iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** III **Month of publication:** March 2021

DOI: https://doi.org/10.22214/ijraset.2021.33483

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue III Mar 2021- Available at www.ijraset.com

Sentiment Analysis and Review Classification using Deep Learning

Dr. P. Saveetha[1], J. Charanya[2], R. Gowsiki Devi[3], N. Kowsalya[4], G. Maheshwari[5], N. Thiripura Sundari[6]
[1]*Professor*, [2]*Assistant Professor*, [3, 4, 5, 6]*UG Students – Final Year, Department of Information Technology, Nandha College of Technology, Perundurai, Tamilnadu, India*

Abstract: ***Sentiment analysis or opinion mining is the method of deciding the emotions, opinions behind series of words, its accustomed gain an understanding of the attitudes, opinions and emotions expressed by the people. In recent years, research work is being performed in these fields by applying numerous methodologies. Sentiment analysis of social media content has become one of the most sought area among researchers because the number of product review sites, social networking sites, blogs, and forums are developing extensively. This field mainly utilizes supervised, unsupervised and semi-supervised technique for this sentiment prediction and classification task. This project studies the inability of widely used feature selection method like Information gain (IG) on machine learning and deep learning approaches. Initially, feature selection method called information gain has been used to select the feature subsets. In addition Naïve Bayes classification is done to find the probability of features found in all sub categories of reviews. Deep learning approach has been used in exact classification of new review from the given dataset and R language is used to develop the application.***

Keywords: Sentiment Analysis, movie reviews, naive bayes, supervised and unsupervised technique, machine learning, deep learning.

I. INTRODUCTION

Sentiment analysis, the method outlined as "aims to work out the emotions, perspective of an individual. An opinion may be a judgement a couple of explicit factor that acts as a key influence on a private method of deciding. People's belief and therefore the selection they create are always dependent on how others see and judge the thing. Therefore opinion holds high values in several facet of life. In recent years, this field is widely appreciated by researchers because it has dynamic range of application in various numbers of fields. There are several areas which are benefited from the result of this sentiment analysis such as marketing; politics etc. Now-a-days it has become difficult to select their preferred genre of movie due to vast range of movies. Movie review turn out to be very useful reference. The solution obtained can be roughly classified into machine learning approach and lexicon-based approach for solving the problem of sentiment classification. The previous approach was applied to classify the feelings supported trained moreover as take a look at information sets. Before this few researchers applied hybrid approaches by combining both approaches machine learning and lexical to improve the sentiment classification performance.

This field become more challenging because of many demanding and interesting problems are still remains in this field to solve. In this section we provide a brief overview of the previous studies regarding supervised multiple Machine Learning (ML) algorithms.

II. LITERATURE SURVEY

Sentiment Classification Using Rough Set Based Hybrid Feature Selection -Basant Agarwal, Namita Mittal - This work focus on that sentiment analysis suggests that to extract opinion of users from review documents. Sentiment classification victimization Machine Learning ways faces the matter of high spatial property of feature vector. Therefore, a feature choice methodology is needed to eliminate the impertinent and noisy options from the feature vector for efficient working of ML algorithms. Rough Set Theory based feature choice methodology finds the best feature set by eliminating the redundant options. Rough Set Theory (RST) based feature selection methodology is applied for sentiment classification. A Hybrid feature selection methodology supported RST and knowledge Gain (IG) is projected for sentiment classification. Proposed ways are evaluated on four customary datasets viz. moving-picture show review, product (book, video disk and electronics) review dataset. Experimental results show that Hybrid feature selection methodology outperforms than different feature choice ways for sentiment classification. Sentiment analysis is used for extracting the opinion by analysing the text documents. These days individuals are using web for writing their opinion on blogs, social networking websites, discussion forums etc. Hence, it's considerably required to investigate these web contents. Rough Set Theory has been used for reducing the feature vector size for text classification. However, it's not been investigated for sentiment analysis nevertheless. Contribution of this paper:

1) Rough Set Theory primarily based feature choice methodology is applied for sentiment classification.
2) Hybrid Feature choice methodology is projected supported Rough Set and Information Gain that performs higher than different feature choice ways.
3) Projected ways square measure experimented with four completely different customary datasets.

