International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

A Swarm Intelligence Based Weighted Feature Extraction and Classification using SVM for Sentimental Exploration

P. V. Naga Srinivas, M. V. P. Chandra Sekhara Rao

Abstract: ***The goal of Sentiment Exploration (SE) is used for mining the accurate sentiments which are very beneficial for businesses, governments, and individuals, the opinions, recommendations, ratings, and feedbacks are becoming an important aspect in present scenarios. The proposed methodology likewise attempts to introduce a swarm intelligence based sentimental supervised methodology. In order to obtain a relevant feature data set from a large number of data samples, this method used particle swarm optimization to attain the utmost optimum feature set. The evaluation of the optimum feature set is obtained by means of using Minimum Redundancy and Maximum Relevancy measure as the fitness function. The categorization of the extracted feature set is accomplished with the Support Vector Machine classification technique. The experimental outcome for the suggested method is evaluated using four performance measure like precision, recall, accuracy, and f-measure and showed that proposed swarm intelligent based classification method has better performance using IMDB, Movie Lens and Trip Advisor Data Samples.***

Keywords: ***Sentimental Exploration, Particle Swarm Optimization, Minimum Redundancy Maximum Relevancy, Support Vector Machine, Classification, Feature Selection.***

I. INTRODUCTION

Social networks and other devices connected to the present world have increased the sources and knowledge drastically. Therefore, the capability of having computers to move at higher speed using innumerable information accessible and mining sentiment or thoughts would be significantly advantageous. SA also is known as Opinion Mining or polarity classification, which is the most deliberately investigated domain. Its goal is to examine people sentiments, feelings, methodes, etc., in the direction of dissimilar properties like themes, products, entities, organizations, and amenities. SA is an amalgamation of natural language processing and data mining tools. The foremost application of SA is employed in the polarity classification of movie, product, or shopping evaluations [2].

Manuscript published on November 30, 2019.
* Correspondence Author
P. V. Naga Srinivas*, Research scholar- CSE Department, University College of Engineering & Technology, Acharya Nagarjuna University, Guntur, Andhra Pradesh, India.
M. V. P. Chandra Sekhara Rao, Professor, CSE Department, RVR & JC College of Engineering, Guntur, Andhra Pradesh, India.
E-mail: srinivas.scet.ithod@gmail.com

This could be attained through supervised, semi-supervised, or unsupervised machine learning methodes and lexicon aided techniques [3]. The Supervised methodes are domain related, and categorized group of training and experimental information are obligatory for its even task [4]. SVM, Naïve Bayes, Neural Network, k-Nearest Neighbour, and Decision Tree are certain instances of supervised learning methodes [7, 8]. Alternatively, unsupervised methodes do not need any domain-related training information and typically depends on emoticons [9], a bag of words, or vocabulary methodologies [6]. Lexical aided sentiment exploration could be additionally categorized into vocabulary aided and corpus aided labels. Sentiment vocabularies are comprised of sentiment terms with preceding polarity scores, while corpus aided methodologies determine the possibility of a term co-appearing with positive or negative terms through enormously sized information from search appliances [5]. Numerous techniques have been already introduced for sentimental exploration under various categories using supervised/unsupervised methodologies. However, there exist numerous issues in the traditional classification methodes where, the algorithms even though extract the features from the documents and accomplish classification techniques, could not obtain accurate output at the end. The method results in more complexity with less accuracy and varies with noise in data. Therefore as to overcome these, the author proposed a robust method which is also not sensitive data. The proposed sentimental classification method prior to classification, particularly signifies on the selection of feature set from the large data samples. This is mainly accomplished by means of swarm intelligent technique known as particle swarm optimization. This novel methodology for sentiment exploration likewise has validated to be efficient both for normal transcripts and transcripts with a higher amount of noise.

The study on sentiment exploration specifies a worth amount of investigation that is attained through numerous authors depending on document-level sentiment categorization. In [13], the categorization of Chinese remarks depending on word2vec and SVM perf was suggested. Their methodology is dependent on two partitions. In initial partition, they employed word2vec device to identical group characters so as to seize the semantic characteristics in the preferred area. Subsequently, in the next partition, the lexicon aided and POS aided feature selection methodology is employed to produce the training information. Word2vec device assumes Continuous Bag-of-Words (CBOW) prototype and uninterrupted skip-gram prototype to study the vector illustration of terms in [14].

SVM perf is an execution of SVM for multiple variation performance methods that follows a substitutional structural design of SVM optimization issues for binary categorization given in [15].

A novel methodology is suggested to transform the textual information into lower magnitude emotional space (ESM) given in [18].

