International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-2, December, 2019

Text Classification of Cornell Movie Data using Data Mining with Feature Selection

A. K. Shrivas, S. M. Ghosh, Amit Kumar Dewangan

Abstract: ***Text Classification is branch of text mining through which we can analyze the sentiment of the movie data. In this research paper we have applied different preprocessing techniques to reduce the features from cornell movie data set. We have also applied the Correlation-based feature subset selection and chi-square feature selection technique for gathering most valuable words of each category in text mining processes. The new cornell movie data set formed after applying the preprocessing steps and feature selection techniques. We have classified the cornell movie data as positive or negative using various classifiers like Support Vector Machine (SVM), Multilayer Perceptron (MLP), Naive Bayes (NB), Bays Net (BN) and Random Forest (RF) classifier. We have also compared the classification accuracy among classifiers and achieved better accuracy i. e. 87% in case of SVM classifier with reduced number of features. The suggested classifier can be useful in opinion of movie review, analysis of any blog and documents etc.***

Keywords: Classification, Cornell Movie Dataset, Feature Selection, and WEKA Tool.

I. INTRODUCTION

In this present digitalization era, all the human being are dependent on machines and it is also called as computer age. All the information is saving as digital data like text, audio, video etc. The information [1] is covered in broad fields such as educational information, opinion and feedback of products quality, view about social issues etc where text categorization plays very important and critical issues. It helps people to think and take decision in many real time problems. The most of people always taken final decision from others opinion [1]. Sentimental analysis, opinion mining and text classifications are the broad research area where we can extract useful knowledge. Sentimental analysis is not that much easy task as name suggest. It had taken number of preprocessing steps to get the classification of the cornell movie data from the data set. [17]. We have discussed the prominent related research being carried out in the area of sentimental analysis and text mining. R. Ullah et al. (2015) [2] have the understanding of emotional content and significant implications in word of mouth for online retailers and consumers is discussed. They have used natural language processing (NLP) technique for positive or negative bag of word sand Coefficient correlation is used to measure the dependence between a term polarity range. M. Bilal et al. (2016) [3] have suggested Naıve Bayesian Decision Tree and KNN for text classification using WEKA environment. The opinions have written written in Roman, Urdu and English and extract from blog. S. Liao et al.(2017) [4] have discussed about the sentiment analysis of twitter data base using deep learning technique. They have used convolution neural network (CNN) for analysis of twitter data and it is also useful for extracting information from a larger piece of text, hence sentiment analysis with convolution neural network. They have suggested CNN technique outperforms SVM and Naive Bayes methods. P. Baidet al.(2017) [7] have also suggested Naïve Bayes, K-Nearest Neighbor and Random Forest for analyzing the polarity of the tweets about movie reviews and compared the performance in terms of accuracy. K.M.A. Kumar et al.(2015) [21] have discussed the Sentiment Analysis of the regional languages like kannada, mainly spoken in Karnataka. They have used the machine learning and semantic approaches for sentimental analysis and text classification on English language data set, from Kannada web documents. N. B. Khanna et al. (2018) [9] have suggested a new softmax based attitude detection approach for analyzing nature of users. The suggested approach is experimented on tweets collected from micro blogging website twitter. S. Zhang et al. (2018) [10] have presented a sentiment analysis method for Chinese micro-blog text that is depend on sentiment dictionary to support network regulators. M. Kang et al.(2018) [11] have suggested a novel sentiment analysis technique as text-based hidden Markov models(TextHMMs) for text classification. J. Chambua et al.(2018)[12] have proposed tenor factorization model which complements the factor's central function for identifying similar users, unwrap the fundamental features of the data, and forecasting user priority using text semantic similarity. M. Malik et al. (2018) [13] have worked on customers reviews. Random reviews were collected for the purpose of sentiment analysis. The opinion polarity was calculated using the weight method. X. Li et al. (2018) [14] have proposed a Common Semantic Subject Model (CSTM) for text analysis. The proposed CSTM given better result than existing short text subject model. L. Wang et al. (2018) [15] have proposed a cross-domain sentiment classification technique to analyze sentiment polarity for short texts.N. Öztürk et al. (2018) [16] have investigated public opinion and feelings of Syrian refugee crisis which influenced by millions of people worldwide through social media site.

