ISSN (ONLINE): 2456-5717

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Vol.3, Issue.9, September 2017

A Novel Framework for Mining Sentiments Using Hybrid Classification Algorithm

[1]Nagamanjula R, [2]A. Pethalakshmi
[1]Research Scholar, Mother Teresa Women's University, Kodaikanal, Tamilnadu, India.
[2]Associate Professor and Head, M.V.Muthiah Government Arts College for Women, Dindigul, Tamilnadu, India.

Abstract:- Opinion Mining or Sentiment Mining is a kind of Natural Language Processing (NLP) which supports the tracking of moods from the users about a particular product. It mainly focuses to collect the information of products and categorize the opinions of product. A marketer uses this information for evaluating the success rate of specific product and enhances the next versions of products. Here main challenge of opinion mining is that detecting the optimal reviews of customers and analyzing from newer sources. Many existing researchers proposed several ways for sentiment analysis but their performance doesn't show the stability of emotional symbols. To solve these problems, we present a novel framework using hybrid classification algorithm called Modified Genetic with Particle Swarm Optimization (MG-PSO). We also involve preprocessing, aspects extraction, sentiment analysis before classifying the dataset. We experimentally evaluate our proposed system and provide effective result by analyzing with performance metrics such as fitness function value, accuracy and computation time.

Index Terms:- Opinion or Sentiment Mining, Genetic Algorithm (GA), Particle Swarm Optimization, Aspects.

I. INTRODUCTION

Nowadays, there is an increasing usage of internet and online activities by many users on real world. This leads to extract, transform, analyze and load large data based on several data mining approaches. Here opinion mining from data mining is one of the approaches for extracting opinions of several users' information from numerous blog spots, social networks, e-commerce websites and several news reports. This is related to customer reviews and provides the summary of opinion for a specific product [1]. Sentiment analysis is that works based on natural language processing for tracking the moods or reviews of peoples about a particular commercial product. There are several challenges are there for analyzing the sentiments such as 1) In a comment or review, a word is considered as positive in one situation and negative is next situation [2]. For example, consider a sentence "A smart is looking good" is different from "A smart phone is not looking good". 2) Many people don't express their feelings in single way. Normally people provide comment on any online environment like Facebook, twitter, blogs, etc. In order to buy a product, at that time opinion mining guides the specific customer by providing the questions like "which brand of best phone, I should buy at lower cost", "which camera should I buy for getting higher clarity", "which novel is the best novel for purchasing", etc. Here opinionated data are fundamentally specifying the internal emotions of a person for a product by the user. These

opinions are the major source for evaluating the product's success or failure and that also plays a major role for performance evaluation [3]. Figure 1 specifies the nature of opinion mining.