They have interpreted lesser dimension terms, which have specific and strong significance. They likewise employed Ekman Paul's investigation to categorize the terms into six elementary groups like annoyance, panic, disgust, sorrow, pleasure, and astonishment. They again have deliberated two diverse methods to allocate values to terms through expressive tagging. The overall weight of entire expressive labels are evaluated, and depending on these weights, the information is categorized into diverse classes. Even though its methodology harvests a noble outcome judiciously for the stock message board, the writers entitle that it could be employed on data sample or field. In [19], a Multi-View Sentiment Exploration (MVSA) is suggested comprising a group of textual image group using physical explanation gathered from Twitter. This methodology of sentiment exploration could be classified into two stages, such as lexicon aided and statistical learning. Considering, the lexicon aided study, a group of opinion terms or axioms are deliberated having priori determined sentiment value. Whereas in statistical learning, numerous machine learning methodes are employed using committed textual characteristics.

Machine learning methodologies are categorized as supervised and un-supervised classes. An estimation of the administered sentiment classification method is presented in [1]. An amalgamation of unigram and bigram characteristics accomplished better for movie and product review data samples. Feature selection methodes comprise of Chi-Square, Information Gain, Mutual Information, Odds Ratio, and Relevancy Score that are matched in the cases where mutual information is good. The estimation was performed on Czech language data samples. Even though the data samples are obtainable openly, it is fascinating to perceive the outcomes of English language data samples. A performance assessment for machine learning prototypes was performed in [10] who employed Multi-layer Perceptron (MLP), Naïve Bayes (NB), and SVM. SVM was perceived to accomplish finest on movie and product review data samples. Naïve Bayes presumes characteristics individuality that might not be the condition in every circumstance.

Sentiment exploration for big data was investigated [11], where Naïve Bayes classifier was exploited for textual classification. Movie review data samples are integrated to perform the implementation, and it was perceived that NB balances fine. The cause of selecting Naïve Bayes for big data categorization was not deliberated. State of the art assessment was likewise misplaced, making it complicated to validate the execution of the suggested method. A machine learning depending on feature selection technique was given in [12] where Information Gain and GA was employed to assist the needy. Naïve Bayes, Logistic Regression, SVM, Bagging, and Bayesian Boosting were employed on standard data samples for estimation. SVM accomplished in diverse experimental situations. The selection of individual categorization was not defensible. The authors did not clarify the motivation for employing Information Gain even though there are numerous other feature deduction techniques like Gain Ratio, Chi-Square, etc.

II. RELATED WORK

The previous work on sentiment exploration specifies a worth amount of investigation that is attained through numerous authors depending on document-level sentiment categorization. In [13], the categorization of Chinese remarks depending on word2vec and SVM perf was suggested. Their methodology is dependent on two partitions. In initial partition, they employed word2vec device to identical group characters so as to seize the semantic characteristics in the preferred area. Subsequently, in the next partition, the lexicon aided and POS aided feature selection methodology is employed to produce the training information. Word2vec device assumes Continuous Bag-of-Words (CBOW) prototype and uninterrupted skip-gram prototype to study the vector illustration of terms in [14]. SVM perf is an execution of SVM for multiple variation performance methods that follows a substitutional structural design of SVM optimization issues for binary categorization given in [15].

A novel methodology is suggested to transform the textual information into lower magnitude emotional space (ESM) given in [18]. They have interpreted lesser dimension terms, which have specific and strong significance. They likewise employed Ekman Paul's investigation to categorize the terms into six elementary groups like annoyance, panic, disgust, sorrow, pleasure, and astonishment. They again have deliberated two diverse methods to allocate values to terms through expressive tagging. The overall weight of entire expressive labels are evaluated, and depending on these weights; the information is categorized into diverse classes. Even though its methodology harvests a noble outcome judiciously for the stock message board, the writers entitle that it could be employed on data sample or field. In [19], a Multi-View Sentiment Exploration (MVSA) is suggested comprising a group of textual image group using physical explanation gathered from Twitter. This methodology of sentiment exploration could be classified into two stages, such as lexicon aided and statistical learning. Considering, the lexicon aided study, a group of opinion terms or axioms are deliberated having priori determined sentiment value. Whereas in statistical learning, numerous machine learning methodes are employed using committed textual characteristics.

Machine learning methodologies are categorized as supervised and un-supervised classes. An estimation of the administered sentiment classification method is presented in [1]. An amalgamation of unigram and bigram characteristics accomplished better for movie and product review data samples. Feature selection methodes comprise of Chi-Square, Information Gain, Mutual Information, Odds Ratio, and Relevancy Score that are matched in the cases where mutual information is good. The estimation was performed on Czech language data samples. Even though the data samples are obtainable openly, it is fascinating to perceive the outcomes of English language data samples. A performance assessment for machine learning prototypes was performed in [10] who employed Multi-layer Perceptron (MLP), Naïve Bayes (NB), and SVM.

SVM was perceived to accomplish finest on movie and product review data samples. Naïve Bayes presumes characteristics individuality that might not be the condition in every circumstance.