Revised Manuscript Received on December 30, 2019.
* Correspondence Author
A. K. Shrivas*, Department of Information Technology, Dr. C.V. Raman University, Bilaspur (C.G.), India.
S. M. Ghosh, Department of Computer Science Engineering, Dr. C.V. Raman University, Bilaspur (C.G.), India.
Amit Kumar Dewangan, Department of Computer Science Engineering, Dr. C.V. Raman University, Bilaspur (C.G.), India.

He has analyzed public sentiments about the topic on twitter, relevant tweets in two languages, including Turkish and English. He has also analyzed the comparative sentiment of retrieved tweets.

II. PROPOSED ARCHITECTURE

In this section, we are exploring the proposed architecture of data preprocessing and text classification.

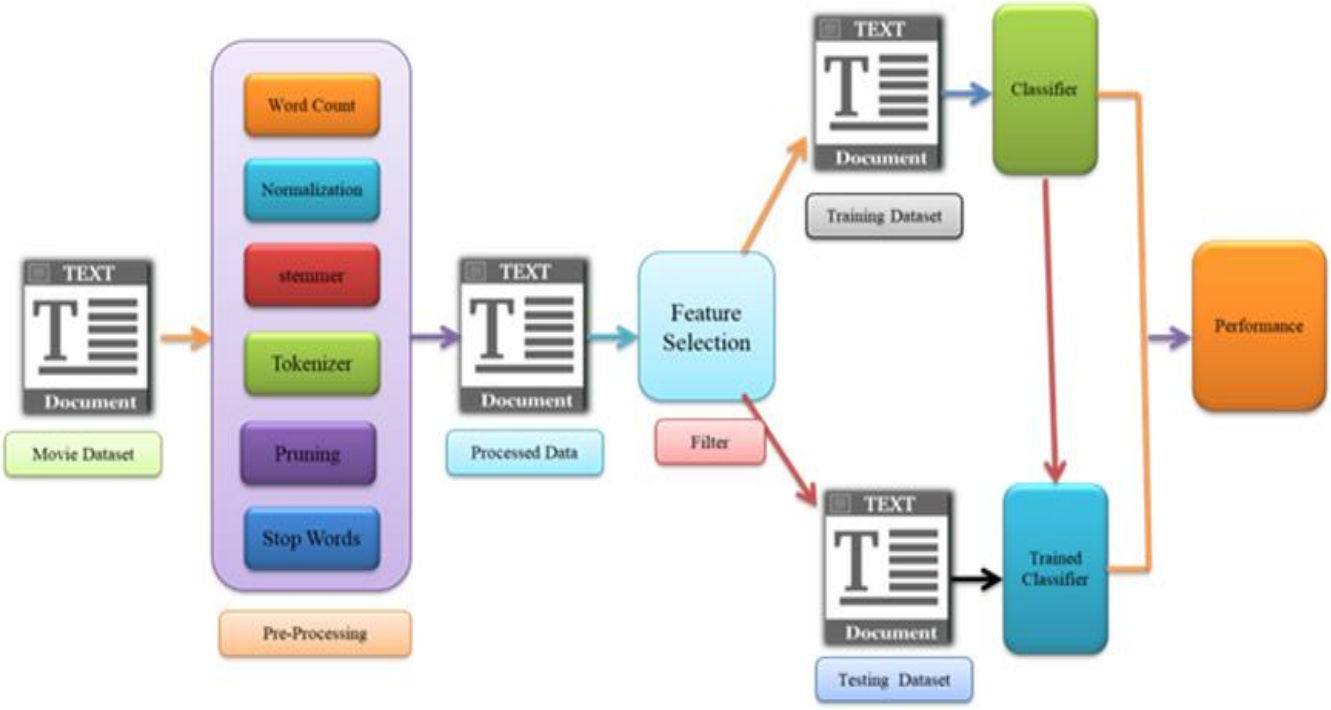

Figure 1. Proposed model for text classification

The proposed architecture for data preprocessing and text classification as shown in Figure 1. The proposed architecture is divided in six sections: Data set collection, pre-process the data set, reduced data set with feature selection, partition of data set into training and testing, training and testing of classifiers and compare the performance of classifiers. In first section, we have collected cornell movie sentiment polarity dataset version 2.0from cornell.edu web site [5]. In second section applied the various preprocessing operation like stemmer, tokenizer, stopwords remover and pruning on the cornell movie data set to remove the noise and inconsistent data and prepare the smooth dataset. In third section, we have used the different filter techniques like ChiSquaredAttributeEval with Ranker and usingCfsSubsetEval with Best Fit to extract the feature from dataset and select important features from the data set. In fourth section, partitioned the data set into training and testing with different data partitions. In fifth section, training and testing the classifiers. In last section ,we have compared the classifiers in terms of various performance measures like accuracy, precision, recall, F-measures and ROC curve.

III. DATASET

We have collected dataset from Internet Movie Database .The dataset contains 2000 user-created cornell movie data files on the IMDb (Internet Movie Database)[5].This data set are equally partitioned as positive and negative samples.

IV. DATA PREPROCESSING TECHNIQUES

Preprocessing is the important data filtering techniques for removing the inconsistent and noise data from database. To achieve the preprocessed data, first import the data from database to WEKA data mining tool. WEKA [6] provides a simple import procedure for textual datasets using TextDirectoryLoadercomponent filter. We have also [6] used the StringToWordVector filter from the package weka.filters.unsupervised.attributefor creating positive and negative document into vector form. In this research work we have applied the various preprocessing techniques like word counts, normalization, steaming, tokenizer, pruning and stop word as shown in figure 1.The description of all preprocessing techniques as shown in table 1.

