E-ISSN 2581 - 7957
P-ISSN 2277 - 3916

CVR Journal of Science and Technology, Volume 19, December 2020
DOI: 10.32377/cvrjst 1912

# Proficient Machine Learning based Scheme for Classifying User Reviews

Dr. M. Deva Priya[1] and T. Akash[2]
[1]Assoc. Professor, Sri Krishna College of Technology/CSE Department, Coimbatore, Tamil Nadu, India
Email: m.devapriya@skct.edu.in
[2]UG Scholar, Sri Krishna College of Technology/CSE Department,Coimbatore, Tamil Nadu, India
Email: akashthiruvenkatam@gmail.com

***Abstract:*** **Each user represents his view on a product or issue using several aspects or features. A user's opinion may be conveyed based on diverse levels of positivity or negativity. In this paper, a proficient technique is propounded for categorizing customer reviews taken from Twitter dataset. The features are extracted using Bag of Words (BoW) algorithm and the information is categorized using modified Support Vector Machine (SVM) and ensemble classification. Further, the results are optimized using BAT algorithm which improves the classification accuracy. It is notable that the scheme proposed offers better accuracy, precision, recall and involves less time.**

***Index Terms:*** **Text mining, Bag of Words (BoW) algorithm, Support Vector Machine (SVM), ensemble classification, Twitter dataset**

## I. Introduction

Users exhibit their views on the social media and other related applications about a particular topic. They either grant their support or non-support [1]. These views and opinions collectively form a huge volume of data with huge dimensions and velocities [2]. The data from social media constituting a huge volume is processed for exposing the views of the users [3]. Many researchers have concentrated on using huge volumes of data to describe, estimate and predict human insights and features of important areas of applications [4]. In the data available over the internet, about 82% are text and hence text analysis plays a dominant role in the process of expounding views, emotions and sentiments of users [5].

Sentiment analysis, well-known as opinion mining is appropriate for measuring the users' sentiments connected to a subject by examining their views and posts on the social media [6]. The polarities of posts are classified into emotions that include positive feedback, negative feedback and similar ones [7]. Sentiment analysis or opinion mining is involved in observing sentiments of people. Internet is a collection of sentiments. People post their views on social media including forums, blogs and other sites. Social media includes reviews about different issues, news and products etc., People update their views through the internet. Sentiment analysis deals with categorization of these reviews based on the opinion as positive or negative.

Sentiment analysis is categorized into lexicon and machine learning analysis.

- ✓ Lexicon analysis emphases on computing the degree of polarity of a particular content based on the words' semantic positioning or the phrases seen in the document. This method does not consider the context of investigation [8].
- ✓ Machine learning based sentiment analysis is based on constructing models collected from the training dataset. These data are labeled for measuring the degree of orientation of the document. This scheme is extensively used in the method of opinion collection for evaluating the satisfying sentiments relating to a topic in the data. This method can be applied to items, themes, folks, events and services in varied domains [9, 10]. This method of analysis yields extremely variable accuracies.

These methods do not seem to be successful due to issues in the semantic positioning of words that varies based on the context. Deep learning based schemes may be used in opinion mining to effectively estimate users, emotions, sentiments, likes and dislikes.

The learnt information is shared among users through tools like chats, forums, comments etc., of social media. The information on such websites is direct and is unstructured and fuzzy. In everyday debates, the users do not focus on spelling, grammar and sentence forms prompting diverse types of ambiguities. Phrase mining focuses on removing phrases from a text corpus which is good in quality and has diverse downstream applications like extraction and retrieval of information, building of taxonomy and modeling topics.

Opinion mining is highly employed in very recent years. Studies which are not optimal for offering better precision rate focus on the precision rate of feature reviews and extraction of opinion words. With the development of textual data available on the Internet, automatic text classification is a reasonable solution for organizing information and managing knowledge. Feature selection is the rudimentary phase in statistical text classification, which is significantly based on the term weighting methods.

The amount of observations in a class of imbalanced data is considerably rarer in another class leading to significant focus in the area of data mining. As multi-class imbalanced learning is hardly specified, it emphases on binary imbalanced cases.

In this paper, a model is propounded to classify user reviews posted on social media. Features are extracted using the Bag of Words (BoW) algorithm. Classification is performed using Support Vector Machine (SVM) algorithm and ensemble classification. The results are optimized using the BAT Algorithm.

E-ISSN 2581 - 7957
P-ISSN 2277 - 3916

DOI: 10.32377/cvrjst 1912

## II. Related Work

This section discusses about the researches of authors done in feature extraction, classification and optimization of mining texts.

### A. Feature Extraction

Supervised Protein-Protein Interaction Extraction (PPIE) deals with richly selected features and kernels yielding high accuracy [11]. Features and kernels focus on the domain knowledge and analysis of natural language. This converts a supervised model into a costly, heavy and fragile one. Additionally, the representation methods like one-hot encoding and vector space model do not keep the semantic likeness between words. The instance representation architecture of PPIE includes word representation, vector configuration and feature choice to reduce manual intervention.

In the study of Salloum et al (2017) [12] textual data from Facebook is analyzed to determine interesting information and represent in diverse ways. The posts are extracted using diverse mining techniques and examined which shows that the Fox news is the mostly used channel with maximum on Facebook, followed by CNN and ABC News respectively.