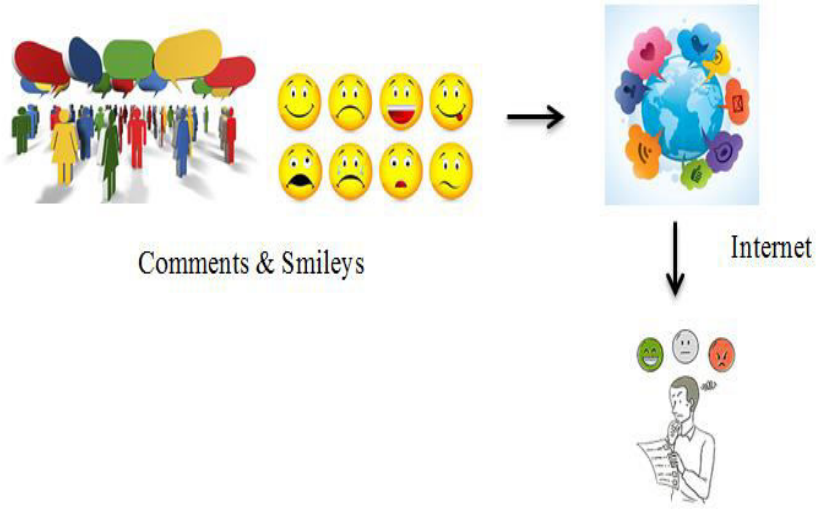

Fig 1. Nature of Opinion Mining

Using the opinion of a customer, the products behavior is analyzed for the evaluation and makes the important decision for the organization. This organization allows several approaches for reviewing the documents. The approaches are natural language processing, machine learning algorithms like maximum entropy, Support Vector Machine (SVM), K-nearest Neighbor (KNN), decision tree algorithms, Genetic Algorithm (GA), Particle Swarm Optimization (PSO) and Naïve Bayes. These supervised machine learning algorithms are used for gathering user reviews, emotions and opinions in the form of Neutral, Positive and Negative [4]. Aspects are the main task on the sentimental analysis process. The main complicated task for sentimental analysis is that only way to predicting the aspects of a sentence. This sentimental analysis model is split into two types such as a) aspect Model and b) sentimental model. Aspect model works in sentence vector and outputs a probabilistic distribution over the aspects whereas the sentimental model involves its process on sentence and its outputs on a probabilistic distribution through polarities [5]. There are several challenges and difficulties in opinion mining that mainly arise on languages which is main problem for understanding according to grammatical problems on several languages. Usually grammar-noun words are also said to be featured words whereas the verbs and adjective can also be used as a feature words that is main challenge for judging because adjective words like "great" are used for both positive sentence and negative sentence. Human are allowed to write the feedback in any format and they are free to write in any of the manner [6]. Users are allowed to use symbols, abbreviations, small letters, capital letters, shortcut methods (such as word "before" is shorted as "b4"), regional language based feedbacks such as pictures as pics, you can be written as "U", etc. Several application domains of opinion mining involve 1) shopping (E-commerce), 2) Entertainment such as movies reviews, 3) Business, 4) Research and Development (R & D), 5) Health, 6) Academics, 7) Transportation and 8) Politics.

Our proposed work involves a hybrid machine learning based classification approach for sentiment analysis of review comments of a product. Here we involve four phases on our proposed

framework for opinion mining. The first phase is preprocessing which involve the removal of repeated words, special characteristics, URL, audio, video and question marks. The second phase involves extraction of aspects. Here we split the joined word, tenses, Parts of speech, words spell check, and Lexicons. Next in third phase, we analyze the strength based on positive comments, negative comments, neutral comments and emotions using Sentiment Analysis Algorithm (SAA). This SAA provides the results in ranges for positive comments has 70 % - 100 %, Negative comments ranges below 40% and Neutral comments ranges from 40 % - 70 %. In the final phase, we involve MG-PSO classification of sentiments using the strength results.

The main contribution of our proposed work is as follows:

- Proposed a novel Sentimental Analysis Algorithm for analyzing the strength of the user comments
- Proposed a novel MG-PSO classification algorithm for comments.

Paper Organization

Section 2 has the existing techniques and its literatures that are done on opinion mining. Section 3 describes the problem definition for classification mechanism. Section 4 specifies out proposed framework and its performance evaluation are done with existing system are described on section 5 with graphical results. In section 6, we conclude our proposed work.

II. LITERATURE SURVEY

Avinash Chandra Pandey, Dharmveer Singh Rajpoot, Mukesh Saraswat proposed a twitter sentiment analysis based on hybrid cuckoo search method [7]. Here the authors used data set twitter data for analyzing the sentiments. The twitter data sets are subjective in nature, and they proposed a novel metaheuristic method called CSK which is based on K-mean and Cuckoo Search algorithm. This method supports for finding optimum cluster heads from sentimental contents of twitter dataset. The feature extraction methods results in several features on twitter dataset.

A sentiment analysis was proposed in paper [8] by identification of human agent interaction and discussed about the growing opportunities of cross-disciplinary work which increases individual advances. This opinion detection method allows human agent interactions at rare manner and designed for socio affective interactions such as sentiment analysis as input, timing constraint for the interactions, output for interactions. They also proposed a comparative study that analyzes sentiment related phenomena and sentiment detection methods.

In paper [9], authors proposed a deep learning and sub-tree mining and a document level sentiment classification mechanism. This paper combines the deep learning methods and sub tree mining in order to solve the sentiment classification problem. Here the Stanford parser was used for extracting the relations from the starting and ending of the sentences. Finally the sentences are represented as the tree. Find best sub-tree algorithm is proposed with sub-tree mining algorithm that eliminates the outliers and Depth first search was used for arranging the sentences in their order.

