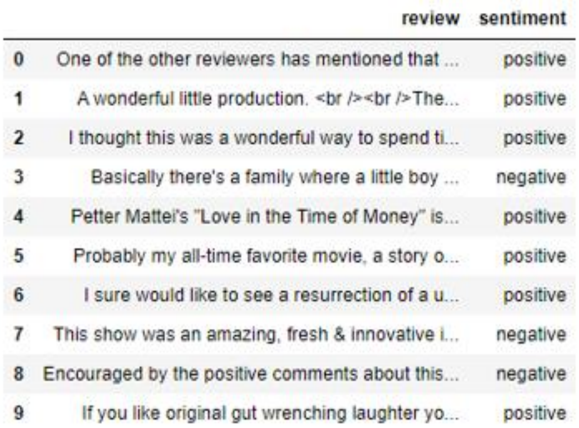
IV. Results and Discussion

This work has implemented using python programming language and platform used is python (version 6.3.1). Here, we have used IMDb datasets for perform the experiment. The description of such datasets and achieved results of the proposed model has given below.

A. Dataset description

The Large Movie Review dataset on IMDb has 50,000 reviews, evenly split over 25000 training and 25000 test datasets. There are a total of 25000 positive and negative opinions. There are additionally 50,000 data sets that are unlabeled. Only the tagged data are considered for analysis in this study. There are no more than 30 reviews for any any movie because this could affect the classification's overall result.

Moreover, train & test sets include disjoint series of films, such that the memorization & their association with observed labels do not achieve substantial performances.[15]. There are two names train & test directories which compare training & test datasets. Every list also has two additional directories, "pos" and "neg," which include the relevant reviews. These evaluations are collected as text files.



	review	sentiment
0	One of the other reviewers has mentioned that ...	positive
1	A wonderful little production. <br /><br />The...	positive
2	I thought this was a wonderful way to spend ti...	positive
3	Basically there's a family where a little boy ...	negative
4	Petter Mattei's "Love in the Time of Money" is...	positive
5	Probably my all-time favorite movie, a story o...	positive
6	I sure would like to see a resurrection of a u...	positive
7	This show was an amazing, fresh & innovative i...	negative
8	Encouraged by the positive comments about this...	negative
9	If you like original gut wrenching laughter yo...	positive

Figure 2. Sample of dataset

B. Evaluation Parameters

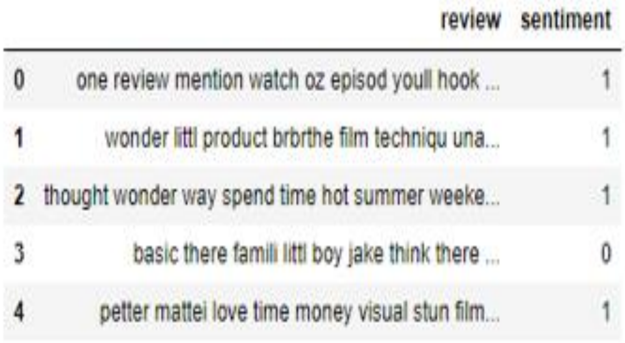
Term or confusion matrix components can be used to evaluate the performance of supervision ML algorithms on a set of test data. 'True positive' (TP), 'False positive' (FP), 'True negative' (TN), and 'False Negative' (FN) are the four terms. The Precision, Recall, F-Score, and ROC test matrices are used to calculate any classifier's output score based on the values of these components.

A ROC curve is a graphical plot that shows the classification model performance at all thresholds, taking into consideration parameters True positive (TF) on the X-axis & True negatives on the y-axis.

Precision (π): $\frac{TP}{TP+FP}$ (2)

Recall (ρ): $\frac{TP}{TP+FN}$ (3)

F- Score: $\frac{2*Precision*Recall}{Precision+Recall}$ (4)



	review	sentiment
0	one review mention watch oz episod youll hook ...	1
1	wonder littl product brbrthe film techniqu una...	1
2	thought wonder way spend time hot summer weeke...	1
3	basic there famili littl boy jake think there ...	0
4	petter mattei love time money visual stun film...	1

Figure 3. Pre-Process dataset

Results of the proposed Methodology

This subsection visualizes all simulated results using the proposed classifier IMDb dataset. It also represents the accuracy, precision, recall, and F1- measure for datasets.

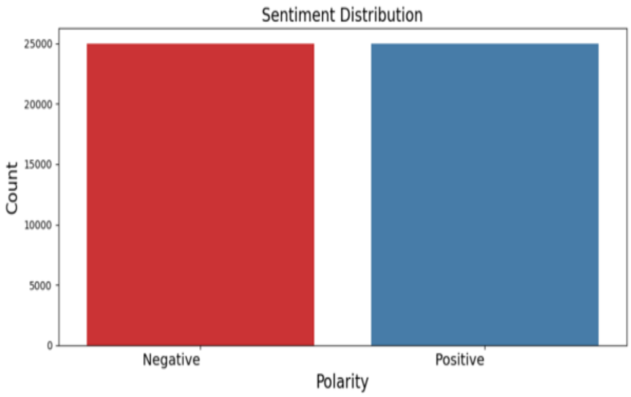


Figure 4. Sentiment Distribution in IMDb Dataset

From the output, we can see that the majority of the tweets are negative and positive tweets.