The methods which exist for phrase mining rely on multifaceted, skilled linguistic analyzers, and so likely have an unacceptable performance on text quantities of novel areas and genres with exclusive adjustment. The state-of-the-art models and even data-driven models are not fully automatic. This is due to the requirement of human experts for scheming rules or classifying phrases. Shang et al (2018) [13] have proposed Auto Phrase which is suitable for all languages until a general knowledge base is available, while benefiting from a POS tagger. In contrast to other methods, this method is efficient and can be extended to model single-word phrases.

### B. Classification

Text classification which is vital in information retrieval is applied to categorize documents into a collection of predefined groups. Numerous practices are available to classify data and many researches have dealt with English text classification. Classification is done using Support Vector Machine (SVM) algorithm, Modified SVM algorithm and Ensemble Classification.

#### i. Support Vector Machine (SVM) algorithm

Machine learning techniques are mostly used for extracting text. SVM is the best supervised learning method used in text classification. Ramesh & Sathiaseelan (2015) [14] have proposed Advanced Multi Class Instance Selection based Support Vector Machine (AMCISSVM) to increase the effectiveness of SVM. The performance of AMCISSVM is analyzed by relating it with Multi-class Instance Selection (MCIS) which shows high accuracy to multi-datasets and Neighbourhood Property based Pattern Selection (NPPS) algorithm.

The healthcare industry gathers enormous amount of healthcare data that is not extracted to find the concealed information for efficient analysis and decision making. Discovery of concealed patterns and relationships frequently goes inactive. Latest data mining schemes can assist and offer solution to deal with these circumstances. Data mining schemes play a dominant role in domains like text, graph, medical, multimedia and web mining. Vijayarani et al (2015) [15] have predicted kidney ailments by using SVM and Artificial Neural Network (ANN). The proposed scheme offers better accuracy and involves less execution time. It is also seen that the performance of ANN is better in contrast to SVM.

Nowadays, news is posted in blogs and social networks and the major role of text mining is to classify the news posted. Dadgar et al (2016) [16] have dealt with classifying news based on their popularity, country and time. The authors have propounded a classification scheme using Term Frequency-Inverse Document Frequency (TF-IDF) and SVM. It involves pre-processing of text, extracting features related to TF-IDF and classification based on SVM. The scheme is analyzed for BBC and diverse news group datasets.

Few works have focused on Arabic text classification. Mohammad et al (2016) [17] have discussed about three famous techniques used to classify data which are applied to Arabic datasets. This study uses fixed quantity of documents for training and testing and shows that SVM yields better performance.

Networking messages, news and reviews of products involve much effort in extracting opinions and sentiments from natural language. Rana & Singh (2016) [18] have dealt with sentiment orientation by taking the positive and negative sentiments from reviews of films. The scheme uses Naive Bayes Classifier (NBC), Linear SVM and synthetic words followed by Synthetic words approach to provide better accuracy.

The studies by Ali et al (2016) [19] are based on NBC, SVM, K-Nearest Neighbours (KNN) and classical ontology which are not suitable for categorizing feature reviews into better levels of polarity. Moreover, the prevailing classical ontology-based systems are not capable of retrieving distorted information from reviews and hence provide pitiable results. They have proposed a classification scheme for identifying features and semantic knowledge based on SVM and Fuzzy Domain Ontology (FDO) for hotel dataset. SVM is used to remove inappropriate reviews (noises) and FDO is used for finding the polarity of features. The mixture of FDO and SVM improves the precision rate.

Power load forecasting shows time and spatial distribution of upcoming power loads. The accuracy directly manipulates the trustworthiness of the power system. Niu & Dai (2017) [20] have propounded Empirical Mode Decomposition-Grey Relational Analysis-Modified Particle Swarm Optimization-Least Squares Support Vector Machine (EMD-GRA-MPSO-LSSVM) load predicting model which uses a de-noising method merging empirical mode decomposition and Grey Relational Analysis (GRA) to deal with the actual load series. It shows the processed results of modified Particle Swarm Optimization (PSO) and Least Square-SVM (LS-SVM).

#### ii. Modified SVM algorithm

In the work done by Jadav & Vaghela (2016) [21], primarily, the dataset is pre-processed to convert unstructured reviews into structured ones. Lexicon based

method is involved to change structured review into a numerical score. The dataset is pre-processed using feature selection and semantic analysis. Stop word removal, stemming, POS tagging and computation of sentiment score using SentiWordNet dictionary are carried out in the pre-processing step. The opinions are classified either as positive or negative. SVM categorizes reviews wherein, RBF kernel SVM is adapted using hyper parameters. Optimized SVM offers better results when compared to SVM and NBC.

Sabbah et al (2017) [22] have proposed improved frequency-based term weighting mechanism namely, Mtf, Mtfidf, TFmIDF, and mTFmIDF. The propounded term weighting schemes consider the quantity of missing terms by operating the weight of existing terms. Moreover, from the results, it is understood that by using SVM and KNN for classification, better results can be obtained when compared to the weighting schemes like TF, TFIDF and Entropy.

Non-informative sequence features and class disparity in the training data are some of the issues faced. To focus on these issues, Abidine et al (2018) [23] have propounded a new scheme based on an amalgamation of Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and the modified weighted SVMs. Initially, the dominating features are retrieved using LDA and added to the set. An appropriate sequence feature set joint with the modified WSVM based on criterion classifier offers better development and efficacy over existing methods.