III. PROBLEM DEFINITION

Opinion mining involves many classification algorithms and several machine learning approaches for analyzing the sentiments. A dictionary based classification algorithm was proposed in [10] that allows sentiment and polarity levels analyzing the dataset. This results in lesser accuracy, recall, precision, F-score and error rate based on confusion matrix while comparing with Naïve Bayes and SVM. In paper [7], a twitter dataset based sentiment analysis was proposed based on k-means clustering algorithms and cuckoo search algorithm. Here the K-means clustering algorithm takes more time for selecting the centroid points based on its iteration process. K-mean clustering algorithm is processed with Euclidean distance, so the accuracy of whole process is reduced.

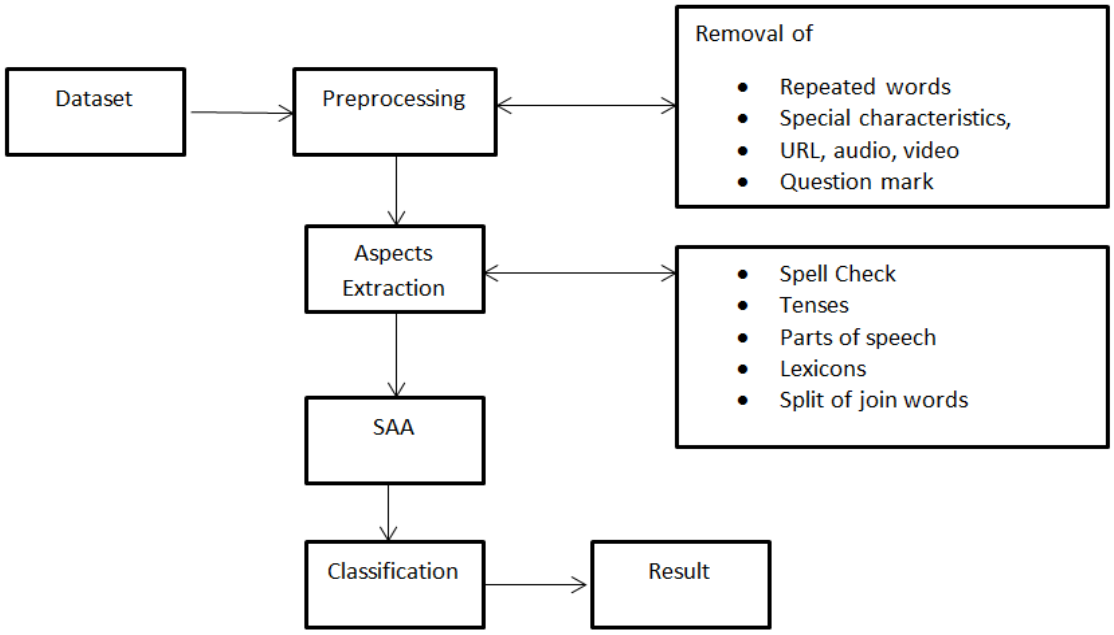

Fig 2. Proposed Framework for Sentiment Analysis

IV. PROPOSED WORK

In our proposed framework, we involve mainly four phases for sentiment analysis. Initially we involve preprocessing procedure for the twitter dataset. This twitter data set contains several positive comments and negative comments for several products. We involve mainly four phases on our proposed framework. Here preprocessing is the initial phase that allows removal of question tags, URL, audios, Videos and special characters in a sentence. Then the second phase is that extraction of aspects on the sentences which involves splitting of joined words, correction of spelling and error corrections, tenses, parts of speech and lexicons. Next third phase is that, involves our proposed SAA algorithm that analyses the positive reviews and negative reviews on the aspects. Finally we classify the dataset based on hybrid classification approach named "Modified Genetic Particle Swarm Optimization" algorithm. Thus our proposed approach results

in effective performance metrics and that verified in performance metrics sections. We discuss our proposed approach in one by one manner.

a) Preprocessing

This section is the most important section for analyzing the sentiments where involve the removal of noisy data and obstacles that defeat the analysis of sentiments. Here we remove the repeated words on the tweets (For example: "Rose is a very very beautiful flower"), URL on the comments area, audio comments, video comments, special characteristics like brackets and curly braces. Thus preprocessing result in effective retrieval of character and words on the tweets and that helps in extraction of aspects.

b) Extraction of Aspects

Aspects are defined as the verb form on the English language which is related to time characteristics such as completion, duration, repetition of several actions. In our proposed system we involve extraction of aspects from tweets. We involve extraction of aspects from splitting of join words, tenses, and spell check. For lexicons, we involve the sentence that must have the sentiment words like happy, sad, ugly, awesome and etc. These words have the specialty for expressing the positive, negative and neural sentiments. Then for parts of speech, we involve a standard POS tagger for extraction of sentiments from the sentence. For spell checking, we involve checking of words with optimal spelling from word dictionaries such as "Awwweeeesomeeeeee", "Beauuuuuutifullll", "verrrrrrrrrrry bad", etc.