Rough Sets Theory (RST) may be a mathematical tool to form attribute reduction by eliminating redundant condition attributes (features). The rough set is that the approximation of obscure idea (set) by a combine of precise ideas, referred to as lower and higher approximations. Rough Set Attribute Reduction (RSAR) may be a filter primarily based methodology by that redundant options square measure eliminated by keeping the number of data intact within the System. Proposed Hybrid Approach to Feature choice The utility of an attribute is set by each its relevancy and redundancy. An attribute has relevancy if it's prognosticative to the category attribute, otherwise it's unsuitable. An attribute is taken into account to be redundant if it's related to with different attributes. Hence, The Aim is to search out the attributes that square measure extremely related to with the category at tribute, however not with different attributes for an honest attribute set. Information Gain primarily based feature choice ways confirm the importance of a feature within the documents. But, its disadvantage that threshold price is needed at the start that isn't legendary typically. This methodology doesn't think about the redundancy among the attributes. Hybrid feature selection methodology is projected that is predicated on RSAR and IG. Experimental results show that Hybrid feature selection methodology with terribly less range of options produces higher results as compared to different feature choice ways. All the ways are experimented mistreatment four customary datasets. In future, a lot of ways are often explored for creating rough set primarily based feature choice methodology computationally a lot of economical by incorporating organic process approaches in choosing feature subsets. Sentiment Classification of Movie Reviews using Hybrid Method - M.Govindarajan - This work explicit that the realm of sentiment mining (also referred to as sentiment extraction, opinion mining, opinion extraction, sentiment analysis, etc.) has seen an oversized increase in tutorial interest within the previous couple of years. Researchers within the areas of natural language processing, data mining, machine learning, and all others have tested a range of ways of automating the sentiment analysis method. During this analysis work, new hybrid classification technique is planned supported coupling classification ways mistreatment arcing classifier and their performances area unit analysed in terms of accuracy. A Classifier ensemble was designed mistreatment Naive Bayes (NB), Support Vector Machine (SVM). Within the planned work, a comparative study of the effectiveness of ensemble technique is created for sentiment classification. The ensemble framework is applied to sentiment classification tasks, with the aim of expeditiously integration completely different feature sets and classification algorithms to synthesize additional correct classification procedure. The practicability and also the advantages of the planned approaches area unit incontestable by suggests that of motion picture review that's wide employed in the sector of sentiment classification. A large vary of comparative experiments area unit conducted and at last, some in-depth discussion is given and conclusions area unit drawn regarding the effectiveness of ensemble technique for sentiment classification. Recently, several websites have emerged that supply reviews of things like books, cars, snow tires, vacation destinations, etc. They describe the things in some detail and valuate them as good/bad, most popular/not preferred. So, there's motivation to reason these reviews in an automatic manner by a property apart from topic, namely, by what's referred to as their 'sentiment' or 'polarity'. That is, whether or not they suggest or don't suggest a selected item.

A. Data Pre-processing

A data Pre-processing completely different pre-processing techniques were applied to get rid of the noise from our knowledge set. It helped to cut back the dimension of our knowledge set, and therefore building a lot of correct classifier, in less time. the most steps concerned area unit i) document pre-processing, ii) feature extraction / choice, iii) model choice, iv) coaching and testing the classifier. Data pre-processing reduces the scale of the input text documents considerably. It involves activities like sentence boundary determination, tongue specific stop-word elimination and stemming. Stop words area unit useful words that occurs off-times within the language of the text (for example, „a", "the", "an", "of" etc. in English language), so they're not helpful for classification. Stemming is that the action of reducing words to their root or base kind.

B. Document Indexing

Document categorization making a feature vector or different illustration of a document could be a method that's illustrious within the IR community as categorization. There area unit a spread of how to represent matter knowledge in feature vector kind, but most area unit supported word co-occurrence patterns. In these approaches, a vocabulary of words is outlined for the representations, that area unit all attainable words which may be necessary to classification.

This is often sometimes done by extracting all words occurring higher than a particular range of times (perhaps three times), and process your feature area so every dimension corresponds to at least one of those words. Once representing a given matter instance (perhaps a document or a sentence), the worth of every dimension (also referred to as Associate in Nursing attribute) is allotted supported whether or not the word similar to that dimension happens within the given matter instance. If the document consists of just one word, then solely that corresponding dimension can have a worth, and each different dimension (i.e., each different attribute) are going to be zero. This is referred to as the ``bag of words'' approach. One necessary question is what values to use once the word is gift. May be the foremost common approach is to weight every gift word exploitation its frequency within the document and maybe its frequency within the coaching corpus as a full. The most common coefficient perform is that the tfidf (term frequency inverse document frequency) live, however different approaches exist. In most sentiment classification work, a binary coefficient perform is employed. Assignment one if the word is gift, zero otherwise, has been shown to be handiest.