Table 1: Description of Preprocessing Techniques

Preprocessing Techniques	Description
Words Counts	Word Count means total number of words are present in the documents. We calculate this words by using wordsTokeep attribute.
Stremmer	Stemming is the process through which reduce same root of different words. We have used Iterated LovinsStemmerto reduce the same root words.
Tokenizer	The process in which breaking up a statement into pieces such as words, keywords, phrases and symbols are called tokenization. This is done by tokenizer. Tokens are the individual letter, words, phrases or even whole sentences.
Prunning	Pruning is a method in which removing the term which is least suited for discrimination in document.
Stop Words	Stop words are commonly used words in any language like a, an, the, is etc. just like in English language. Stop words are very basic set of words which are necessary but it has not any important meaning in any sentence.
Document Frequency(DF)	Document Frequency means to recognize the frequency of individual words in the document. DF provides the sentiments of identified words. It is also provides the importance of words which is very important across the documents.
Inverse term Document Frequency(ITDF)	IDF stands for inverse document frequency which value or weight depends on frequency of words. The value of IDF increases when frequency of words are less while the the value of IDF decrease when frequency of word is high.

A.Feature Selection

Feature Selection is a process to extract important words or feature from the dataset for constructing the computationally efficient model which achieve better classification accuracy. We have used two feature selection technique namely Correlation-based feature subset selection and Chi Squared.

- **Correlation-based feature subset selection**

The Correlation-based feature subset selection technique is a ranking based feature selection technique which rank the features based on its important. In this feature selection technique, we choose CfsSubsetEval [20] (Correlation-based feature subset selection) evaluation technique combine with BestFit search strategy. The working methodology of the CfsSubsetEval is to choose those attribute which are highly correlated with any class attribute and left the low correlation with other attribute. CfsSubsetEval evaluates the value of a subset of features by considering the degree of redundancy between them as well as the individual predictive capacity of each feature; A subset of features that are highly correlated with class is preferred when there is low inter-correlation. After applying the Correlation-based feature subset selection evaluation technique combination with BestFit search strategy. The number of attributes has measurably reduced.