c) Sentiment Analysis

We involve the sentiment analysis based on the words form aspects which can express the reviews in terms of positive, negative, neutral and in terms of emotions. Here the emotions are analyzed based on smileys. The emotions are the symbolic descriptions of mind, moods, emotions and feelings which are used in many social networks and activities. The list of emotions are taken from a set of 145 emotions which is extended to 230 that has 120 of positive emotions and 110 of negative emotions, thus finally the sentiment score of an emotions [11] are computed based on,

$$\text{Pol-E (Score)} = \begin{cases} 1, if\ (r \in R\ \wedge e \in E(positive)) \\ 0, if\ (r \in R\ \wedge e \in E(negative)) \end{cases} \quad (1)$$

Where, the scoring of emotions are said to be 1 for positive and neutral whereas 0 for negative reviews. Here R is the set of reviews and r is a review. 'e' is the polarity value between 0 and 1. E (positive) and E (Negative) is defined as the total list of positive and negative emotions.
Modifiers play a major role for analyzing the sentiments which enhance the strength of the words in terms of positive and negative. We analyze the strength of reviews based on positive, negative and neutral based on modifiers. The modifiers are analyzed by the polarity scoring that is computed as follows:

$$\text{Pol\_score (M)} = \begin{cases} P\_sc\ (W) + (P_{sc}(w) * pm_{sc(W_x)})), \\ if\ (r \epsilon R\ \wedge W_x \epsilon M(positive)) \\ P\_sc\ (W) + (P_{sc}(w) * nm_{sc(W_y)})), \\ if\ \left(r \epsilon R\ \wedge W_y \epsilon M(negative)\right) \end{cases} \quad (2)$$

Here Wx and Wy belongs eo the set of positive and negative modifiers, W is the opinion word which belongs to a set of words W, r is a review from a set of reviews R, pm_score(Wx) is the percentage score of positive modifiers and nm_score(Wy) is the percentage score of negative modifiers.

Using the SentiWordNet, we analyze the POS tagger and perform the analysis of speech tag. Here we compute the sentiment scoring for the words based on,

$$P_{Sc(SW)} = \begin{cases} P\_sc\ (WP), \\ if\ max\left(p_{Sc}(WP), P_{sc}(WN), P_{SC}(WNU)\right) = P_{SC}(WP) \\ P_{sc}(WN), \\ if\ max\left(p_{Sc}(WP), P_{sc}(WN), P_{SC}(WNU)\right) = P_{SC}(WN) \\ P_{sc}(WNU),\ \ else \end{cases} \quad (3)$$

Based on these formulae we compute the scoring of the sentiments. The following algorithm specifies the proposed SAA algorithm.

Algorithm 1: Sentiment Analysis Algorithm

Input : Aspects
Output : Scoring for positive, negative, neutral words

Start

1. *Give Aspects*
2. *For (sentence)*
   {
3. *Calculate the score for emotions using equation (1)*
4. *Calculate modifiers score from the equation (2)*
5. *Calculate SentiWordNet from equation (3)*
   }
6. *End for*

End

---

d) Hybrid classifier

Our proposed work mainly focus for classifying the sentiments from the specific data. Here based on the words score, we classify the data for getting effective result. Our approach focus on combination of modified Genetic algorithm and Particle Swarm optimization algorithms. Here genetic algorithm works with a population of n chromosomes that specify the candidate solution and it has the process of initialization, fitness evaluation, newer population, analysis. The work flow of genetic algorithm was specified on figure 3.