Sentiment exploration for big data was investigated [11], where Naïve Bayes classifier was exploited for textual classification.

Movie review data samples are integrated to perform the implementation, and it was perceived that NB balances fine. The cause of selecting Naïve Bayes for big data categorization was not deliberated. State of the art assessment was likewise misplaced, making it complicated to validate the execution of the suggested method. A machine learning depending on feature selection technique was given in [12] where Information Gain and GA was employed to assist the needy. Naïve Bayes, Logistic Regression, SVM, Bagging, and Bayesian Boosting were employed on standard data samples for estimation. SVM accomplished in diverse experimental situations. The selection of individual categorization was not defensible. The authors did not clarify the motivation for employing Information Gain even though there are numerous other feature deduction techniques like Gain Ratio, Chi-Square, etc.

III. PROPOSED SWARM INTELLIGENCE BASED FEATURE SELECTION AND CLASSIFICATION (SIFSC)

In this section, an Intelligent Supervised Semantic Exploration Method is introduced by means of employing the optimization algorithm for efficient feature extraction from large data samples. As mentioned in the literature study, there exists numerous techniques for classification and clustering the semantic information from the large data samples. However, in this section, the methodology focused primarily on the extraction of relevant terms or words from the semantic data to perform an accurate and efficient classification method. The architectural representation of the proposed method is given in figure 2. Thus to, accomplish an efficient feature extraction, an optimized methodology such as an evolutionary algorithm aided Particle Swarm Optimization is employed. The proposed method is specifically divided into three modules. They are:

- Pre-processing
- Feature Extraction
- Classification

3.1 Pre-processing

The obtained or collected information is initially given to the pre-processing module so as to transform this data into a further appropriate form of natural language processing as to eventually attain a qualitative sentiment exploration. The following are the steps to be performed:

- Process English language: The proposed method essentially performs its testing and evaluation only on the English language. Consequently, information present in any other language needs to be removed. The WordNet6 is an English language dictionary which is generally used for semantic exploration. This tool is employed to ensure that the data samples comprising merely those terms that are present in WordNet vocabulary and which are predominantly in English. Any term in the data samples is discarded if it does not exist in the WordNet tool.
- Stemming and lemmatization: Stemming is the method of minimizing the terms to its root whereas lemmatization is an method of stemming that performs spell rectification to the stemmed term so as to evade the inconsistency that might come up whenever a vocabulary lookup is accomplished. Porter Stemmer is employed so as to stem the term, and further Word Net and JSPell are used for spelling rectification.
- Spelling correction: So as to obtain better performance, spelling mistakes need to be corrected. This is achieved through incorporating JSpell for spell rectification.
- Removal of stop words, query words, URLs, special characters: The terms that are utmost frequently employed could not be used for sentiment categorization. Consequently, the stop words and query words ought to be detached for additional processing of feature extraction. This minimizes the dimension and complexity of the proposed method. These are acknowledged through a group of stopping words and query words. So as to abandon the stopping words and query words, the term grouped are encumbered from the corresponding documents, and every term from data sample is matched to these group prior it is employed for additional processing. If any of the terms are discovered in this group, it is detached from the data sample. URLs and special characters are recognized through regular expressions and are likewise detached prior to the additional process.
- Enlarge abbreviations and Substitute jargons: The abbreviations need to be enlarged and the jargons to be substituted with the appropriate terms/phrases so that any inconsistency is circumvented whenever a vocabulary is accomplished. The abbreviation and jargon vocabulary is employed for this method. This vocabulary is employed to discover the jargons and abbreviations and subsequently substitute them with the entire term and accurate spell word.
- Part of speech tagging: Part of speech knowledge is important whenever the sentiment recognition is employed since any term employed like a diverse part of speech, could have diverse implications completely. This study employs Stanford POS tagger to recognize part of speech data in the documents beneath contemplation.

3.2 Feature Selection

The part of speech labeled and pre-processed information is now sent to the subsequent feature selection module. Feature selection is a complex job due to feature collaborations and the huge exploration domain. Feature selection purpose is to pick a subset of appropriate attributes which are essential and adequate to define the target notion [20]. Through minimizing the inappropriate and replicated features, feature selection would minimize the magnitude, diminish the quantity of information and preserve merely the essential features for the learning, truncate the execution period, streamline the configuration as to enhance the efficiency of the learned categorizer [20].This module is accomplished by means of particle swarm optimization (PSO) method.

PSO [21, 22, 26] is a comparatively current Evolutionaryiv. Computation method depending on swarm intelligence. The fundamental notion of PSO is that information is augmented through social collaborations, where the collaboration is not merely individual; nevertheless, it is global [22, 23].