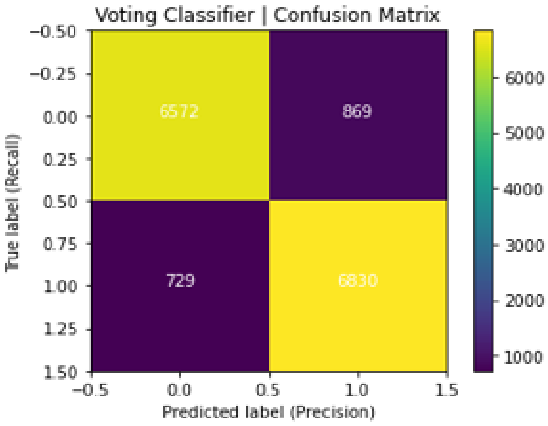


Figure 5. ConfusionMatrix

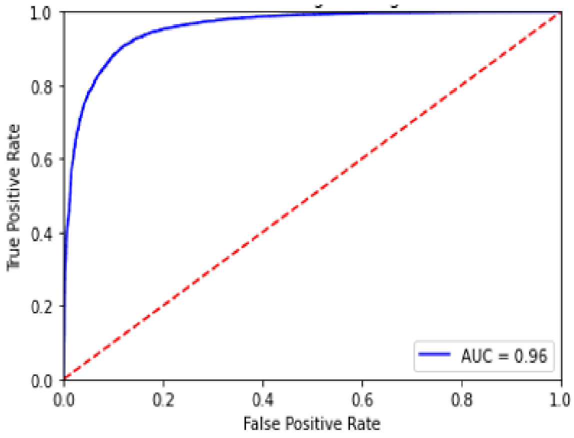


Figure 6. Receiver operating characteristic curve for Voting Classifier

Table 1: The comparison of Performance Parameters

Parameters	**Logistic Regression**	**Voting Classifier (LR+RF)**
Accuracy	88.41	89.34
Precision	87.22	88.71
Recall	89.86	90.35

F1-measure	88.52	89.52
AUC	88.42	89.33

The comparison of Performance Parameters between the Existing Logistic Regression classification technique and the proposed voting classifier technique has been shown in Table 1. The accuracy of the voting classifier (Logistic Regression + Random Forest) is89.34 which is greater than the logistic regression which has an accuracy of only 88.41.

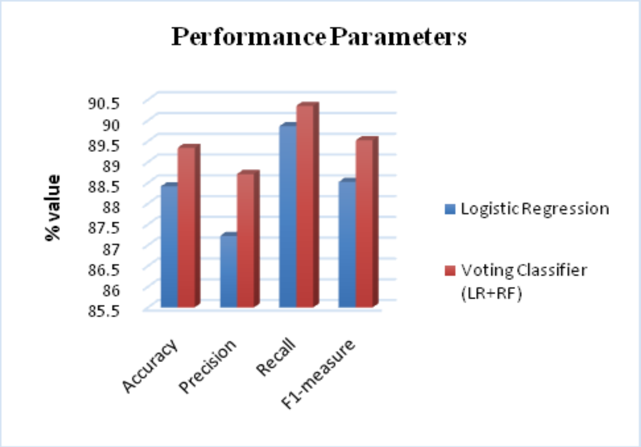


Figure 7. Comparative graphical representation between Existing technique and over proposed technique for IMDb Dataset

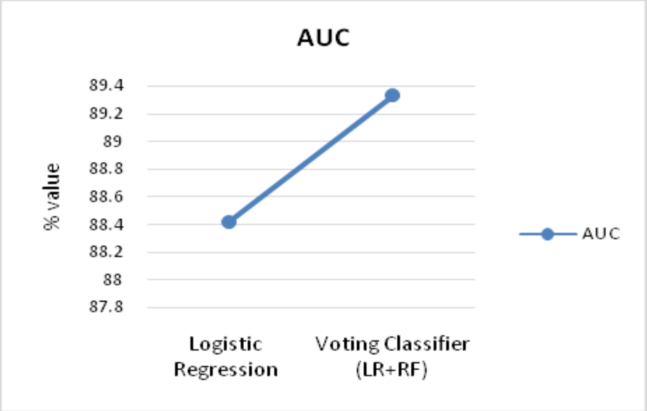


Figure.8. Comparative Line graph representation between Existing technique and over proposed technique for IMDb Dataset

V. CONCLUSION AND FUTURE SCOPE

This article summarizes the study & classification of different movie reviews from numerous film applications. This article applies several techniques for analyzing film reviews such as data preprocessing. This paper presents Voting Classifier Technique based on Logistic Regression and Random Forest to analyze the reviews taken from various Movie-based applications. We have performed SA an IMDb movie review dataset taken from Kaggle's which was crawled and labeled positive/negative. Then, using the Extraction Feature in the data set, classification is done based on the Movie reviews dataset. When classifying, we evaluated the total number of comments and ratings of users on the movies. As a result, we can achieve an accuracy of 89.34% using the proposed approach.. A Part of the accuracy Parameter we have compared these proposed techniques on different Parameters like Precision, Recall, F1-measure, AUC and we have found that over-proposed technique outperformance in comparison base technique.