### *iii. Ensemble Classification*

Ensemble method uses diverse models for better performance. Ensemble methods are widely used in areas like statistics, machine learning and computational intelligence. Ren et al (2016) [24] have reviewed conventional and advanced ensemble methods. The ensemble methods are categorized into traditional methods like bagging, boosting and random forest, negative correlation learning methods, decomposition methods, fuzzy methods, multi-objective optimization based schemes, multiple kernel learning schemes and deep learning based schemes. Deviations, developments and typical applications are examined.

Yijing et al (2016) [25] have dealt with imbalanced data using an adaptable algorithm. The imbalanced data differs in its balanced ratio, number of classes and performance of classifiers for diverse datasets. The authors have proposed Adaptive Multiple Classifier System (AMCS) to manage multi-class imbalanced learning that builds a difference between huge amounts of imbalanced data. It includes feature selection, resampling and ensemble learning chosen discriminatively for different types of data. The adapting principle of selecting components from the pool is inspected through empirical analysis. To confirm the efficiency of AMCS, AMCS with several advanced algorithms are compared.

Pan et al (2017) [26] have propounded an ensemble framework that associates spectral and spatial information that are in diverse scales. It is based on the idea that by merging separate learners, ensemble learning can offer improved generalization than an individual learner. Single learners are chosen using joint spectral-spatial features produced from diverse scales. Hierarchical Guidance Filtering (HGF) and matrix of Spectral Angle Distance (Msad) are used to build an ensemble model. These methods are combined using a weighted ensemble strategy. The spatial contextual information is reclaimed in each hierarchy. With increase in the hierarchy, the pixel spectra become smooth, and the spatial features are improved. HGF offers a series of classifiers. Consequently, Msad aids in computing the variety of training samples in every hierarchy.

Ensemble-based methods are extensively utilized for categorization of data streams. They offer better performance when compared to single learners while being significantly simple to organize real-world applications. Ensemble algorithms are used in data stream learning by combining it with drift identification algorithms and including dynamic updates like choosy removal or inclusion of classifiers. Gomes et al (2017) [27] have performed data stream ensemble learning by studying over 60 algorithms. Significant aspects such as grouping, variety and lively updates are methodically presented. Open-source tools and research challenges related to ensemble learning are also included.

Artificial Neural Networks (ANNs) are used to perform a diversity of machine learning functions like image identification, semantic segmentation and machine translation. Few studies have fully examined collections of ANNs. Ju et al (2018) [28] have examined extensively used ensemble methods comprising of unweighted averaging, majority voting, Bayes Optimal Classifier and Super Learner for image identification with Deep Neural Networks (DNNs). Many experiments with these algorithms are performed using diverse model checkpoints in a single training process, networks with similar structure but trained many times and the ones with varied structure. It is seen that the super learner offers better performance.

Sentiment analysis aids in examining opinion and sentiments prevalent in texts. The opinion mining methods are based on sentiment lexicon that includes a collection of predefined keywords. Opinion mining demands appropriate sentiment words to be mined beforehand and faces a challenge in categorizing sentences that entail an opinion without using sentiment keywords. Kang et al (2018) [29] have proposed a sentiment analysis that is based on text-based Hidden Markov models (TextHMMs) for text categorization making use of word sequences in training texts on behalf of predefined sentiment lexicon. It is essential to know text patterns that represent sentiments in ensemble TextHMMs. The hidden variables in TextHMMs are found using semantic cluster information and considering the co-occurrence of words. The sentiment orientation of sentences by fitted TextHMMs is computed, and to replicate diverse patterns, an ensemble of TextHMM-based classifiers is applied.

## *C. Optimization*

Optimization is carried out using BAT algorithm. Heraguemi et al (2016) [30] have proposed a cooperative multi-swarm bat algorithm for Association Rule Mining (ARM) that mimics bat-inspired algorithm reformed to rule learning problem (BAT-ARM). This involves loss of

communication among bats that reduce the examination of search space. It has an efficient rule generation procedure leading to perfect local search. It sustains a good trade-off between diversification and intensification. Cooperative methods are introduced among the swarms that reveal their competence in multi-swarm optimization algorithm. Alomari et al (2017) [31] have propounded a filter method named Minimum Redundancy Maximum Relevancy (MRMR) and a wrapper method, Bat Algorithm (BA) for gene selection in microarray dataset. MRMR is involved in finding the most significant genes in gene expression data, and BA is used to find the most edifying gene subset from the reduced set produced by MRMR. The wrapper method based on SVM method with 10-fold cross-validation serves as an evaluator of the BA. Alomari et al (2017) [32] have designed a Bat-Inspired algorithm for gene selection for classifying cancer using microarray datasets. Microarray data involves irrelevant, repeated and noisy genes. Gene selection problem deals with finding the most relevant genes taken from microarray data to precisely identify cancer. There are two stages in gene selection namely, filter stage that uses MRMR method, and wrapper stage that involves BA and SVM.

## III. Pre-Processing

Pre-processing of data taken from Twitter includes stop word removal, tokenization and normalization.

### *A. Stop Word Removal*

Stop words are the most frequently used words like ‘a, the, an, in. The search engine ignores them while searching and extracting them based on the query. Removing them reduces the processing time. In the natural language, stop words are removed before processing of text. The most common stop words that support text mining include prepositions, articles and pronouns.