C. Dimensionality Reduction

Dimension Reduction techniques area unit planned as a knowledge pre-processing step. This method identifies an appropriate low-dimensional illustration of original knowledge. Reducing the spatiality improves the procedure potency and accuracy of the information analysis.

Steps:

1) Select the dataset.
2) Perform discretization for pre-processing the information.
3) Apply Best initial Search algorithmic rule to separate redundant & super flows attributes.
4) Using the redundant attributes apply classification algorithmic rule and compare their performance. Identify the most effective One. Best initial Search Best initial Search (BFS) uses classifier analysis model to estimate the deserves of attributes. The attributes with high advantage worth is taken into account as potential attributes and used for classification Searches the area of attribute subsets by augmenting with a backtracking facility. Best initial could begin with the empty set of attributes and search forward, or begin with the total set of attributes and search backward, or begin at any purpose and search in each directions.

D. Existing Classification Methods

Three classification ways (Naïve Bayes, Support Vector Machine and proposed Hybrid NB-SVM Method) are tailored for every training set.

The foremost competitive classification ways are used for a given corpus. The results are evaluated exploitation the cross validation methodology on movie review based on the classification accuracy. The primary metric for evaluating classifier performance is classification Accuracy - the proportion of take a look at samples that are properly classified. The accuracy of a classifier refers to the power of a given classifier to properly predict the label of recent or previously unseen information (i.e. tuples while not category labels information). Similarly, the accuracy of a predictor refers to however well a given predictor will guess the worth of the anticipated attribute for brand new or previously unseen information. The author terminated that during this analysis, a brand new hybrid technique is investigated and evaluated their performance supported the picture show review information then classifying the reduced information by NB and SVM. Next a hybrid NB SVM model and NB, SVM models as base classifiers are designed. Finally, a hybrid system is planned to create optimum use of the most effective performances delivered by the individual base classifiers and therefore the hybrid approach.The hybrid NB-SVM shows higher proportion of classification accuracy than the bottom classifiers and enhances the testing time thanks to information dimensions reduction. The experiment results cause the subsequent observations. SVM exhibits higher performance than NB in□ the vital respects of accuracy. Comparison between the individual classifier and therefore the hybrid classifier: it's clear that the hybrid classifier show the significant improvement over the only classifiers.

III. SYSTEM ANALYSIS

A. Existing System

In existing system, Sentiment classification process has been classified into feature level. Sentiment classification at the given individual record level, considers the individual record as a sentence or paragraph is taken to identify whether the sentence is positive or negative. In addition, this project considers the sentence classification using Naïve Bayes method so that the sentences are feature extracted and probability percent of each feature in the given sentences are found out and displayed. Moreover, Information Gain (IG) method is employed as a single univariate method with low complexity, which ranks the features based on high information gain entropy in decreasing order.

1) Drawbacks of Existing System

a) Using IG method alone cannot handle redundant features.
b) If feature words are present in more than one review type, probability values may not assist in better classification.
c) Naïve Bayes classification is carried out for words present in the dataset instead of information gain based feature words.

B. Proposed System

In addition with all the existing system mechanism, the proposed study also presents deep learning based classification approach. Here, reviews for movie and product domain contents are taken. Then term document matrix is prepared for both domains to get the feature words. Then the words are encoded to prepare input for neural network's input layer. Initial weight values are given for neurons and edges. The model is trained for some fixed repititions. The weight values are updated in intermediate iterations. Then the final weight values are used for further data set model. Thus the recommendation features for combined cross domain can be fetched out to find the test data review whether it is given for book or electronic shopping cart items.

1) Advantages of Proposed System

The proposed system has following advantages.

a) Along with IG method, hot vector encoding is created for input neurons based on IG extracted feature words.
b) Even if feature words are present in more than one review type, deep learning assist in better classification.
c) Neural network model assists in easy review classification.
d) The new review content test data is properly classified based on the training data in the given dataset records.