Feature selection has a huge exploration domain that frequently causes the issue of becoming trapped in local optimum in prevailing methodes. Thus, it requires a universal exploration method. EC methodologies are famous for their universal exploration capability. PSO is an EC method which is capable of exploring huge domains to detect optimum or near-optimum outcomes efficiently.

Particle swarm optimization (PSO) [21] is a swarm intelligence method motivated through communal behavior of birds assembling or fish schooling [21]. In PSO, every individual of the issue is signified as an element that is encrypted through a vector or an array. Individuals roam in the exploring domain to explore for the optimum results. In the course of this association, every individual could recollect its finest experience. The complete group explores the optimum outcomes through modernizing the location of every individual depending on the finest experience of its individual and its adjoining individual [23]. PSO is a humble, however authoritative exploration procedure, which has been positively smeared to resolve issues in multiple regions. Figure 1 characterizes the workflow method of PSO for feature selection.

Steps for Feature Selection using PSO:

i. Initialization of population: The size of the individuals or swarm is initialized by selecting the pre-processed terms in the document that need to be employed for classification along with the target labels that is assigning memories to each swarm. These are known as candidate solutions or particles.

ii. Initializing the term location and velocity: Every term in the document has a location in the exploration domain that is signified using a vector $x_i = (x_{i1}, x_{i2}, \dots \dots x_{iD})$, where D is the magnitude of exploration domain. Individual roam in the exploration domain to pursuit for the optimum outcomes. Thus, every individual has a velocity, which is signified as $v_i = (v_{i1}, v_{i2}, \dots \dots v_{iD})$.

iii. Fitness Evaluation: Minimum Redundancy Maximum Relevance (MRMR) measure is employed to filter the relevant terms from the pre-processed data sample. The fitness of the terms that are considered as particles in the PSO algorithm is used in this step. MRMR uses top-ranking features based on mutual information without considering the relationship between the features. This measure provides the feature or the terms that are minimally repeated in the document and has maximum relevance with the target labels for classification. As to evaluate this measure, the mutual information amongst the terms is estimated to determine the redundant terms, and mutual information amongst the feature and the target label is estimated to determine the relevant terms. This is given as follows:

$Min\, W_1, \quad W_1 = \frac{1}{|S|^2}\sum_{i,j\epsilon S} I(i,j), \quad Max\, V_1, \quad V_1 = \frac{1}{|S|}\sum_{i,\epsilon S} I(h,j)$

S is the group of attributes and $I(i,j)$ is the mutual information amongst the features i and j, h is the target group, then $MRMR = \max(V_1 - W_1)$.

v. Update term position and velocity: During the movement, each term in the document apprises its location and velocity pertaining to its individual capacity and that of its adjacent. The finest preceding location of the individual is noted as the individual best p_{best}, and the finest location attained through the populace thus away is known as g_{best}. Depending on p_{best} and g_{best}, PSO examines for the optimum outcome through updating velocity and the location of every individual pertaining to the given equations:

$$x_{id}{}^{t+1} = x_{id}{}^{t} + v_{id}{}^{t+1} \quad (1)$$

$$v_{id}{}^{t+1} = w * v_{id}{}^{t} + a_1 * ran_1 * (p_{id} - x_{id}{}^{t}) + a_2 * ran_2 * (p_{gd} - x_{id}{}^{t}) \quad (2)$$

where t signifies the t^{th} generation in an evolutionary method. $d\epsilon D$ signifies the d^{th} dimension in the exploratory domain. w is inertia weight that is employed to regulate the influence of the preceding velocities on the present velocity. a_1 and a_2 are acceleration coefficients. ran_1 and ran_2 are arbitrary values consistently disseminated in [0, 1]. p_{id} and p_{gd} signify the attributes of p_{best} and p_{best} in the d^{th} magnitude. The velocity is restricted through a pre-determined maximal velocity, v_{max}, and $v_{id}{}^{t+1}\epsilon[-v_{max}, v_{max}]$.

Termination Criterion: This is nothing but the termination criteria for the PSO method. This methodology terminates whenever a predetermined condition is reached that might be a decent fitness value or a pre-determined extreme number of generation t specified through the individual.

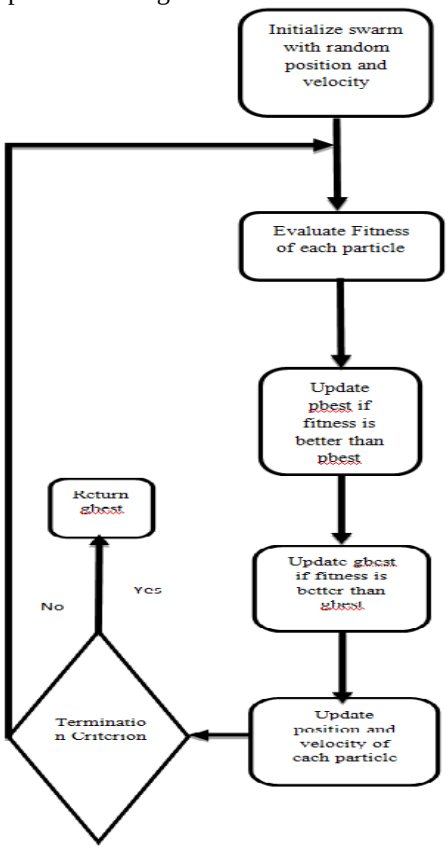

Fig 1: Work Flow Process of Particle Swarm Optimization Algorithm.