### *B. Tokenization*

Tokenization deals with splitting a sequence of strings into tokens which include words, keywords, phrases, symbols etc., Characters like punctuation marks may be discarded and meaningful keywords are identified. In short, it is the process of delineating and categorizing sections of a string. The words in a sentence are explored and consistency is established.

### *C. Normalization*

Normalizing words deals with the process of dropping words to their roots. This depends on the nature of the text taken for analysis. The stems are not challenging if they do not become a part of human interaction.

## IV. Feature Extraction

Features are taken from the token obtained from the previous phase using the Bag of Words (BoW) algorithm.

### *A. Bag Of Words*

The Bag of Words (BoW) algorithm finds its effective use in Information Retrieval (IR) and Natural Language Processing (NLP). The text including sentences is taken as a bag of words, preserving their diversity but excluding their grammar and order of words. It also finds its application in Computer Vision (CV) [33].

It is widely used in classification of documents where the rate of word occurrence is taken as the predominant feature taken for classifier training. It is referred in a linguistic framework designed for distributional structure [34]. BoW is also widely used to generate features. Once the words in the text are added to bags, diverse measures can be taken to describe the text. Out of all the features considered, the commonly seen feature is ‘frequency’ which gives the total number of times a word is found in the document.

Lists are constructed to keep in store the frequencies of all dissimilar words. The order of words as seen in the document is not preserved in the list. This is an important aspect in BoW. This form of demonstration enables it to be applied in email filtering [33]. Nevertheless, frequencies of words that appear in the document do not seem to be the best form of demonstration. Familiar words including ‘the’, ‘a’ and ‘to' are seen to be the ones with greater frequency. This count creates an illusion that the words with high frequency are the most important. To deal with this issue, the frequencies are normalized by assigning weights to words by the inverse of document frequency.

Further, to support classification, supervised substitutes are designed to find the label for the class label. Finally, binary weighting can be used to find the frequencies for problems that are executed in the WEKA software [35]. BoW is used in different types of text mining which includes constructing a classification system involving text that contains tweets, short articles etc., The text is categorized and labelled as positive or negative. A bag containing unigrams and a vector comprising of words seen in the text are built. The text in the training set is considered repeatedly, and ‘1’ is marked in the row vector conforming to the word it holds. The feature size is reduced so as to improve the computation speed and performance of classification by applying different feature selection techniques.

## V. Classification

The customer reviews are classified into positive, negative and neutral reviews using modified SVM algorithm and ensemble classification.

### *A. Modified Support Vector Machine (Svm) Algorithm*

Support Vector Machine (SVM) algorithm is the most appropriate classification algorithm used in text mining. It is a supervised Machine Learning (ML) algorithm that is widely in classification and outlier detection along with problems related to regression.

SVM finds hyperplanes and forms groups based on patterns. The hyperplane with the highest margin is chosen as the best. SVM is also used in classifying text and image, recognizing hand-writing, detecting faces and analyzing bio sequences. SVM is capable of both linear and non-linear data. It maps the training data to higher dimensions and segregates tuples belonging to each class. With non-linear mapping to a high dimension, data is divided using a

E-ISSN 2581 - 7957
P-ISSN 2277 - 3916


DOI: 10.32377/cvrjst 1912

hyperplane found using vectors and margins. SVMs consume more training time but are accurate as they are less subject to overfitting in contrast to other methods.

**Step 1:** Initialize the population size, crossover probability, mutation probability
**Step 2:** Categorize the dataset using SVM and calculate the accuracy rate, that is the fitness function offered by the BAT algorithm
**Step 3:** Frame the fitness function which is the optimization goal
**Step 4:** Set the Population
**Step 6:** Choose the operation and compute the fitness in population and predictive rate for unidentified data
**Step 7:** Execute Crossover
**Step 8:** Implement Mutation
**Step 9:** Compute the fitness value again and move to Step 2

### *B. Ensemble Classification*

Ensemble learning aids in improving the results of Machine Learning (ML) by joining several models. This method yields a better performance in contrast to a single model. A set of classifiers learn and then cast their vote. The predictive accuracy is improved but it is challenging to understand them [36]. It is useful in solving statistical, computational and representational problems. It is not essential to find more precise models, but build models with errors. Ensemble models built to perform classification can misclassify initially.

There are different methods of building ensembles.

- Maximum Vote
- Bagging and Random Forest (RF)
- Chance Injection
- Feature choice Ensembles
- Error Correcting Output Coding (ECOC)

The algorithm is shown below.

**Step 1:** Form the test set ‘T’ using ‘n’ documents in ‘X’
**Step 2:** Form the training set ‘TR’ using the residual documents in ‘X’
**Step 3:** for every classifier in ‘C’
Train the classifier using the categorized documents in ‘T’
Use the trained classifier to categorize the documents in ‘S’
Store the resultant labels in the particular class
**Step 4:** for every ‘o’ in the range 1 to s
for every ‘j’ in the range 1 to s
for every ‘b’ in the range 1 to k
for every ‘i’ in the range b+1 to k
if(class[b,o] == class[i,j])
if(M[o,j]==0)
M[o,j]=1;
else
M[o,j]=M[o,j]*2;
**Step 5:** ‘m’ is fed into the k-means algorithm to form document clusters
**Step 6:** Apply SVM-linear algorithm on ‘T’ for document categorization
**Step 7:** Select the classes conforming to the clusters by finding the class attained in the preceding step

## VI. Optimization

The classification results are optimized using BAT algorithm.