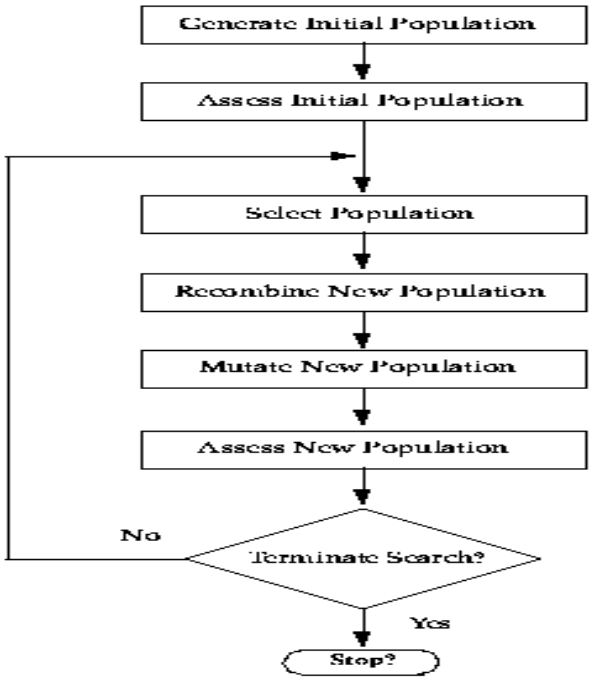

Fig 3. Genetic Algorithm

Then PSO [12] has specific learning factors which several populations and with inertia weights. Here this involves the total number of swarm evolutions. The work flow of PSO was described in figure 4.

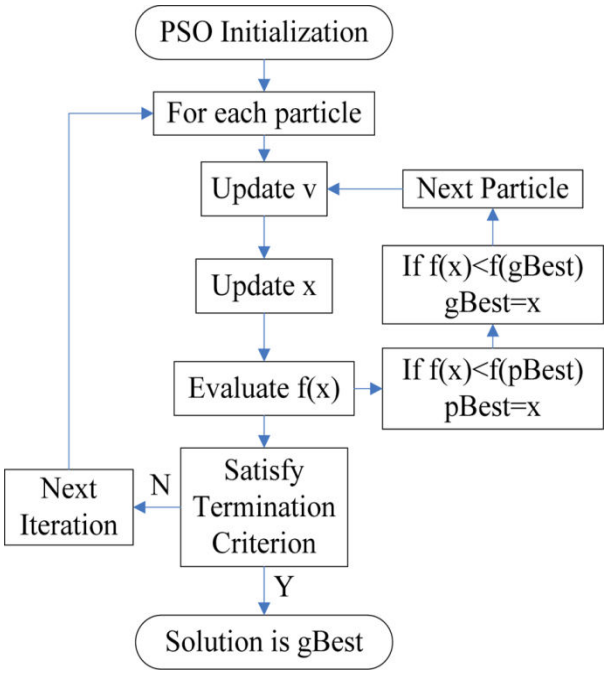

Fig 4. Particle Swarm Optimization

These both algorithms has drawbacks such that PSO has the drawback that it specifies the result based local data whereas the genetic algorithm does not provide results because the fitness function is based on random chosen of data. To solve these problems, we combine these two algorithms for effective classification and provide accurate result. We get the features based on the scores obtained on the SAA algorithm. Figure 5 describes the MG-PSO algorithm and algorithm 2 describes the overall process of MG-PSO.

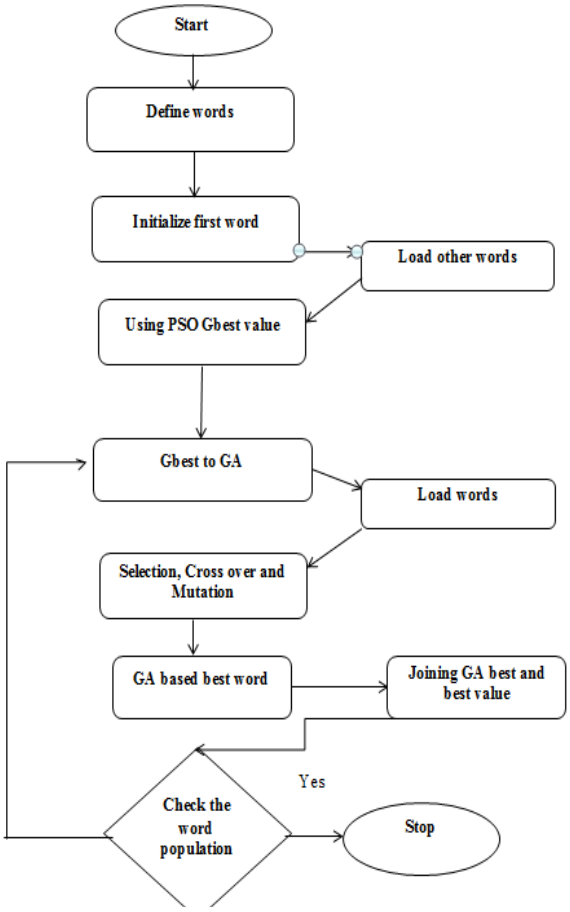

Fig 5. Proposed Hybrid MG-PSO

Algorithm 2: Hybrid MG-PSO classification

Input : Comments $N_i = \{N_1, N_2, N_3 \ldots\}$

Output : best results

Start

1. Read parameters of N_i (reviews)
2. Set initial word
3. For N_i (i= 1, …)
   {
   Calculate FF
   Best FF ← N_i
   Calculate Gbest← word (PSO)
   }
   End For