3.3 Classification Using Linear SVM

The feature obtained from the feature selection module is used for further classification of the data samples. The class labels or target class of the data sample are of five categories depending on the polarities like completely negative polarity, negative polarity, neutral, positive polarity, and completely positive polarity. The complete samples are further segregated into a training set and testing set where the features of testing samples are extracted from the feature selection module employing the Particle Swarm Optimization Algorithm. The training data set with class labels are employed to construct a classification module upon which the testing data samples are employed to classify the documents pertaining to the featured terms.

Support Vector Machine (SVM) is employed for the categorization of the data sample. SVM plots entire numeric vectors in the domain and specifies decision limitations through hyperplanes. This hyperplane splits the vectors into two groups such a way that, the distance from every category to the hyperplane is maximum. The given equation is employed as the discriminatory function for the SVM prototype.

$$g(x) = W^T \phi(x_i) + b \qquad (3)$$

here w is the weight vector, b is the prejudice and $\phi(x_i)$ is the input domain to higher magnitude attribute domain non-linear representation. These constraints are obtained mechanically through the exploited margin principle on the experimented data sample.

$$min \frac{1}{2} W^t w + C \sum_{i=1}^{N} \mathcal{E}_i \, , y_i (W^T \phi(x_i) + b) \geq 1 - \mathcal{E}_i \;\; and \;\; \mathcal{E}_i \geq 0 \; for \; N = 1, 2, \ldots N \qquad (4)$$

where $\mathcal{E}_i$ and C signifies the relaxed element and penalty constant correspondingly. The issue is resolved through quadratic enhancement by means of Lagrange multipliers where support vector is determined through training occurrence satisfying $x_i > 0$. Having an overview of the kernel function, the discriminant function is defined as

$$g(x) = \sum_{i=1}^{N} \phi(x_i) \, y_i K(x_i, x) \qquad (5)$$

The feature domain is huge for any text categorization job. Thus linear kernel is generally employed.

IV. RESULTS AND DISCUSSIONS

The Experimental Outcomes for the proposed method is carried out using one data set, which is the movie reviews from the Internet Movie Database (IMDB). The experimental results of the suggested methodology are matched with the conventional sentimental classification using SVM, Sentimental Classification with the pre-processing module, and SVM method. Four different measures are used so as to estimate the efficiency of the method used in the proposed method. They are:

1. Precision: this is also named as the positive predicted value. This measures the portion of retrieved instances that are appropriate. This is specified as:
$$precision = \frac{TP}{TP+FP} \qquad (6)$$

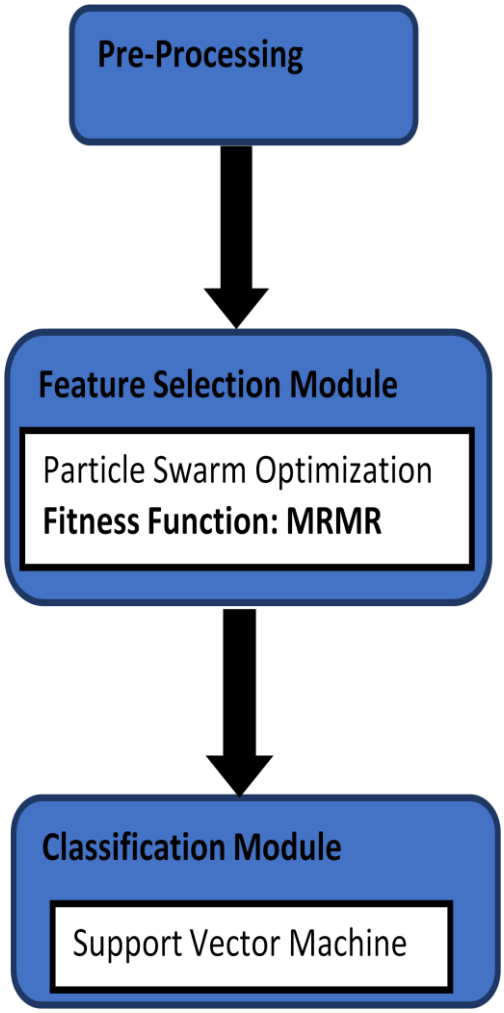

Fig 2: Block diagram of Proposed Swarm Intelligence based Feature Selection and Classification (SIFSC)