### *A. Bat Algorithm*

BAT, founded by Yang (2010) [37] is a meta-heuristic algorithm used in global optimization. The animals with the capacity to echolocate produce calls to the surrounding and pay attention to the echoes that come from diverse objects adjacent to them. Based on this characteristic, the BAT algorithm was built to mimic the characteristics of micro bats. Varying emission rates and intensity of noise are considered [38].

It is assumed that the bats fly arbitrarily at a velocity with changing wavelength and noise. As bats find their prey, they modify the frequency, volume of noise and Pulse Emission Rate (PER). Local random walk is performed to carry out the search. Just as the bats go in search of the prey, the best solution is selected once the criteria to stop the process are satisfied. Frequency is tuned to regulate the vibrant characteristics of a swarm of bats. Investigation and deployment are monitored by the tuning algorithm.

## VII. Results and Discussion

This section discusses about the implementation done to extract features, classify and optimize them. Reviews are evaluated qualitatively and quantitatively. From the results, it is evident that the propounded system outperforms the existing systems. This novel method is proposed to deal with user reviews. It yields better accuracy and involves less time period in contrast to the existing methods and algorithms. Figure 1 shows the dataset taken as input loaded for experimentation.

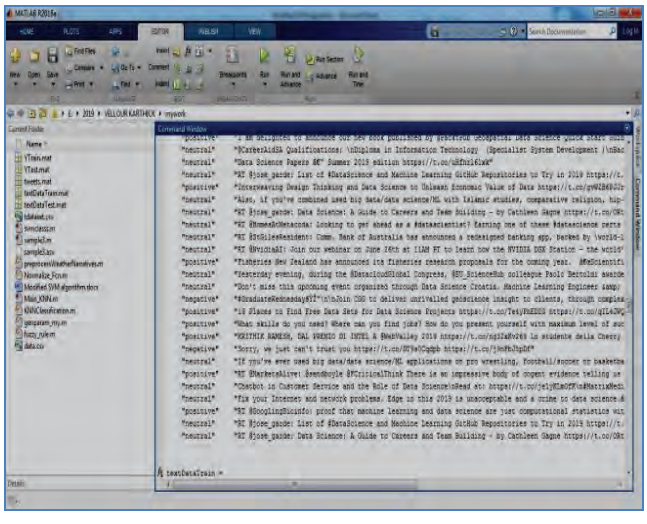

Figure 1. Input Dataset

Figure 2 shows the process of tokenization.

E-ISSN 2581 - 7957
P-ISSN 2277 - 3916

CVR Journal of Science and Technology, Volume 19, December 2020
DOI: 10.32377/cvrjst 1912

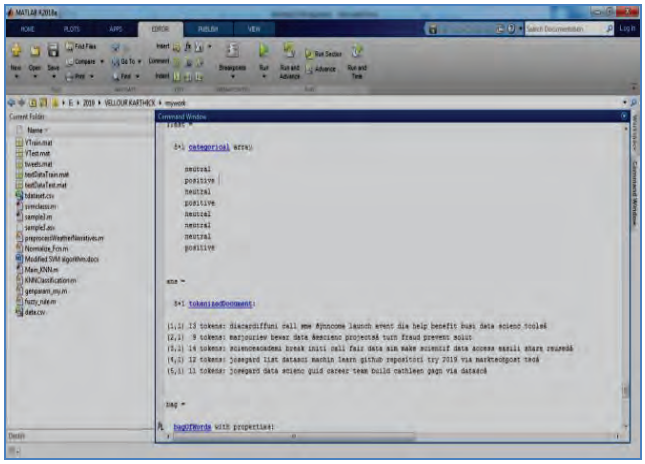

Figure 2. Text Preprocessing Tokenization

Figure 3 shows how Bag of Words (BoW) algorithm is applied for feature extraction.

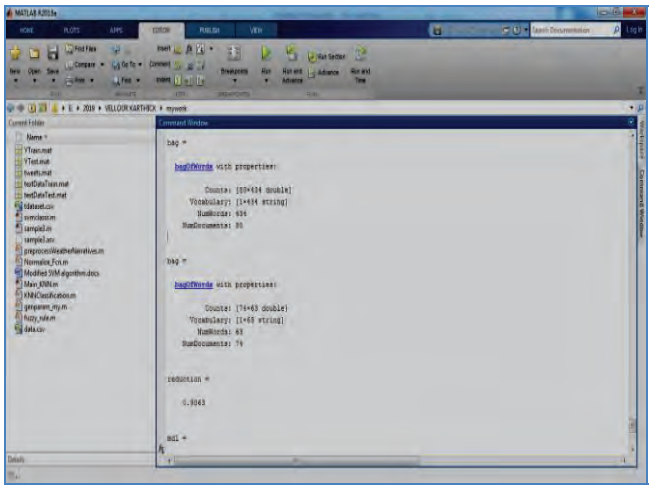

Figure 3. Bag of Words Detection

Figure 4 shows the Confusion matrix.