- **Chi Squared**

WEKA has also provides another combination of attribute selection techniques ChiSquaredAttributeEval evaluation technique combination with Ranker search method.ChiSquaredAttributeEvalmeasures how much expected numbers and observed numbers deviate from each other. This filter computes the chi-squared statistic of each attribute with respect to the class and provide the rank of all attributes.

V. CLASSIFICATION TECHNIQUE AND TOOLS

In this research we have used five classifiers like Support vector machine (SVM), Multilayer Perceptron(MLP), Naive Bayes (NB), Bays Net (BN) and Random Forest (RF) to classify the cornell movie data as positive or negative samples.

A. Support Vector Machine

The support vector machine (SVM) [17] is a classifier to achieve better accuracy in case of text classification. They construct a hyperplane with maximum Euclidean range for maximum exercising cases. The SVM represents cases as factors of the region that are planned for a high-dimensional region where the planned cases of individual sessions are separated by possible tangential range in the form of a hyperplane. New cases are planned in the same area, and depending on what part of the hyperplane they hold, they are expected to fit with a certain range. SVM hyperplanes are established solely by a relatively small fraction of training conditions, known as support vectors. Relaxation of the exercising data has no effect on the qualified classifier. SVM has been implemented efficiently in a large range of text classification and series handling programs.

B. Multi LayerPerception(MLP)

Multilayer Perceptron (MLP) [19] is a type of artificial neural network that is used as classifier for classification task. MLP consist extra hiddenlayers with input and output layers. It contains more than one hidden layers so it is called multilayer perceptron. The MLP structure is formed in which from the input layer to the first hidden layer, from the first hidden layer to the second and so on, to the output layer to the last hidden layer. MLP handle the non-linear data.

C. Naïve Bayes

Naive Bayes (NB) [8] is a supervised learning technique for constructing classifier that assigns class labels to each samples, where class labels are drawn from some finite set. It is based on the Bayes theorem of posterior probability.

D.Bayes Net

Bayes Net [8] is a classifier and based on Bayes theorem. It is also a supervised learning which is classify the samples based on its classes. Bayes Net is useful in that it provides a solution for calculating the posterior probability. Bayes Net[18] is considered a complete model for variables and their relationships. Therefore, a complete joint probability distribution over all variables is specified for a model. In text mining, the computation complexity of Bayes Net is very expensive.

E. Random Forest

Random Forest (RF) [7] is a decision tree technique for classification and regression. It is combination of more than one decision tree. To classify new case it sends the new case to each of the trees. Each tree performs classification and output a class. The output class is chosen based on majority voting that is the maximum number of similar class generated by various trees is considered as the output of the Random Forest. Random forests is combination of decision tree which is easy to learn and useful for both essential research and programming.

F. WEKA Data Mining Tool

WEKA [6] stands for Waikato Environment Knowledge. It is a collection of various data mining algorithms and tools for in depth analysis. This tools is developed using java programming language and GNU provided General Public License. There are mainly three uses of WEKA. First the analysis of data mining algorithm; second for generation of model; and last for comparison of various data mining algorithm to check the robustness of algorithm.

VI. RESULT AND DISCUSSION

The experiment is carried out with window 7 environment using WEKA data mining tool and i3 system. This experiment is divided into two phases. In first phase, we have applied the various preprocessing techniques on cornell movie data while various data mining based classification techniques have applied on preprocessed cornell movie data set in second phase.

Phase I: Data Preprocessing Steps

In this phase, we have applied the five preprocessing steps for filtering the noise of data and reduce the measurable amount of attributes from dataset. We have applied two feature selection techniques like Correlation-based feature subset and chi-square to find the important features that will helpful for computationally improve the performance of classifiers and achieve better accuracy. Table 1 shows that feature reduced from data set using different preprocessing steps. The cornell movie data set contains total 47163 features. We have applied the various preprocessing steps to reduce the features of cornell movie data set. In case of Stremmer, the features are reduced from 47163 to 27573. In case of Tokenizer, the features are reduced from 27573 to 19244. In case of pruning, features are reduced from 19244 to 9012. Similarly, features are reduced from 9012 to 8777 in case of stop words. Finally we have form a reduced cornell movie dataset with 8777 features.