1. Recall: This is also named as sensitivity, which is the portion of appropriate instances that are retrieved. This is specified as:
$$recall = \frac{TP}{TP+FN} \qquad (7)$$
2. Accuracy: This metric defines the general proportion of accurately categorized instances, irrespective of its kind. This metric is the one employed to estimate the efficiency of sentiment exploration. This is specified as:
$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \qquad (8)$$
Where TP, TN, FP, and FN represent the true positive, true negative, false positive, and false negative.
3. F-Measure: This recurrently employed metric allows factoring precision and recall to a unique value, therefore signifying its harmonic mean. The metric is evaluated through the given formula:
$$F - measure = \frac{(1+\beta^2) \times recall \times precision}{(\beta^2 \times recall) + precision} \qquad (9)$$

where β could be employed to specify higher weight in the score to either the precision or the recall, in this study, $\beta = 1$ is used, therefore giving the equivalent load to the two metrics.

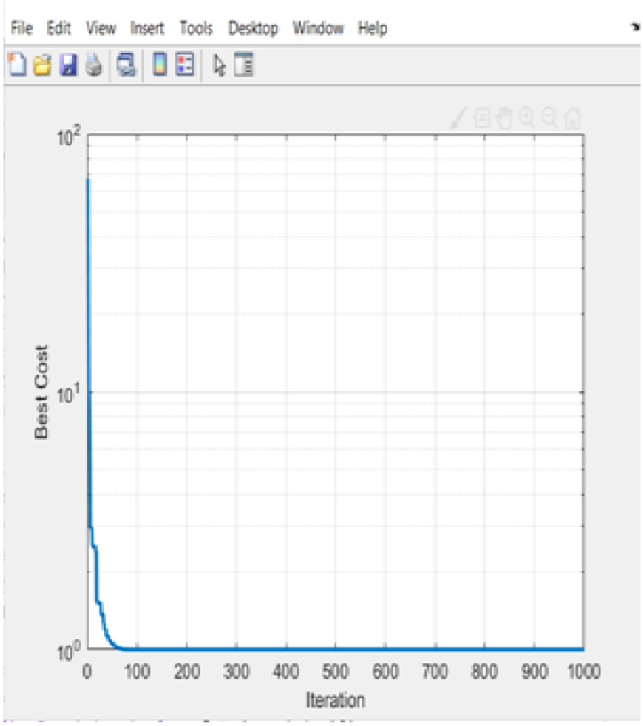

Fig 3: Differentiation of Fitness Value with No. of iteration for IMDB Data Sample

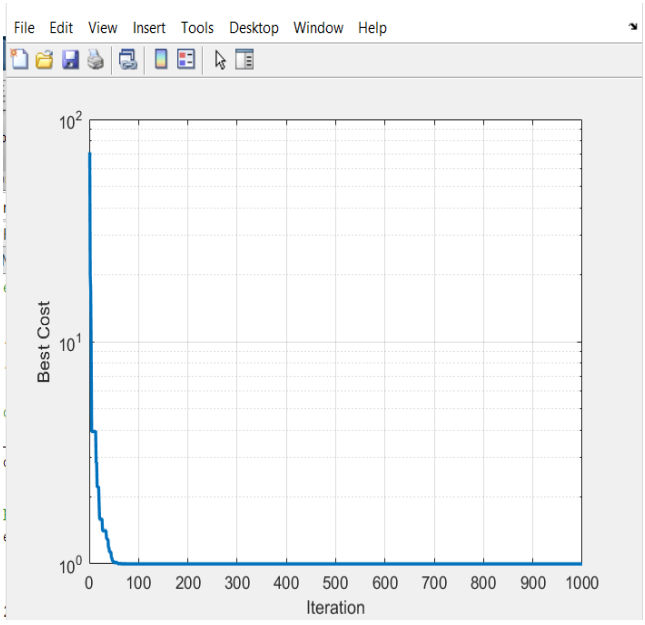

Fig 4: Differentiation of Fitness Value with No. of iteration for Movie Lens Data Sample

Movie Review: The Movie Review Data is a collection of movie reviews retrieved from the imdb.com website in the early 2000s by Bo Pang and Lillian Lee. The reviews were collected and made available as part of their research on natural language processing. The dataset is comprised of 1,000 positive and 1,000 negative movie reviews drawn from an archive of the rec.arts.movies.reviews newsgroup

IMDB: This Dataset contains information about 14,762 movies. Information about these movies was downloaded with wget for the purpose of creating a movie recommendation app. This is a dataset for binary sentiment classification containing substantially more data than previous benchmark datasets. We provide a set of 25,000 highly polar movie reviews for training and 25,000 for testing. There is additional unlabelled data for use as well.