4. Gbest → GA
5. For N_i (i= 1, …)
   {
   6. Selection N_i
   7. Crossover N_i
   8. Mutation N_i
   9. Calculate Gbest1← word
   10. BESTPOST= Gbest*Gbest1
   11. if (BESTPOST >= scoring)
       {
        Goto step 6
          Else
        BESTPOST (Best result)
       }
      End if
End For

In above algorithm, we use N_I is the number of words on the sentences, here we set the first word of the sentences and calculate the fitness function for the PSO algorithm based on standard formulae,

Velocity:

$$vi\ (t + 1) = \alpha\ vi + c1 \times rand \times (pbest\ (t)\text{-}xi\ (t)) + c2 \times rand \times (gbest\ (t)\text{-}xi\ (t)) \quad (4)$$

Position update:

$$Xi\ (t + 1) = Xi\ (t) + Vi\ (t + 1) \quad (5)$$

Then this value is assigned to genetic algorithm, which processes the selection, mutation and crossover operation on the data. Finally we evaluate the best position by combining GA scoring and PSO Scoring. Finally we classify the sentiments. Thus our proposed work provides effective results which are evaluated on the comparative metrics.

V. COMPARATIVE ANALYSIS

a) Dataset

Our experimental procedure involves, twitter dataset which is taken from Twitter social media that contains topics like saints, funny, images, jokes, sports and college students. This dataset contains 1000 tweets with positive and negative tweets. Here the positive tweets are represented by 1 and negative tweets are represented by 0. This dataset can be applied various machine learning approaches which results on performance metrics such as accuracy, computation time and fitness function value.

b) Software Requirement

Our proposed framework is implemented in Java programming language. We used the 1.8 version of Java Development Kit (JDK) and Integrated Development Environment (IDE) we preferred is Netbeans of version 8.0. Java is a programming language expressly designed for use in the distributed environment of the Internet. It was designed to have the "look and feel" of the C++ language, but it is simpler to use than C++ and enforces an object-oriented programming model. Java can be used to create complete applications that may run on a single computer or be distributed among servers and clients in a network. It can also be used to build a small application module or applet for use as part of a Web page. Applets make it possible for a Web page user to interact with the page. We implement our project in windows 7 32-bit operating system. Table 1 specifies the hardware and software requirements.

TABLE I: REQUIREMENTS

OS	Windows 7 (Ultimate 32-bit)
IDE	Netbeans 8.0
Development Kit	Jdk 1.8
Processor	Dual Core
RAM	1 GB

c) Existing Methods Analysis

Our proposed framework is compared with several existing system. Table 2 describes the comparison table of proposed work with existing system, which is explained as follows:

- CSK [7]:

This method performs the sentiment analysis based on hybrid cluster method that analyzes the sentiments of tweets based on k-means clustering algorithm and modified cuckoo search algorithm. This approach performs effective in terms of fitness function and computation time.

- Dictionary Approach [10]:

This method involves the classification approach based on the dictionary based algorithm which support lexicon based analysis on the contexts and sentiment polarity levels.

TABLE II: COMPARATIVE RESULTS

Method	Accuracy	Error Rate	Recall
SVM (10)	0.960	0.04	0.959
MG_PSO	0.968	0.03	0.977
Dictionary based Classification (10)	0.810	0.19	0.835

d) Result Metrics

Our proposed system is evaluated with various performance metrics that are compared with several existing system. Some of the performance metrics are described as follows:

Accuracy

The accuracy is calculated based on overall performance of our proposed hybrid classification process. This can be computed by,

$$\text{Accuracy} = \frac{TN+TP}{TN+TP+FP+FN}$$

Recall

It is defined as the measure of the proportion of positives rates whereas the data are identified correctly, which can be expressed in formulation as follows:

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

Error Rate

The error rate is calculated by based on participation of input data which has average errors,

$$\text{Error rate} = E - \frac{1}{n}\sum_i E_i$$

Computation Time

The computation time is defined as the total number of time taken by the process for completing its task.