IV. SYSTEM ARCHITECTURE

The statistical property is employed as a good answer for feature choice. IG technique is employed to pick important feature supported the category attribute rules of features classification. The IG value of every term will measure the quantity of bits of information acquired for class prediction by knowing presence or absence of that term within the document .Depending on the score of IG it offers a ranking of features. So an explicit range of features can be selected easily. In this module, reviews for movie domain and product domain contents are taken from the 'csv' file. Two columns with review as first column and category like 'horror', 'suspense', 'love' etc as second column. These details are taken as data frame. Then term document matrix is prepared for both domains to get the feature words. Based on information gain value, top feature words are filtered out and then the words are encoded to prepare input for neural network's input layer. Initial weight values are given for neurons and edges. The model is trained for some fixed repetitions. The weight values are updated in intermediate iterations. Then the final weight values are used for further dataset model. Thus the recommendation features for combined cross domain can be fetched out to find the test data review whether it is given for book or electronic shopping cart items.

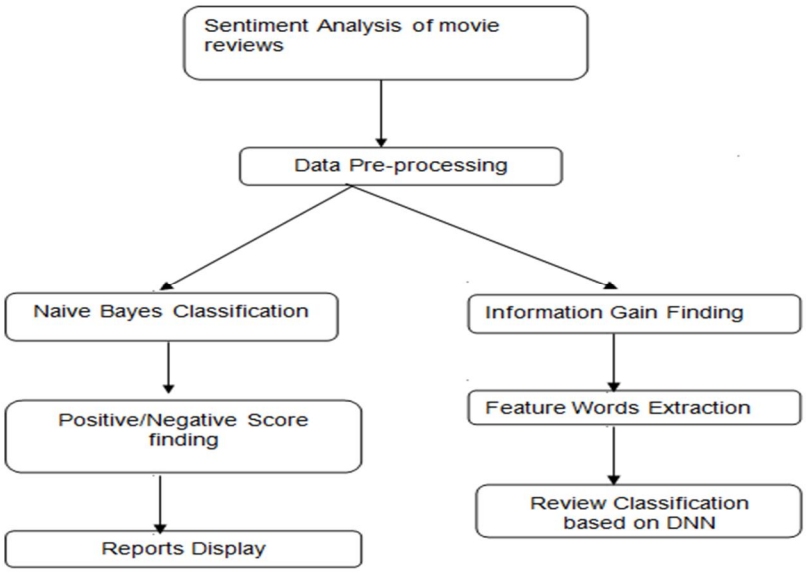


Figure 1 System flow diagram

A. *Naïve Bayes Algorithm*

Naive Bayes classifiers are a set of classification algorithms supported Bayes' theorem. It's not one algorithm but a family of algorithms where all of them share a standard principle, i.e. every pair of features being classified is independent of each other .To start out with, allow us to consider a dataset. Naïve Bayes classifiers are highly scalable, requiring variety of parameters linear within the number of variables (features/predictors) during a learning problem. Maximum-likelihood training are often done by evaluating a closed-form expression, which takes linear time, instead of by expensive iterative approximation as used for several other sorts of classifiers In the statistics and computing literature, naive Bayes models are known under a spread of names, including simple Bayes and independence Bayes. All these names reference the utilization of Bayes' theorem within the classifier's decision rule, but naïve Bayes isn't (necessarily) a Bayesian method.

B. *Support Vector Machine*

Support vector machine is a supervised machine learning algorithm which is used for both classification and regression. Primarily, it's used for classification problems in machine learning.

The goal of the SVM algorithm is to make the simplest line or decision boundary which will segregate n-dimensional space into classes in order that we will easily put the new data point within the correct category within the future. This best decision boundary is called a hyper-plane. SVM chooses the acute points/vectors that help in creating the hyper-plane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyper-plane.

C. *Deep Neural Network*

Deep neural networks have recently become the quality tool for solving a spread of computer vision issues. Whereas coaching a neural network is outside the Open VX scope, commerce a pre trained network and running logical thinking thereon is a vital a part of the Open VX practicality. The conception of the Graph API of nodes representing functions and links representing information is extremely convenient for implementing deep neural networks with Open VX. In fact, every neural network unit is portrayed as a graph node.