Table 2: Feature reduction using preprocessing steps.

S. No.	Preprocessing Steps	Number of Features
1	Stremmer	27573
2	Tokenizer	19244
3	Prunning	9012
4	Stop Words	8777

The new cornell movie data set with 8777 features has been reduced using two feature selection techniques as shown in table 3. When we have applied the Corelation based Feature Selection Technique then feature is reduced from 8777 to 68 and form a new reduced **cornell movie dataset1**while feature is reduced from 8777 to 334 in case of ChiSquare feature selection technique and form another new **cornell movie data set2**. Figure 2 shows that number of features with different preprocessing steps and feature selection.

Table 3: Feature selection technique

Sr. No.	Feature Selection Technique	Number of Features	Data Set
1	CfsSubsetEval with Best Fit	68	Cornell movie data set1
2	ChiSquaredAttributeEval with Ranker	334	Cornell movie data set1

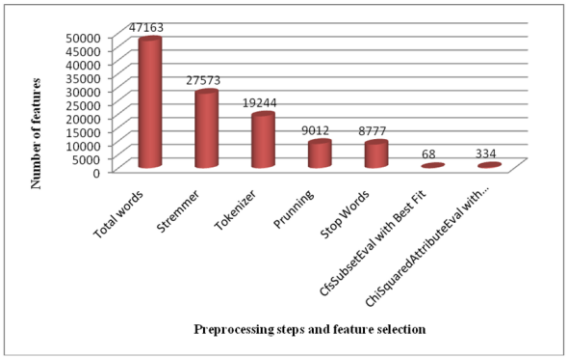

Figure 2. Number of features with different preprocessing steps and feature selection

Phase II: Classification of Cornell Movie Dataset

This research work have used various data mining based classification techniques for classification of positive and negative cornell movie samples. We have applied the preprocessed cornell movie data set on the different classifiers like Support Vector Machine (SVM), Multilayer Perception (MLP), Naive Bayes (NB), Bays Net (BN) and Random Forest (RF) and compared the performance of classifiers to categorize the cornell movie samples as positive or negative. The experimental work done into two sections for classification of cornell movie data set 1 and cornell movie data set 2. Table 4 shows that accuracy of various classification techniques with different partition of dataset like 10 fold cross validation, 70%-30% and 80%-20% of training testing dataset in case of cornell movie data set 1. The accuracy of classifiers are varying from partition to partition as shown in the table 4. The SVM classifier gives better classification accuracy compare to other classification techniques. The SVM gives highest accuracy as 81.75% in case of 80%-20% training testing data partition. Figure 3 shows that graphical representation of classification accuracy with cornell movie data set 1.

Table 4: Accuracy (In %)of classifier with cornell movie data set 1

Model	10 Fold Cross Validation	70%-30%	80%-20%
SVM	**81.55**	**80.83**	**81.75**
NB	79.85	81.5	81.25
BN	79.7	77.5	81.25
RF	78.1	77.83	78.25
MLP	75.9	75	76.75

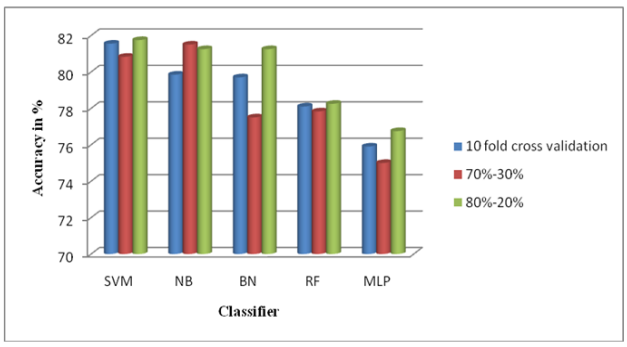

Figure 3. Graphical representation of classifier accuracy with cornell movie data set 1.