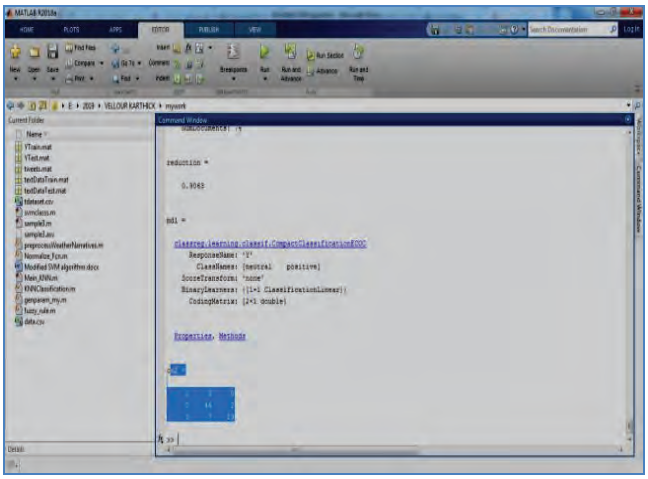

Figure 4. Classification Confusion Matrix

Figure 5 shows the count of positive and negative words.

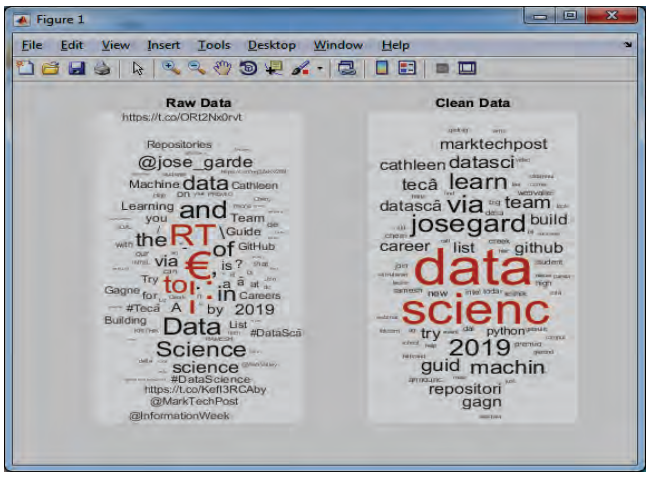


Figure 5. Word Count based on Positive and Negative Words

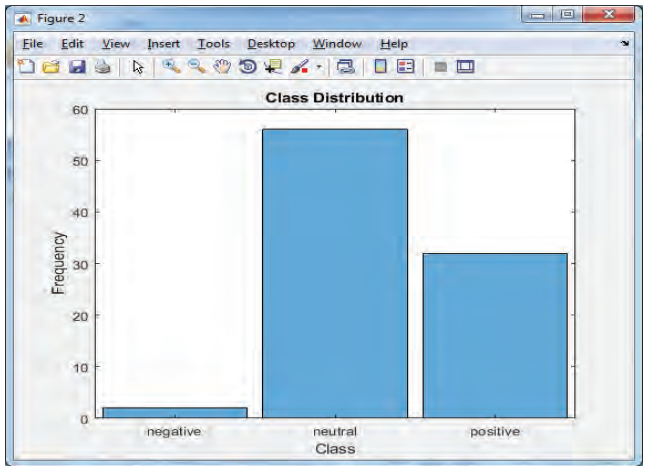


Figure 6. Classification

Figure 6 and Figure 7 show the classification classes and results respectively.

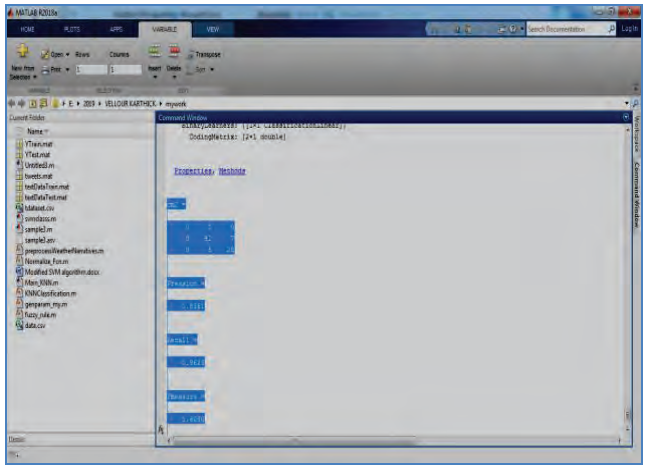

Figure 7. Classification Parameters

Figure 8 to Figure 11 show the accuracy, precision, recall and time period of the proposed method in contrast to the method without BAT optimization.

It is seen that the scheme with feature extraction using Bag of Words (BoW) algorithm and optimization using BAT algorithm yield better results.