Although the evolution of the relatively unreliable system is difficult to speculate, the majority agree that sentimental analysis must go beyond a positive or a negative level. in the future, A more sophisticated, multidimensional dimension is required to better appreciate & capture the full spectrum of human emotions represented in words. This heading would also be subject to further effort

REFERENCES

[1] Tajinder singh, Madhu Kumari, "Role of Text Pre-Processing in Twitter Sentiment Analysis", Procedia Computer Science 89 (2016), pp.549-554.

[2] Isha Gandhi, Mrinal Pandey, "Hybrid Ensemble of Classifiers using Voting", Green Computing and Internet of Things (ICGCIoT), 2015, DOI: 10.1109/ICGCIoT.2015.7380496.

[3] Bin Lu, K.T. Benjamin," Combining A Large Sentiment Lexicon And Machine Learning For Subjectivity Classification", Machine Learning and Cybernetics (ICMLC), 2010,DOI: 10.1109/ ICMLC.2010.5580672.

[4] Zamahsyari, Arif Nurwidyantoro, "Sentiment Analysis of Economic News in Bahasa Indonesia Using Majority Vote Classifier", Data and Software Engineering (ICoDSE), 2016, DOI:10.1109/ICODSE.2016.7936123.

[5] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, Christopher Pos, "Learning Word Vectors for Sentiment Analysis",2011 In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Portland, Oregon, USA, 142–150. http://www. aclweb.org/anthology/P11-1015

[6] Kumar, V., & Subba, B. (2020). A TfidfVectorizer and SVM based sentiment analysis framework for text data corpus. 2020 National Conference on Communications (NCC). doi:10.1109 /ncc48643.2020.9056085

[7] Mahmud, Q. I., Mohaimen, A., Islam, M. S., & Marium-E-Jannat. (2017). A support vector machine mixed with statistical reasoning approach to predict movie success by analyzing public sentiments. 2017 20th International Conference of Computer and Information Technology (ICCIT). doi:10.1109/iccitechn.2017.8281803

[8] Xu, G., Yu, Z., Yao, H., Li, F., Meng, Y., & Wu, X. (2019). Chinese Text Sentiment Analysis Based on Extended Sentiment Dictionary. IEEE Access, 7, 43749–43762. doi:10.1109 /access.2019.2907772

[9] Sahu, T. P., & Ahuja, S. (2016). Sentiment analysis of movie reviews: A study on feature selection & classification algorithms. 2016 International Conference on Microelectronics, Computing and Communications (MicroCom). doi:10.1109/microcom.2016.7522583

[10] Hourrane, O., Idrissi, N., & Benlahmar, E. H. (2019). An Empirical Study of Deep Neural Networks Models for Sentiment Classification on Movie Reviews. 2019 1st International Conference on Smart Systems and Data Science (ICSSD). doi:10.1109/icssd47982.2019.9003171

[11] Manjunath, D. R., & Hadimani, B. S. (2019). Hierarchical Clustering and Regression Classification based Review analysis on Movie based Applications. 2019 1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE). doi:10.1109/icatiece45860.2019.9063861

[12] Gladence L, Karthi M, Anu V. A statistical comparison of logistic regression and different Bayes classification methods for machine learning. ARPN J Eng Appl Sci. 2015;10(14):5947–53.

[13] Parmar, A., Katariya, R., &amp; Patel, V. (2018). A Review on Random Forest: An EnsembleClassifier. Lecture Notes on Data Engineering and Communications Technologies,758–763. doi:10.1007/978-3-030-03146-6_86.

[14] https://github.com/jalbertbowden/large-movie-reviews-dataset/tree/master/acl-imdb-v1

AUTHORS PROFILE

Mis. Karishma kaushik pursued a Bachelor of engineering from the University of Madhav institute of science, in 2016. I am currently pursuing a M.TECH main research work focuses on Data mining and Image Processing.

Mr. Mahesh Parmar as an Assistant Professor in CSE&IT Department in MITS Gwalior and having 10 years of Academic and Professional experience. He received M.E. degree in Computer Engineering from SGSITS Indore. He has guided several students at Master and Under Graduate level. His areas of current research include Data mining and Image Processing. He has published more than 30 research papers in the journals and conferences of international repute. He has also published 02 book chapters. He is having the memberships of various Academic/ Scientific societies including IETE, CSI, and IET etc.