ISSN (ONLINE): 2456-5717

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Vol.3, Issue.9, September 2017

VI. GRAPHICAL RESULTS

This section describes graphical result of proposed which are compared with existing system. This specifies effective results of our proposed system when comparing several other machine learning approaches such as PSO, CSK, etc.

a) Accuracy

Figure 6 describes the graphical result of proposed system by comparing several existing techniques like CS, ICS, PSO, SVM, etc. Here the graph comprises the effectively higher result when compared with all other approaches. Table 3 describes the values of accuracy.

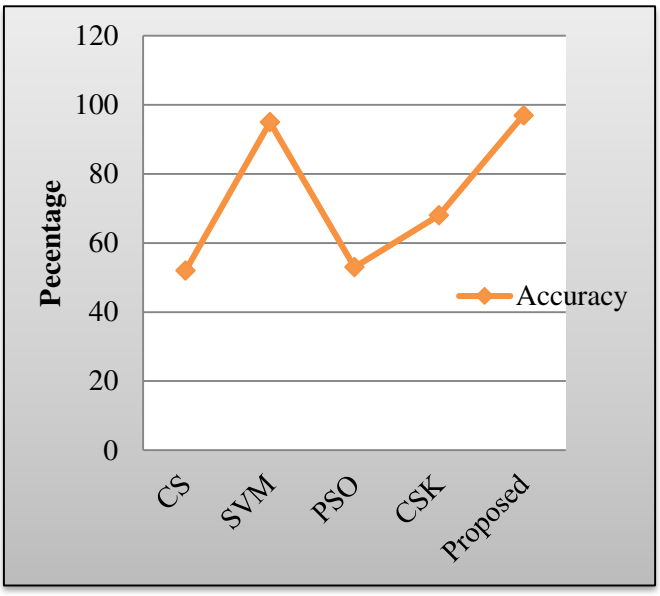

Fig 6. Comparison of Accuracy with existing system [7]

TABLE III: COMPARISON OF ACCURACY

Techniques	Accuracy
CS	50
SVM	95
PSO	55
CSK	65
Proposed (MG-PSO)	97

b) Computation Time

The computation time is described in graphical analysis in figure 7. Our proposed analysis has less timing when compared with other existing system. Table 4 describes the comparison of computation time.

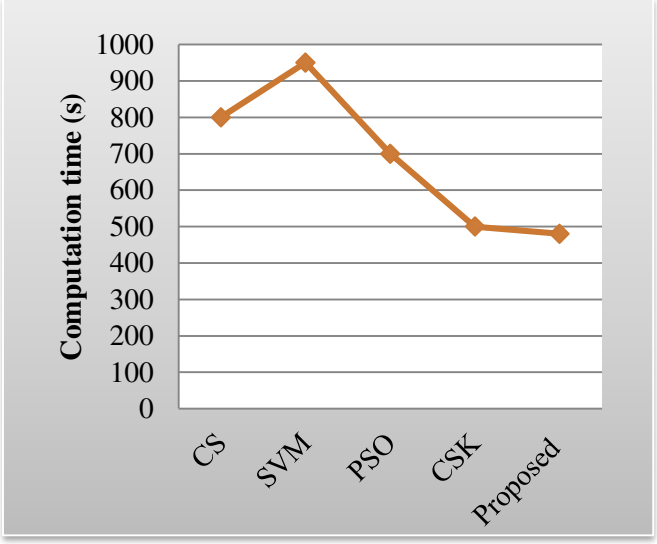

Fig 7. Comparison of computation time with existing system [7]

TABLE IV: COMPARISON OF COMPUTATION TIME

Techniques	Computational Time
CS	800
SVM	950
PSO	700
CSK	500
Proposed (MG-PSO)	480

c) Fitness Function Value

The Fitness function value is described in graphical analysis in figure 8. Our proposed analysis has lesser value when compared with other existing system. Table 5 describes the comparison of fitness function value.