DOI: 10.32377/cvrjst 1912

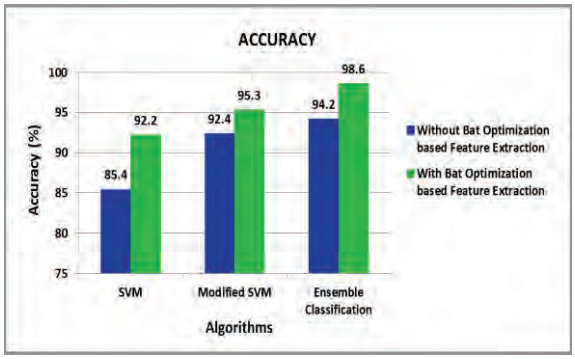


Figure 8. Accuracy

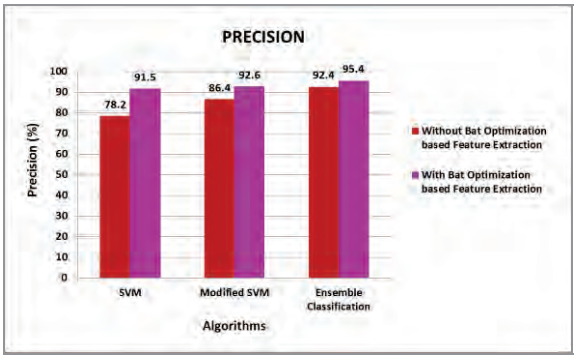


Figure 9. Precision

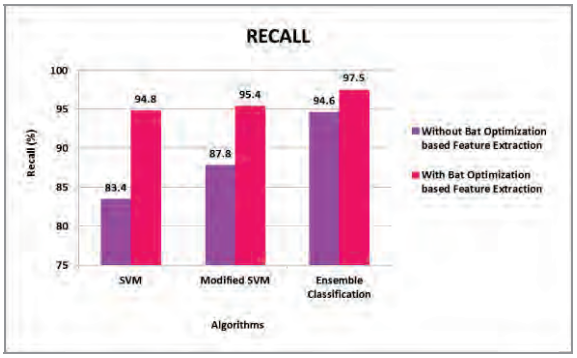


Figure 10. Recall

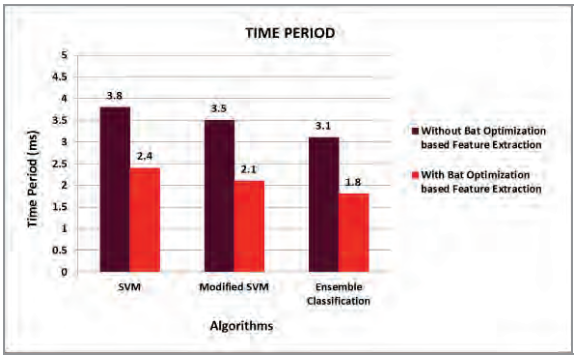


Figure 11. Time Period

## VIII. Conclusions

A novel method is proposed to classify customer reviews taken from Twitter. The Bag of Words (BoW) algorithm is used to extract relevant features. In addition, the data is classified using modified Support Vector Machine (SVM) and ensemble classification. Further, the results are optimized using BAT algorithm which improves the classification accuracy.

## References

[1] Kalra, V., & Agrawal, R. "Challenges of Text Analytics in Opinion Mining", Advances in Data Mining and Database Management, vol .1, no. 2, pp. 268-282, 2019.

[2] Ramteke, J., Shah, S., Godhia, D., & Shaikh, A., "Election result prediction using Twitter sentiment analysis", International Conference on Inventive Computation Technologies (ICICT), vol. 1, no. 1, pp. 45-56, 2016.

[3] Fang, X., & Zhan, J., "Sentiment analysis using product review data", Journal of Big Data, vol. 2, no. 1, pp. 67-78, 2015.

[4] Wang, H., & Castanon, J. A. "Sentiment expression via emoticons on social media", IEEE International Conference on Big Data (Big Data), vol. 1, no. 1, pp. 78-86, 2015.

[5] Ghosh, M., & Sanyal, G., "An ensemble approach to stabilize the features for multi-domain sentiment analysis using supervised machine learning", Journal of Big Data, vol. 5, no. 1, pp. 12-23, 2018.

[6] Singh, P., "Sentiment Analysis Using Tuned Ensemble Machine Learning Approach", Advances in Data and Information Sciences, vol. 1, no. 1, pp. 287-297, 2018.

[7] Shrote, K. R., & Deorankar, A., "Review based service recommendation for big data", 2nd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), vol. 1, no. 2, pp. 34-43, 2016.

[8] Nhlabano, V., & Lutu, P., "Impact of Text Pre-Processing on the Performance of Sentiment Analysis Models for Social Media Data", International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD), vol. 1, no. 2, pp. 65-74, 2018.

[9] Patil, M., & Darokar, M. S., "A Supervised Joint Topic Modeling Process Using Sentiment Analysis", Journal of Advances and Scholarly Researches in Allied Education, vol. 15, no. 2, pp. 720-725, 2018.

[10] Kaur, G., "Text Mining Based Approach to Customer Sentiment Analysis Using Machine Learning", Journal of Advances and Scholarly Researches in Allied Education, vol. 15, no. 6, pp. 58-65, 2018.

[11] Jiang, Z., Li, L., & Huang, D., "A general protein-protein interaction extraction architecture based on word representation and feature selection", International Journal of Data Mining and Bioinformatics, vol. 14, no. 3, pp. 276-291, 2016.

[12] Salloum, S. A., Al-Emran, M., & Shaalan, K., "Mining social media text: extracting knowledge from Facebook", International Journal of Computing and Digital Systems, vol. 6, no. 02, pp. 73-81, 2017.

[13] Shang, J., Liu, J., Jiang, M., Ren, X., Voss, C. R., & Han, J., "Automated phrase mining from massive text corpora", IEEE Transactions on Knowledge and Data Engineering, vol. 30, no. 10, pp. 1825-1837, 2018.

[14] Ramesh, B., & Sathiaseelan, J. G. R., "An advanced multi class instance selection based support vector machine for text classification", Procedia Computer Science, vol. 57, pp. 1124-1130, 2015.