Table-I Performance comparison of algorithms in terms of accuracy

Algorithm	Accuracy
Naïve Bayes	78.88 %
Support Vector Machine	85.22 %
Deep Neural Network (DNN)	91.36 %

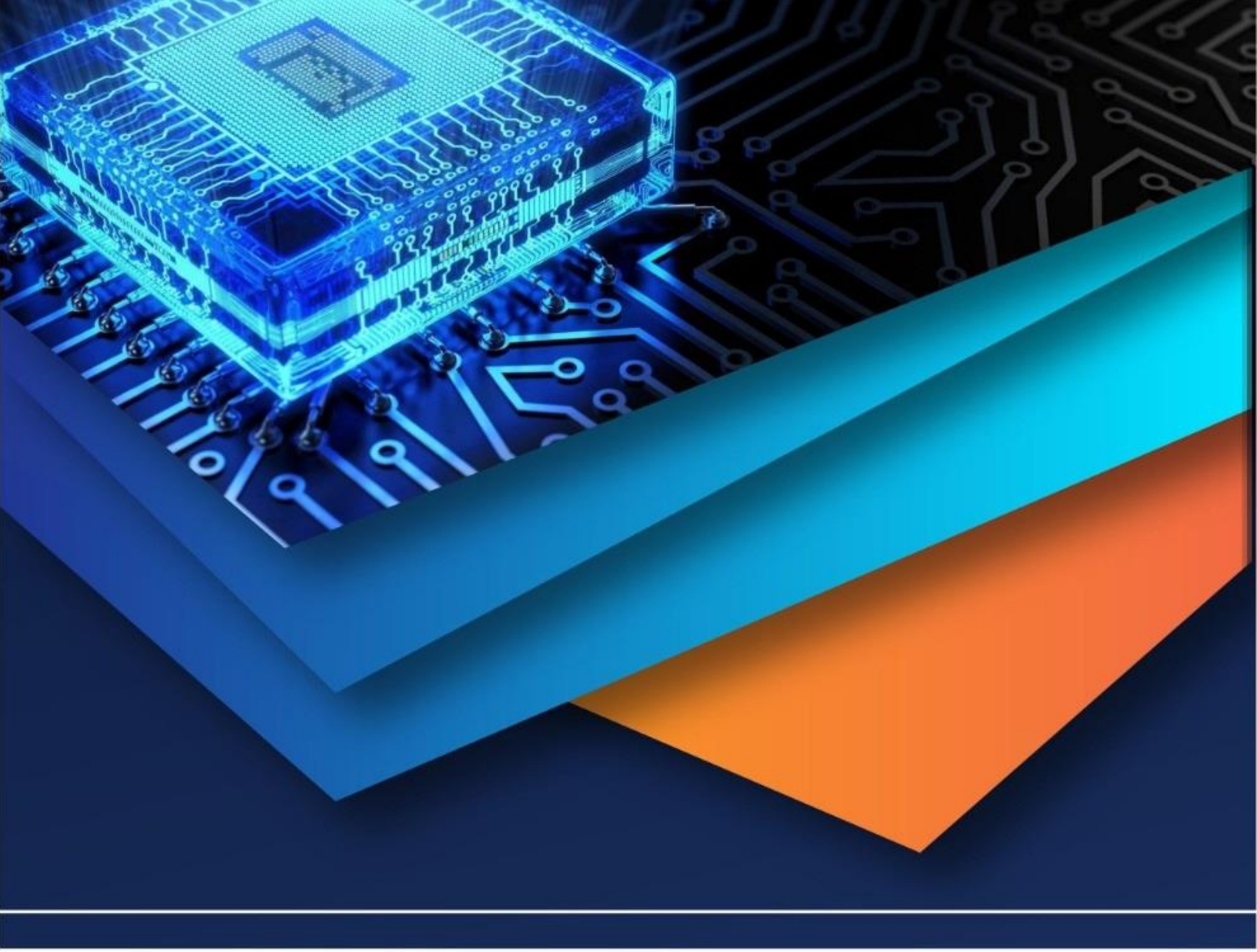
While using Support Vector Machine algorithm, accuracy level has been increased by 7% compared to Naïve Bayes algorithm. However, while comparing with both SVM and Naïve Bayes algorithms, DNN has increased the accuracy level by 12% than existing system.

V. CONCLUSION

This project investigates the inability of the widely used feature selection method like Information Gain (IG) on machine learning and deep learning approaches. The proposed methods are evaluated on datasets viz. movie reviews, and product reviews dataset. Initially, select the feature subsets from a feature selection method called information gain. Naïve Bayes classification is carried out in addition to find the probability of features found in all sub categories of reviews. Deep learning approach is used in exact classification of new review content from the given dataset content. R Language is used to develop the application. It is believed that almost all the system objectives that have been planned at the commencements of the software development have been net with and the implementation process of the project is completed. An attempt run of the system has been created and is giving smart results the procedures for process is straightforward and regular order. The method of making ready plans been left out which could be considered for additional modification of the application.