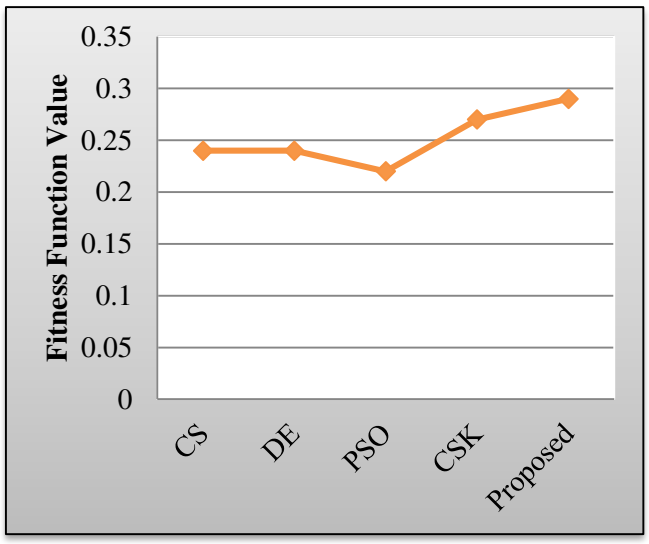

Fig 8. Comparison of fitness function value with existing system [7]

TABLE V: COMPARISON OF FITNESS FUNCTION VALUE

Techniques	Fitness Function value
CS	0.24
DE	0.24
PSO	0.22
CSK	0.27
Proposed	0.29

VII. CONCLUSION

Due to rising of internet application and several social media, opinion mining plays a major role on extracting the opinions from various users. Though it has lot of challenges such as several languages but achieves good results based on several data mining approaches and classification approaches. The categorization of positive, negative comments in a text is written in natural languages. Our proposed framework involves a hybrid approach for classifying the sentiments from various data. We implemented our idea on java language and used twitter dataset from the twitter social media which provide efficient results in terms of accuracy and computation time when compared with several existing techniques such as PSO, CSK, SVM, etc.

REFERENCES

[1] Kumar Ravi, Vadlamani Ravi, "A survey on opinion mining and sentiment analysis: tasks, approaches and applications", Knowledge-Based Systems, 2015.

[2] G.Vinodhini, RM.Chandrasekaran, "Sentiment Analysis and Opinion Mining: A Survey", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 6, pp. 1-11, 2012.

[3] Brindha G. R., Santhi.B, "Application of Opinion Mining Technique in Talent Management", International Conference on Management Issues in Emerging Economies (ICMIEE), IEEE, pp.127-132, 2012.

[4] P. Kalaivani and K. L. Shunmuganathan, "Feature Reduction Based on Genetic Algorithm and Hybrid Model for Opinion Mining", Hindawi Publishing Corporation Scientific Programming, pp.1-15 pages, 2015

[5] Bo Wang, Min Liu, "Deep Learning for Aspect-Based Sentiment Analysis", pp. 1-9 , 2016.

[6] Kazi Mostafizur Rahman and Aditya Khamparia, "Techniques, Applications and Challenges of Opinion Mining", I J C T A, volume 9(41), pp. 455-461, 2016.

[7] Avinash Chandra Pandey, Dharmveer Singh Rajpoot, Mukesh Saraswat, " Twitter sentiment analysis using hybrid cuckoo search method", Information Processing and Management, Elsevier, vol 53, pp. 764-779, 2017.

[8] Chlo´e Clavel and Zoraida Callejas, "Sentiment analysis: from opinion mining to human-agent interaction", IEEE Transactions on Affective Computing, pp. 1-22 , 2015.

[9] Ngoc Phuong Chau, Viet Anh Phan, Minh Le Nguyen, "Deep Learning and Sub-Tree Mining for Document Level Sentiment Classification", International Conference on Knowledge and Systems Engineering (KSE), pp.1-6, 2016.

[10] Santanu Mandal, Sumit Gupta, "A Novel Dictionary-Based Classification Algorithmfor Opinion Mining", IEEE Computer society, pp.175-180, 2016.

[11] Muhammad Zubair Asghar, Aurangzeb Khan, Shakeel Ahmad, Maria Qasim, Imran Ali Khan, "Lexicon-enhanced sentiment analysis framework using rule-based classification scheme", RESEARCH ARTICLE, pp. 1-22, 2017.

[12] Ming Li,Wenqiang Du, and Fuzhong Nian, "An Adaptive Particle Swarm Optimization Algorithm Based on Directed Weighted Complex Network", Hindawi Publishing Corporation, Mathematical Problems in Engineering, pp. 1-6, 2014.