[15] Vijayarani, S., Dhayanand, S., & Phil, M., "Kidney disease prediction using SVM and ANN algorithms", International Journal of Computing and Business Research (IJCBR), vol. 6, no. 2, 2015.

[16] Dadgar, S. M. H., Araghi, M. S., & Farahani, M. M., "A novel text mining approach based on TF-IDF and Support Vector Machine for news classification", IEEE International

Conference on Engineering and Technology (ICETECH), pp. 112-116, 2016.

[17] Mohammad, A. H., Alwada'n, T., & Al-Momani, O., "Arabic text categorization using support vector machine", Naïve Bayes and neural network, GSTF Journal on Computing (JoC), vol. 5, no. 1, pp. 108, 2016.

[18] Rana, S., & Singh, A., "Comparative analysis of sentiment orientation using SVM and Naive Bayes techniques", 2nd IEEE International Conference on Next Generation Computing Technologies (NGCT), pp. 106-111, 2016.

[19] Ali, F., Kwak, K. S., & Kim, Y. G., "Opinion mining based on fuzzy domain ontology and Support Vector Machine: A proposal to automate online review classification", Applied Soft Computing, vol. 47, pp. 235-250, 2016.

[20] Niu, D., & Dai, S., "A short-term load forecasting model with a modified particle swarm optimization algorithm and least squares support vector machine based on the denoising method of empirical mode decomposition and grey relational analysis", Energies, vol. 10, no. 3, pp. 408, 2017.

[21] Jadav, B. M., & Vaghela, V. B., "Sentiment analysis using support vector machine based on feature selection and semantic analysis", International Journal of Computer Applications, vol. 146, no. 13, 2016.

[22] Sabbah, T., Selamat, A., Selamat, M. H., Al-Anzi, F. S., Viedma, E. H., Krejcar, O., & Fujita, H., "Modified frequency-based term weighting schemes for text classification", Applied Soft Computing, vol. 58, pp. 193-206, 2017.

[23] Abidine, B. M. H., Fergani, L., Fergani, B., & Oussalah, M., "The joint use of sequence features combination and modified weighted SVM for improving daily activity recognition", Pattern Analysis and Applications, vol. 21, no. 1, pp. 119-138, 2018.

[24] Ren, Y., Zhang, L., & Suganthan, P. N., "Ensemble classification and regression-recent developments, applications and future directions", IEEE Computational intelligence magazine, vol. 11, no. 1, pp. 41-53, 2016.

[25] Yijing, L., Haixiang, G., Xiao, L., Yanan, L., & Jinling, L., "Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data", Knowledge-Based Systems, vol. 94, pp. 88-104, 2016.

[26] Pan, B., Shi, Z., & Xu, X., "Hierarchical guidance filtering-based ensemble classification for hyperspectral images", IEEE Transactions on Geoscience and Remote Sensing, vol. 55, no. 7, pp. 4177-4189, 2017.

[27] Gomes, H. M., Barddal, J. P., Enembreck, F., & Bifet, A., "A survey on ensemble learning for data stream classification", ACM Computing Surveys (CSUR), vol. 50, no. 2, pp. 23, 2017.

[28] Ju, C., Bibaut, A., & van der Laan, M., "The relative performance of ensemble methods with deep convolutional neural networks for image classification", Journal of Applied Statistics, vol. 45, no. 15, pp. 2800-2818, 2018.

[29] Kang, M., Ahn, J., & Lee, K., "Opinion mining using ensemble text hidden Markov models for text classification", Expert Systems with Applications, vol. 94, pp. 218-227, 2018.

[30] Heraguemi, K. E., Kamel, N., & Drias, H., "Multi-swarm bat algorithm for association rule mining using multiple cooperative strategies", Applied Intelligence, vol. 45, no. 4, pp. 1021-1033, 2016

[31] Alomari, O. A., Khader, A. T., Al-Betar, M. A., & Abualigah, L. M., "MRMR BA: a hybrid gene selection algorithm for cancer classification", J Theor Appl Inf Technol, vol. 95, no. 12, pp. 2610-2618, 2017.

[32] Alomari, O. A., Khader, A. T., Al-Betar, M. A., & Abualigah, L. M., "Gene selection for cancer classification by combining minimum redundancy maximum relevancy and bat-inspired algorithm", International Journal of Data Mining and Bioinformatics, vol. 19, no. 1, pp. 32-51, 2017.

[33] Sivic, J., & Zisserman, A., "Efficient visual search of videos cast as text retrieval", IEEE transactions on pattern analysis and machine intelligence, vol. 31, no. 4, pp. 591-606, 2008.

[34] Harris, Z. S., "Distributional structure", Word, vol. 10, no. 2-3, pp. 146-162, 1954.

[35] Ko, Y., "A study of term weighting schemes using class information for text classification", 35th International ACM SIGIR conference on Research and Development in Information Retrieval, pp. 1029-1030, 2012.

[36] Dietterich, T. G., "Ensemble learning, The handbook of brain theory and neural networks", MA Arbib, vol. 2, pp. 110-125, 2002.

[37] Yang, X. S., "A new metaheuristic bat-inspired algorithm", In Nature inspired cooperative strategies for optimization, pp. 65-74, Springer, Berlin, Heidelberg, 2010.

[38] Richardson, P., "Bats", Natural History Museum, London, 2008.