In cornell movie data set 2, the accuracy of various classification techniques with different partition of dataset like 10 fold cross validation, 70%-30% and 80%-20% of training testing dataset as shown in table 5 where the accuracy of classifiers are varying from partition to partition. Similarly cornell movie data set 1, the SVM classifier gives better classification accuracy compare to other classification techniques. The SVM gives highest accuracy as 87.00% in case of 80%-20% training testing data partition. Figure 4 shows that graphical representation of classification accuracy with cornell movie data set 2. From phase I and phase II, we conclude that SVM is suggested for cornell movie data classification due to achieve better accuracy compare to other classifiers.

Table 5: Accuracy(In %) of classifier withcornell movie data set 2.

Model	10 Fold Cross Validation	70%-30%	80%-20%
SVM	**84.30**	**86.66**	**87.00**
NB	83.15	82.50	83.75
BN	79.85	77.83	82.25
RF	82.05	81.33	83.50
MLP	81.50	78.50	78.75

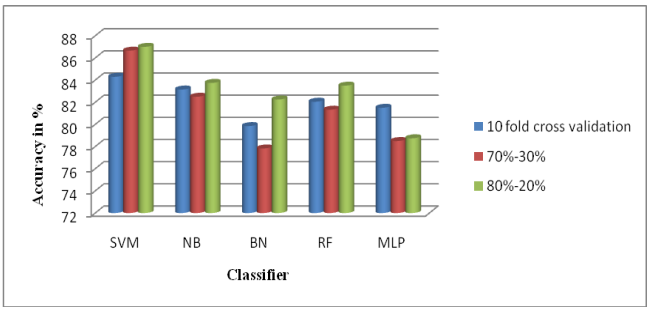

Figure 4. Graphical representation of classifier accuracy with cornell movie data set 2.

VII. CONCLUSION

Sentimental analysis play very important role for classification of text, opinion mining, and reviewcomments etc. This paper presents the problem of sentimental analysis and classification of cornell movie data set. Firstly, we have applied the different preprocessing steps for selecting the relevant features from cornell movie data set. We have suggested SVM with Correlation-based feature subset selection evaluation technique combination with Best Fit search strategy achieved 81.55% accuracy using 10 cross fold validation and 81.75% accuracy using 80%-20% training-testing data partition in case of cornell movie data set1. The SVM is also suggested with Chi-Squared attribute evaluation technique combination with Ranker search method achieved 84.30% classification accuracy using 10 cross fold validation and 87.0% classification accuracy using 80%-20% training-testing data partition in case of cornell movie data set1. In feature, we will use ensemble technique and new proposed feature selection technique to achieve better classification technique.

REFERENCES

1. P. H. Shahana and B. Omman "Evaluation of Features on Sentimental Analysis", , Elsevier, ScienceDirect, Procedia Computer Science, vol. 46, 2015, PP 1585 – 1592.
2. R. Ullaha, A. Zeband W. Kim, "The impact of emotions on the helpfulness of movie reviews", Journal of Applied Research and Technology, vol. 13, 2015, PP 359-363.
3. M. Bilal, H. Israr, M. Shahid and A. Khan, "Sentiment classification of Roman-Urdu opinions using Naive Bayesian, Decision Tree and KNN classification techniques", Journal of King Saud University –Computer and Information Sciences, vol. 28, 2016, PP 330–344.
4. S. Liao, J. Wang , R. Yu , K. Sato and Z. Cheng , "CNN for situations understanding based on sentiment analysis of twitter data", Elsevier, Science Direct, Procedia Computer Science, vol. 111, 2017, PP 376–381.
5. www.cs.cornell.edu/people/pabo/movie-review-data/. (Browsing Date :July 2019)
6. http://www.cs.waikato.ac.nz/ml/weka/(Browsing Date :July 2019).
7. P. Baid, A. Gupta and N. Chaplot, "Sentiment Analysis of Movie Reviews using Machine Learning Techniques", International Journal of Computer Applications, vol. 179 , 2017, PP 45-49.
8. J. Han, M. Kamber and J. Pei, "Data Mining Concept and Technaiues", 3rd Edition, Morgan Kaufmann, 2012.
9. N. B. Khanna, J. S. Moses and M. Nirmala, "SoftMax based User Attitude Detection Algorithm for Sentimental Analysis", Elsevier, ScienceDirect, Procedia Computer Science, vol. 125, PP 313–320.
10. S. Zhang , Z. Wei , Y. Wang and T. Liao, "Sentiment Analysis of Chinese Micro-Blog Text based on Extended Sentiment Dictionary", Future Generation Computer Systems, vol. 81, PP 395-403.
11. M. Kang , J. Ahn and K. Lee, "Opinion mining using ensemble text hidden Markov models for text classification", Elsevier, Expert Systems With Applications, vol. 94, 2018, PP 218-227.
12. J. Chambua, Z. Niu , A. Yousif and J. Mbelwa, "Tensor Factorization Method based on Review Text SemanticSimilarity for Rating Prediction", Elsevier, Expert Systems With Applications, vol. 114, 2018, PP 629-638.
13. M. Malik, S. Habib and P. Agarwal, "), A Novel Approach to Web-Based Review Analysis using Opinion Mining,", Elsevier, ScienceDirect, Procedia Computer Science, vol. 132, 2018, PP 1202-1209.
14. X. Li, Y. Wang, A. Zhang , C. Li, J. Chi and J. Ouyang, "Filtering out the Noise in Short Text Topic Modeling", Information Sciences, vol. 456, 2018, PP 83-96.
15. L. Wang, J. Niu, H. Song and M. Atiquzzaman, "), Senti Related: a Cross-Domain Sentiment Classification Algorithm for Short Texts through Sentiment Related Index", Journal of Network and Computer Applications, vol. 101, 2018, PP 111-119.
16. N. Öztürk and S. Ayvaz , "Sentiment Analysis on Twitter: A Text Mining Approach to the Syrian Refugee Crisis", Telematics and Informatics, vol. 35, 2018, PP 136-147.
17. 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", SciVerse Science Direct, Procedia Engineering vol. 53, 2013, PP 453 – 462.
18. W. Medhat , A. Hassan and H. Korashy , "Sentiment analysis algorithms and applications A survey", Elsevier, Ain Shams Engineering Journal, vol. 5, 2014, PP 1093–1113.
19. A. K. Pujari, "Data Mining Techniques". Universities Press (India) Private Limited. 4th ed.,2001.
20. P. MohanaChelvan and K. Perumal, " A Comparative Analysis of Feature Selection Stability Measures" , International Conference on Trends in Electronics Informatics, 2017, PP. 124-128.
21. K. M. A Kumar, N. Rajasimha , M. Reddy, A. Rajanarayana and K. Nadgir, "Analysis of Users' Sentiments from KannadaWeb Documents", Elsevier, Science Direct, Procedia Computer Science, 54, 2015, PP 247 – 256.

AUTHORS PROFILE

Dr. A. K. Shrivasis a Assistant Professor at Department of IT&CS, Dr. C. V. Raman University, Bilaspur (C. G.), India. He has published more than 80 research papers in national and international journals. His research area are Data Mining and Information Security. He is member of IAENG.

Dr. S. M. Ghosh, is a professor at department of computer science and engineering, Dr. C. V. Raman University, Bilaspur(C. G.),India. He has published more than 100 research papers in national and international journals. He is Director of Auropath.

Mr. Amit Kumar Dewanganis a research scholar at department of computer science and engineering, Dr. C. V. Raman University, Bilaspur(C. G.),India. He has published more than 17 research papers in national and international journals. His research area is Data Mining, Text Mining. He is member of Vijnana Bharati.