Trip Advisor: Three equally experienced annotators provided sentence-level annotations of a subset of 500 randomly selected reviews from the publicly available TripAdvisor dataset. The full TripAdvisor dataset consists of 235,793 hotel reviews crawled over a period of one month. Full reviews of hotels in 10 different cities (Dubai, Beijing, London, New York City, New Delhi, San Francisco, Shanghai, Montreal, Las Vegas, Chicago). There are about 80-700 hotels in each city. Extracted fields include date, review title, and the full review. A total number of reviews: ~259,000.

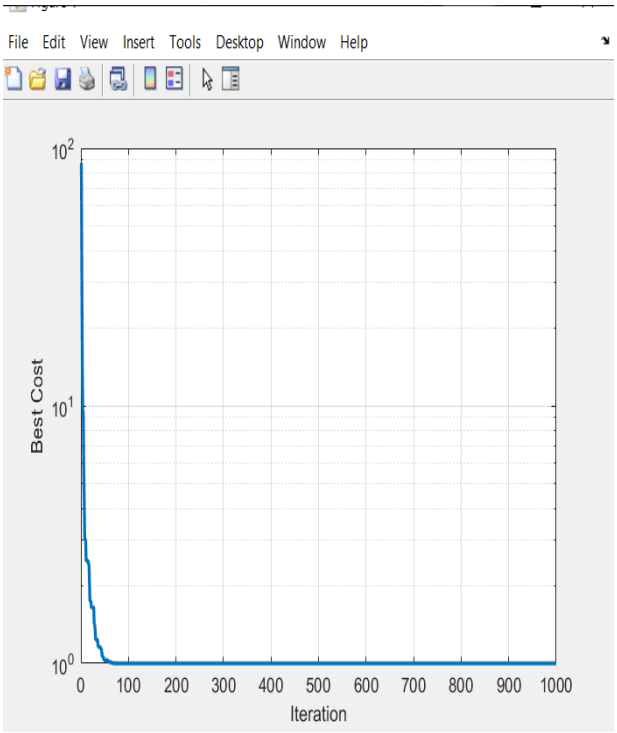

Fig 5: Differentiation of Fitness Value with No. of iteration for Trip Advisor Data Sample

Fig 3, Fig 4 and Fig 5 represents the Differentiationof Fitness Value against the No of Iterations for Particle Swarm Optimization Method for all the three data samples. From the fig, it can be observed that the cost value is increasing the number of iterations for the algorithm is growing linearly.

Table 1: Differentiation of Performance Measure of the Movie Review Database

Performance Measure	IMDB Data Sample	Movie lens Data Sample	Trip Advisor Data Sample
Accuracy	86.23	98.87	85.96
Precision	72.41	98.38	72.39
Recall	100	99.35	100
F-Measure	84.00	98.86	83.98

Table 1 showed the precision, recall, and f-measures of the three different Data Samples using the proposed Classification method. From table 1, it is clearly observed that movie lens data samples have higher accuracy, precision, recall, and F-Measure value compared to the other two data samples using the proposed swarm intelligence-based. Sentimental classification method is higher when compared to the other two existing methodes. Similarity it is perceived that SVM aided Sentimental Algorithm along with the pre-processing has higher performance compared to the traditional method since some part of noisy data is removed by means of pre-processing. Finally, the proposed method has higher performance as the dimensionality of the proposed method is reduced by selecting the optimized and relevant features and minimizing the redundancy between the features through swarm intelligent technique such particle swarm optimization. Here the computational complexity is also minimized as the dimensionality is optimized. It is witnessed that the accuracy of the suggested methodology is higher for both the data sets considered.

V CONCLUSIONS

In this paper, a swarm intelligence based sentimental classification algorithm is employed to obtain accuracy when compared with the existing sentimental classification algorithms. In this paper, foremost concentration is given on the appropriate selection of features from the large data sample by carefully evaluating the terms in the data samples by means of the fitness estimation. This fitness evaluation that is employed in this method is the minimum redundancy and maximum relevancy measure in the particle swarm optimization. Initially, the pre-processing of the documents is accomplished followed relevant feature selection and classification using support vector machine. The experimental results of the suggested method are carried out on three different data samples such as IMDB, Movie Lens, and Trip Advisor, whose performance is evaluated using precision, recall, f-measure, and accuracy. It is shown that this method has good performance compared with the other existing methodes.

REFERENCES

1. I. Habernal , T. Ptáˇcek , J. Steinberger , Supervised sentiment exploration in Czech social media, Inf. Process. Manag. 50 (5) (2014) 693–707.
2. M.D. Molina-González , E. Martínez-Cámara , M.T. Martín-Valdivia , L.A. Ure- ña-López , A Spanish semantic orientation method to domain adaptation for polarity classification, Inf. Process. Manag 51 (2015) 520–531 .
3. W. Medhat , A. Hassan , H. Korashy , Sentiment exploration algorithms, and applications: A study, Ain Shams Eng. J. 5 (4) (2014) 1093–1113 .
4. H. Saif , Y. He , M. Fernandez , H. Alani ,Contextual semantics for sentiment exploration of Twitter, Inf. Process. Manag. 52 (1) (2016) 5–19 .
5. K. Ravi , V. Ravi , A study on opinion mining and sentiment exploration: tasks, methodes and applications, Knowl. Based on Syst. 89 (2015) 14–46 .
6. M. Taboada , J. Brooke , M. Tofiloski , K. Voll , M. Stede , Lexicon-based methods for sentiment exploration, Comput. Ling. 37 (2) (2011) 267–307 .
7. H. Kang , S.J. Yoo , D. Han , Senti-lexicon and improved Naive Bayes algorithms for sentiment exploration of restaurant reviews, Expert Syst. Appl. 39 (2012) 60 0 0–6010 .
8. N. Fan , Y.S. An , H.X. Li , Research on analyzing sentiment of texts based on a k-nearest neighbor algorithm, Comput. Eng. Des. 33 (3) (2012) .
9. F.H. Khan , S. Bashir , U. Qamar , TOM: Twitter opinion mining framework using a hybrid classification scheme, Dec. Support Syst. 57 (2014) 245–257 .
10. P.K. Singh , M.S. Husain , Methodological study of opinion mining and sentiment exploration techniques, Int. J. Soft Comput. 5 (1) (2014) 11 .
11. B. Liu , E. Blasch , Y. Chen , D. Shen , G. Chen , Scalable sentiment classification for big data exploration using naive Bayes classifier, in Proceedings of the IEEE International Conference on Big Data,,IEEE, 2013, October, pp. 99–104 .
12. P. Kalaivani, K.L. Shunmuganathan, Feature reduction based on genetic algorithm and hybrid model for opinion mining, Sci.Program. vol. 2015 (2015) 15 Article ID 961454, DOI: 10.1155/2015/961454 .
13. Zhang, D. Xu, H., Su, Z., & Xu, Y. (2015). Chinese comments sentiment classification based on word2vec and SVM perf. Expert Systems with Applications, 42, 1857–1863.
14. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
15. Joachims, T. (2006). Training linear svms in linear time. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 217–226). ACM.
16. Liu, S. M., & Chen, J.-H. (2015). A multi-label classification based method for sentiment classification. Expert Systems with Applications, 42, 1083–1093.
17. Zhang, M.-L., & Zhou, Z.-H. (2007). Ml-knn: A lazy learning method to multi-label learning. Pattern recognition, 40, 2038–2048.
18. Luo, B., Zeng, J., & Duan, J. (2016). Emotion space model for classifying opinions in the stock message board. Expert Systems with Applications, 44, 138–146.
19. Niu, T., Zhu, S., Pang, L., & El Saddik, A. (2016). Sentiment exploration on multi-view social data. In MultiMedia Modeling (pp. 15–27). Springer.
20. M. Dash and H. Liu (1997). Feature selection for classification. Intelligent Data Exploration, vol. 1, no. 4, pp. 131–156.
21. J. Kennedy and R. Eberhart (1995). Particle swarm optimization. In IEEE International Conference on Neural Networks, vol. 4, pp. 1942–1948.
22. Y. Shi and R. Eberhart, (1998), A modified particle swarm optimizer. In IEEE International Conference on Evolutionary Computation (CEC'98), pp. 69–73.
23. J. Kennedy, R. C. Eberhart, and Y. Shi (2001), Swarm Intelligence. Evolutionary Computation Series. San Francisco: Morgan Kaufman.
24. H. Wang, Y. Lu, C. Zhai (2010). A latent aspect rating exploration on review text data: a rating regression method. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM.
25. B. Pang, L. Lee, A, (2004). A sentimental education: sentiment exploration using subjectivity summarization based on minimum cuts. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, Barcelona, Spain, p. 271.
26. Dr. M. V. P. Chandra Sekhara Rao et al., (2013). Swarm Optimization Algorithm for Privacy-Preserving in Data Mining. IJCSI International Journal of Computer Science, Vol. 10, Issue 2, No 3.

AUTHORS PROFILE

P V Naga Srinivas, received B. Tech. degree in Computer Science and Engineering from JNTU, Hyderabad, India, in 2003. He received M. Tech. degree in Computer Science and Engineering from JNTU, Hyderabad, in 2006. His main area of interests in research are Data mining, Machine Learning and Deep Learning. Presently a Research scholar- CSE Department, University College of Engineering & Technology, Acharya Nagarjuna University, Guntur, Andhra Pradesh, India.

Dr. M.V.P. Chandra Sekhara Rao, Working as Professor, RVR & JC College of Engineering, Guntur-19, A.P. Research Interests: Data Warehousing and Data Mining