REFERENCES

[1] Agarwal B, Mittal N. Sentiment classification using rough set based hybrid feature selection. In: Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis (Atlanta), 2013. p. 115–9.
[2] Govindarajan M. Sentiment classification of movie reviews using hybrid method. Int J Adv Sci Eng Technol. 2014; 3:139.
[3] Liu B. Sentiment analysis and subjectivity. In: Indurkhya N, Damerau FJ, editors. Invited chapter for the handbook of natural language processing. 2nd ed. England: Taylor & Francis; 2010.
[4] Dhaoui C, Webster CM, Tan LP. Social media sentiment analysis: lexicon versus machine learning. J Consumer Mark. 2017; 34(6):480–8.
[5] Samal BR, Behera AK, Panda M. Performance analysis of supervised machine learning techniques for sentiment analysis. In: Proceedings of the 1st ICRIL international conference on sensing, signal processing and security (ICSSS). Piscataway: IEEE; 2017. p. 128–3. http:/northcampus.uok.edu.in/downloads/20161105144024077.pdf
[6] Singh, J.P., et al., Predicting the "helpfulness" of online consumer reviews, Journal of Business Research (2016), http:// dx.doi.org/10.1016/j.jbusres.2016.08.008.
[7] Dave K, Lawrence S, Pennock DM. Mining the peanut gallery: opinion extraction and semantic classifcation of product reviews. In: Proceedings of the 12th international WWW conference. Budapest; 2003. p. 519–8.
[8] BiswaRanjan Samal, Mrutyunjaya Panda, HumanBeing Character Analysis from Their SocialNetworking Profiles A Semisupervised Machine Learning Approach, (IJCSIS) International Journal of Computer Science and Information Security, Vol. 14, No. 5, May 2016 .
[9] Bing Liu, Xiaoli Li, Wee Sun Lee and Philip S. Yu, "Text Classification by Labeling Words" , American Association for Artificial Intelligence. 2004.
[10] Semi-Supervised Learning—O. Chapelle, B. Schölkopf, and A. Zien, Eds. (London, U.K.: MIT Press, 2006, pp. 508, ISBN: 978-0-262-03358-9).R eviewed by Philippe Thomas.
[11] Trevor Hastie, Robert Tibshirani, Jerome Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd edition) (Springer Series in Statistics), 2009.
[12] Sebastian B. Thrun, Efficient Exploration In Reinforcement Learning (1992).
[13] Stiglitz, Joseph E. "Learning to learn, localized learning and technological progress." Economic policy and technological performance (1987): 125-153.
[14] Freitag, Dayne. "Machine learning for information extraction in informal domains." Machine learning 39.2-3 (2000): 169-202.
[15] Bing Liu. Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, May 2012.
[16] Aidan Finn, Nicholas Kushmerick, and Barry Smyth. Genre classification and domain transfer for information filtering. In Fabio Crestani, Mark Girolami, and Cornelis J. van Rijsbergen, editors, Proceedings of ECIR-02, 24th European Colloquium on Information Retrieval Research, Glasgow, UK. Springer Verlag, Heidelberg, DE.
[17] Janyce Wiebe, Rebecca Bruce, Matthew Bell, Melanie Martin, and Theresa Wilson. A corpus study of evaluative and speculative language. In Proceedings of the 2nd ACL SIGdial Workshop on Discourse and Dialogue, 2001.
[18] Vasileios Hatzivassiloglou and Kathleen R. McKeown. Predicting the semantic orientation of adjectives. In Proceedings of the 35th Annual Meeting of ACL, 1997.
[19] Vasileios Hatzivassiloglou and Janyce M. Wiebe. Effects of adjective orientation and gradability on sentence subjectivity. In Proceedings of the 18th International Conference on Computational Linguistics, 2000.
[20] P. Subasic and A. Huettner. Affect analysis of text using fuzzy semantic typing. IEEE-FS, 9:483–496, Aug. 2001
[21] M. Hearst. Direction-Based Text Interpretation as an Information Access Refinement. 1992. Alexios Chouchoulas, Qiang Shen, "Rough set-aided keyword reduction for text categorization", Applied Artificial Intelligence, Vol. 15, No. 9, pp. 843-873. 2001.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY