International Journal of Engineering & Technology, *7 (4) (2018) 5887-5894*

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i4.20393
Research paper

Towards improving performance of tweet sentiment analysis by optimized feature weight using metaheuristic discriminative classifier approach

Stuti Mehla [1] *, Sanjeev Rana [2]

[1] *Research Scholar of Computer Science & Engineering, Maharishi Markandeshwar University, Mullana, Ambala. India*
[2] *Professor, Department of Computer Science &Engineering, Maharishi Markandeshwar University, Mullana, Ambala, India*
**Corresponding author E-mail: Stuti21mehla@gmail.com*

Abstract

Technology is exponentially rising day by day and one of the most booming product of this gradually updated technology is social media. Social media has become that platform where user can share his experience or views and can communicate to a mass in a single instance. Expanding technology has allowed the user to post its views anytime and from anywhere. These posted views can be made useful by the companies for their product reviews and this is the reason new fields like Sentiment analysis, Text mining have come into existence. Challenge in text mining is extraction of weight given to features because features weights is highly sparse in nature which increase the false positive rate and reducing the accuracy. In this paper proposed Optimized Feature Sentiment Classifier (OFSC) System gives the optimized weight by convergence of error or minimizing error. After that we improve the learning through classifier. Proposed approach focus on hybridization of optimization techniques ACO, PSO and BAT with Naïve Bayes Classifier to enhance the parametric values of performance matrix. In our work we have taken Twitter data as a sample dataset for optimized feature classification.

Keywords: *Sentiment analysis; feature optimization; ACO; PSO; Naïve Bayes; BAT.*

1. Introduction

Technology has put its feet in every field whether it is industrial sector, manufacturing, media, healthcare, communications or social media. Trillions of data is coming from every sector at an abrupt rate. So to handle this large volume of data having huge variety and coming at a single instance is very difficult and it is termed as Big Data. Interrogation include different fields like acquistion analysis, repository search, transfer rate of data, sharing, perception, renewing and querying the data with confidentiality[1].Basic purpose of Big Data is predictive analysis, user behaviour analysis or use different other progressive data analytic methods in which query data is derived from a sample of data set. Scrutiny of sample data can explore new interrelationships, helpful for realizing market trends, warfare crime, prohibit diseases. Scientists, business executives, doctors and government generally meet problems to handle large sample data-sets providing information about areas i.e metropolitan services, Internet search, and business trends [11]. As to spot business trends Social media is playing a major role. Social media is that media in which user share its ideas, views, career interests and information to virtual society. According to a survey more than 100,000,000 users are registered in social media and these users communicate to their friends, followers and communities in a single instance. Twitter is most popular social networking site in which user post views or communicate with messages known as tweets not more than 140 characters. User can tweet through any electronic medium ie mobile app or through website[2].As twitter is secure and easy to join and use, it is considered as one of the most visited social site in which 200 million are monthly active users post tweets and these tweets can create a huge amount of dynamic data content[2].People related to different fields and different regions use twitter, so that people can follow them and this results a huge amount of audience which can be targeted for business purpose[11]. Companies can directly communicate to the customers and know their product reviews and customer experience for the market research. So it is maintaining a producer consumer relationship and due to this relationship sentiment analysis and text mining have come into existence [9].Sentiment analysis is that process in which sentiments are identified from a text unit through some technique ie natural language processing, statistics or machine learning[3]. It can be used in different fields like politics, sociology, psychology, sports, brands, entertainment because tweets include mass opinion. For eg. In politics through sentiment analysis we can analyse the trends, can know views about party policies. It helps in predicting polarity because tweets include the public opinion. Similarly it can be used by companies to know the market trends and public poll for products. In simple terms, predicting whether the given sentiment or review is positive or negative at a huge rate is known as sentiment analysis. Since sentiment analysis is a classification problem. We have considered sentiments as polar i.e either positive or negative. Most of the researchers have focused on the traditional classifiers like maximum entropy, SVM, Naïve Bayes and there is a need to optimize these classifiers so that we can get maximum output and improve the performance of classifiers. In this research work we have developed an analytical framework termed as Optimized feature sentiment classifier(OFSC) in which we optimize the features from a sample of tweets through meta heuristic techniques like Particle Swarm Optimization(PSO), Ant Colony Optimization(ACO) and BAT by giving relative weights to features and then these extracted features are feed into classifier ie Naïve Bayes to enhance the learning. These three of optimization techniques have idea from nature. They indicate how swarms, ants and bats use their speed, distance and echos to get maximum output. Learning of classifier is evaluated through performance matrix. In this research work we have also focused on, to enhance the parameteric values of Performance matrix i.e accuracy, precision and recall.

Remaining paper explains the following sections: section 2 explains the related work section 3 describes the system model in which techniques are explained, we have followed, section 4 describes results and section 5 explains the conclusion and future work.

2. Related work

Many studies are made which are related to different analytics and classifiers worked on different types of microblogging data. Some of them are explained below:

In [3] authors introduced an approach which abstract the twitter data from different sources and convert them into positive and negative tweets with the help of classifiers like maximum entropy, SVM, Naive Bayes and lexicons. Authors described that how much these classified tweets are useful for those companies who consider these tweets for their product review.

In [4] gave a diagram of the present state of research in reasonable data modeling. Specifically, it examined its hypothetical and test perspectives in large-scale data-intensive fields, identifying with: (1) show vitality productivity, counting computational necessities in learning, and conceivable methodologies, and (2) data-intensive areas' structure and plan, counting the connection between data models and attributes, With the surge in e-science data, manageable data modeling has been appeared to offer a path forward because of its ease in dealing with large amounts of data. It is likewise conceived that such data-modeling revolution can be promptly reached out to different areas in e-science. These recently composed feasible data models won't just have the capacity to adapt to the rising large-scale data paradigm, additionally give a methods in augmenting its arrival for the different e-science areas. In [5] venture gathered constant tweets from U.S. soccer fans amid five 2014 FIFA World Cup diversions (three amusements between the U.S. group and another rival and two recreations between different groups) utilizing Twitter seek API. They utilized sentiment analysis to look at U.S. soccer fans' enthusiastic reactions in their tweets, especially, the enthusiastic changes after objectives (either possess or the opponent's). They found that amid the matches that the U.S. group played, fear and anger were the most widely recognized negative feelings and by and large, expanded when the rival group scored and diminished when the U.S. group scored. Suspicion and satisfaction were likewise for the most part reliable with the objective outcomes and the related conditions amid the amusements. Besides, we found that amid the matches between different groups, U.S. tweets indicated more satisfaction and expectation than negative feelings (e.g., anger and fear) and that the examples in light of objective or misfortune were misty. This venture uncovered that games fans utilize Twitter for passionate purposes and that the big data way to deal with break down games fans' sentiment demonstrated outcomes for the most part steady with the expectations of the aura hypothesis when the fanship was clear and demonstrated great prescient legitimacy. In [6] paper examines political homophily on Twitter. Utilizing a mix of machine learning and social network analysis it arrange clients as Democrats or as Republicans based on the political substance shared. They at that point examine political homophily both in the network of responded and non-responded ties. They find that structures of political homophily contrast unequivocally amongst Democrats and Republicans. By and large, Democrats show more elevated amounts of political homophily. In any case, Republicans who take after official Republican records show more elevated amounts of homophily than Democrats. Moreover, levels of homophily are higher in the network of responded devotees than in the non-responded network. It propose that examination on political homophily on the Internet should take the political culture and practices of clients truly. In [7] winnowed 2.5 million feeling tweets covering 7 feeling classes for programmed feeling recognizable proof. This is one of the biggest datasets for programmed feeling distinguishing proof. The trial comes about demonstrate that the component blend of unigrams, bigrams, existing sentiment and feeling dictionaries, and grammatical form accomplishes the best accuracy, despite the fact that lexicon based also, grammatical feature highlights turn out to be less compelling in recognizing fine-grained feelings than in sentiment analysis. In [8] examine the advancement of infrastructure and the improvement of abilities for data mining on "big data". One vital lesson is that fruitful big data mining in rehearse is about significantly more than what generally scholastics would consider data mining: life "in the trenches" is possessed by much preliminary work that goes before the utilization of data mining calculations and taken after by considerable push to transform preliminary models into vigorous arrangements. In this specific situation, they talk about two points: First, patterns play a critical part in helping data researchers comprehend petabyte-scale data stores, yet they're insufficient to give a generally speaking "big picture" of the data accessible to create insights. Second, they watch that a noteworthy test in building data examination stages comes from the heterogeneity of the different parts that must be integrated together into creation work processes—we allude to this as "plumbing". In [9] display the engineering behind Twitter's ongoing related question proposal and spelling rectification benefit. Despite the fact that these errands have gotten much consideration in the web look writing, the Twitter context introduces a constant "twist": after huge breaking news occasions, they expect to give pertinent outcomes within minutes. This paper gives a contextual investigation illustrating the difficulties of ongoing data processing in the time of "big data". They recount the narrative of how framework was constructed twice: its first execution was manufactured on a run of the mill Hadoop-based analytics stack, yet was later supplanted since it didn't meet the inactivity prerequisites fundamental to create meaningful constant outcomes. The second execution, which is the framework sent underway, is a custom in-memory processing engine particularly intended for the assignment. This experience showed that the current average use of Hadoop as a "big data" stage, while incredible for experimentation is not appropriate to low latency processing. In [10] LinkedIn's Hadoop-based analytics stack is presented, to extract insight and build product features from massive amount of data which allow data scientists and machine learning researchers. In particular, in providing a rich developer ecosystem they presented solution to the "last mile" issue. To online system this includes easy ingress from and egress, and as production processes managing the workflows. From the researcher a key characteristic of solution is that these distributed system concerns are completely abstracted away. For example, into the online system deploying data back is simply a 1-line Pig command that a data scientist can add to the end of their script. In [11] they have executed a social media data mining framework equipped for determining occasions identified with Latin American social agitation. This strategy straightforwardly removes few tweets from openly accessible data on twitter.com, consolidates comparative tweets into lucid estimates, and amasses a point by point and effectively interpretable review trail which permits end clients to rapidly gather data around an upcoming occasion. Its framework capacities by constantly applying multiple textual and geographic filters to a huge volume of data gushing from twitter.com through people in general API and a commercial data encourage. To be particular, scan the whole of twitter.com for a couple of painstakingly picked watchwords, seek inside those tweets for notices of future dates, channel again utilizing different strategic regression classifiers, and at long last allot an area to an occasion by geocoding retweeters. Moreover, they recognize socioeconomics likely intrigued in an upcoming occasion via scanning retweeter's current posts for statistic particular watchwords. In [12] paper tended to issues around the big data nature of Twitter analytics and the requirement for new data administration furthermore, inquiry dialect structures.

They looked into the tweet analytics space by investigating systems fundamentally for data gathering, data administration furthermore, dialects for questioning and breaking down tweets. They have distinguished the fundamental fixings required for a brought together structure that address the constraints of existing frameworks. The paper plots explore issues and recognize some of the difficulties related with the advancement of such incorporated stages.

3. System model

System model explains the phases of Optimized Feature Sentiment Classifier(OFSC) System. This system explains how the input tweet data is collected, preprocessed and how extracted features are optimized and then feed to classifier for the final prediction. System model include the phases through which input data passes through and how every phases of our optimized feature sentiment classifier system works. In OFSC system we are using relative weight of optimized feature set and then these feature set is trained by classifiers to enhance the values of performance matrix.

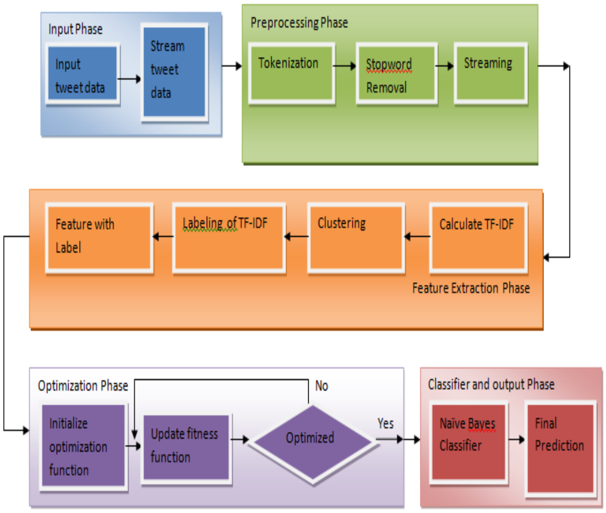

Fig. 1: Architecture of Optimized Feature Sentiment Classifier System.

Input Phase: In this phase it is explained how the twitter data is collected for the following experimentation. Data set is collected through REST APIs. It is explained below:
REST APIs
The programmatic access is provided by REST APIs for reading and writing Twitter data.

1) With the use of OAuth, REST API identifies users and Twitter applications;
2) JSON format is followed for response. Default entities and retweets:
3) By default returned the entities, applicable and retweets. As Tweet objects parts, returning entities unless setting false the include_entities. Timeline including native retweets unless setting false include_rts parameter.
4) Authentication on all end points:

All requests are authenticated by applications with application-only authentication or OAuth 1. 0a.
This Helps in understanding the API use of application category and the prevention of abusive behavior. This understanding is utilized for meeting the developers need in a better way and for evolving platform.
Preprocessing Phase:
Preprocessing phase includes the following steps tokenization, stopword and stemming.

1) In tokenization input tweet is converted into tokens. Tokens are the small chunks of data which are logically same and have some meaning.
2) In stopword process, we just remove stopwords like comma, full stop, hash(#) or any special sign which have no meaning and these are just added by users to emphasise its tweet.

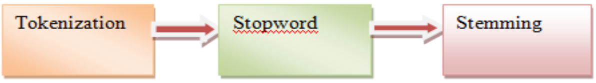

Fig. 2: Pre-processing Phase.

3) In Stemming process unuseful data i.e the data which logically don't have meaning are removed. It includes helping verbs, prepositions or any another grammatical form.

Feature Extraction Phase:
In this phase feature is extracted from the pre-processed data by calculating its TF-IDF and then clustering it into positive and negative tweets.

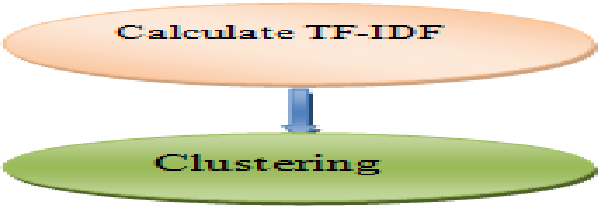

Fig. 3: Feature Extraction Phase.

1) First we calculate TF-IDF of every preprocessed data and keep them into dynamic arrays where TF stands for term frequency and IDF stands for Inverse Document frequency.TF-IDF tells us how much frequently a data item is used by customers and then this TF-IDF of every data is arranged into array having weights.
2) This weighted data is then clustered into positive and negative tweets.

Optimization Phase:
Optimization is a process in which best output is generated through the maximum use of resource. In optimized-classifier system we use different optimization algorithms like ACO,PSO, BAT. Every optimizer has different working, but basically each optimizer works on two phases updation and prediction for eg in NBACO algorithm if minimum error is obtained than optimization is done otherwise data is again feed into optimizer. The working of optimizer is explained below:

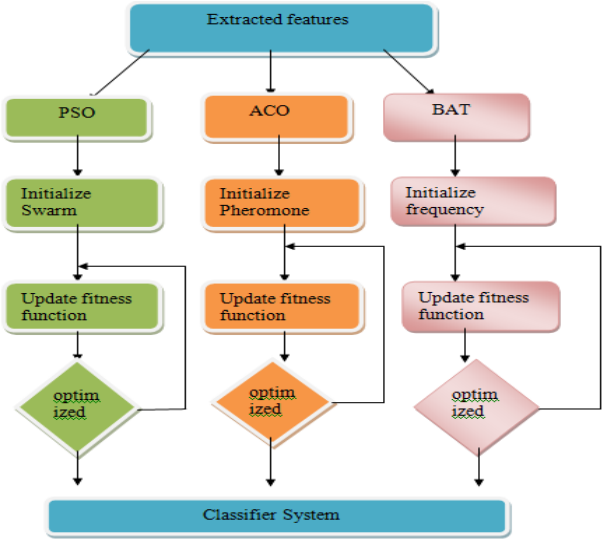

Fig. 4: Optimization Phase.

In this phase extracted features are optimized by using three of techniques and after optimization it is classified with the help of Naive Bayes Classifier. Basically these techniques are different on the basis of updation function.

Classifier Phase: In this phase we have used Naive Bayes Classifier which basically works on the bayes classification rule. It is used for training, classification and prediction. First this optimized data is trained, classified and then final prediction is done oin the basis of accuracy, prediction and recall features.

Working:

Our optimized-classifier system contain the different phases through which tweet data passes. First Input tweet data is preprocessed and then clustered. After categorizing data into positive and negative, it is feed into optimizers. The main purpose of optimizers is to extract maximum required features with the maximum use of resources. Then optimized data is feed into training classifiers for predicting the final results. Extracted features from input tweet data are optimized and then feed into classifiers.

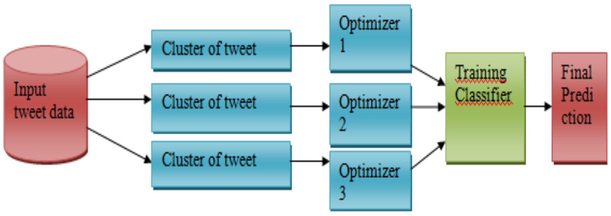

Fig. 5: Working of Optimized-Classifier System.

Prediction is done on the basis of three parameters i.e accuracy, precision and recall. Accuracy is that measure which tells the user how correct predictions are made by classifier. Precision is defined as that measure which tells us about relevant data from the extracted data and recall is that feature which tells the user most useful relevant data from the total amount of relevant data.

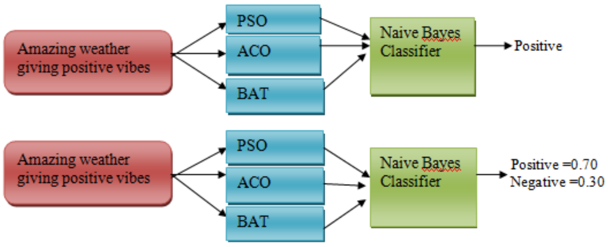

Fig. 6: Percentage of Tweet Data.

Suppose we have an example 'amazing weather' from the input tweet data. First we feed this extracted feature into optimizer and then give its weight. After giving weight we feed this optimized extracted weighted data into Naive bayes classifier and then predict the polarity of data. Optimized classifier system implement three algorithms Naive Bayes PSO,NB ACO and NB BAT. These algorithms are implemented on Eclipse tool. In this research paper NB ACO,NB BAT and Naïve Bayes PSO module are explained with their respective algorithm and flowchart. In NB ACO,NB BAT and Naïve Bayes PSO experiment is conducted having parameter values number of ants=100,number of iteration=500,evaporation coefficient=0.5,pheromone intensity=100,initial value of pheromone=100,alpha=(0.7,1),beta=(2,5) and the population size=100,loudness Ai= (0.4 to 0.9),pulse rate ri=(0.7,1),minimum frequency Qmin=10,maximum frequency Qmax =100 respectively.

Algorithm 1: NB ACO Module

```
Step 1:Introducing ants, where for individual $ant_n$
$\forall\, m \in M \; where \; M = (1,2 \ldots \ldots N)$
Step 2: In $ant_n$ individual variable $z_m^a$ where a
$\in A \; where \; A = (1,2 \ldots \ldots A)$
Step 3:From the pheromone table with probability in eq. (1), we select $\mu_r^a$
to integrate pheromone ,
where, r ∈ {1, 2, 3…K}.
Step 4: There is higher probability if there is minimum or less number of
errors.
Step 5: By using uniform distribution V (0,1) we used to generate a standard deviation $\sigma_r^a$ if ls ≤ $z_1$ from equation 2.
Where ls is a random value which lies between $z_1$ and the predefined
threshold 0 and 1.
Step 6: We use normal distribution N ($\mu_r^a$, $\sigma_r^a$) to generate new values for
variable $z_m^a$ if ls≤ $z_2$.
Step 7: uniform distribution generates random value and its solution for
$z_m^a$.
Step 8: The acquired variable $z_m^a$ denotes the examined value to certain
class label y.
Step 9: Evaluating probability for individual class
$P(z_m^a) = \frac{P(c_r)P(c_e)}{\sum_{i=1}^{c} P(y_i)P(y_j)}$, e=1,2……..y
Where,
$P(c_r)$is the prior probability of $c_r$,
$P(c_e)$is the conditional class probability density function.
sStep 10: Evaluate probability distribution over the set of features:
$P(z) = \prod_{r=1}^{h} P(y_r)P(z_m^a/y_r)$
Where
h is the number of classes,
$y_r$ is the r^th class.
Step 11: Calculate accuracy, precision and recall.
```

Fig. 7: Algorithm 1.

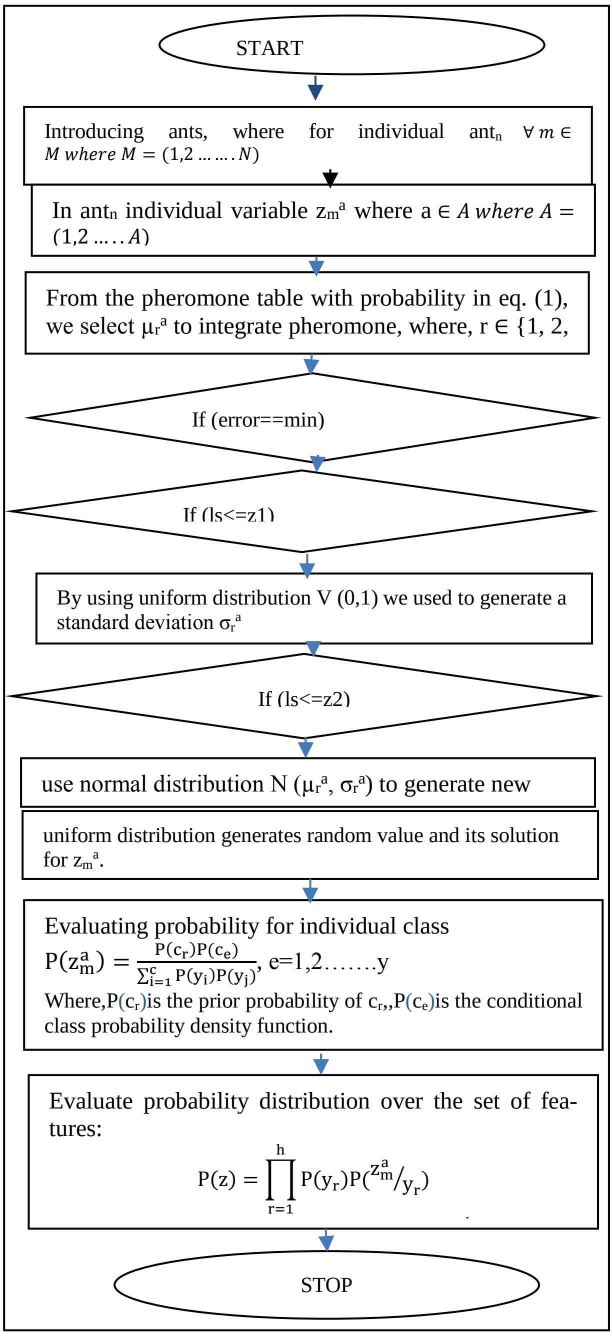
START
Introducing ants, where for individual antₙ ∀ m ∈ M where M = (1,2 …….N)
In antₙ individual variable zₘᵃ where a ∈ A where A = (1,2 …..A)
From the pheromone table with probability in eq. (1), we select μᵣᵃ to integrate pheromone, where, r ∈ {1, 2,
If (error==min)
If (ls<=z1)
By using uniform distribution V (0,1) we used to generate a standard deviation σᵣᵃ
If (ls<=z2)
use normal distribution N (μᵣᵃ, σᵣᵃ) to generate new
uniform distribution generates random value and its solution for zₘᵃ.
Evaluating probability for individual class
P(zₘᵃ) = P(cᵣ)P(cₑ) / Σᵢ₌₁ᶜ P(yᵢ)P(yⱼ), e=1,2…….y
Where,P(cᵣ)is the prior probability of cᵣ,,P(cₑ)is the conditional class probability density function.
Evaluate probability distribution over the set of features:
P(z) = ∏ᵣ₌₁ʰ P(yᵣ)P(zₘᵃ/yᵣ)
STOP

Fig. 8: Flow Chart 1.

Algorithm 2: NB_BAT Module

```
Step 1: Evaluating probability for individual category:
$P(z_m^a) = \frac{P(c_r)P(c_e)}{\sum_{r=1}^{y} P(c_r)P(c_e)}$, e∈ E where E = (1,2 …. y)where,
$P(c_r)$earlier probability of $c_r$,
$P(c_e)$ is the conditional category probability density work .
Step 2: evaluate probability appropriation over the arrangement of features

$P(z) = \prod_{r=1}^{h} P(y_r)P\left({z_m^a}/{y_r}\right)$

h is the no. of category,
$Y_r$ is the $r^{th}$ category.
Step 3: This step includes the Bat algorithm in which it start with the Following process.
Introducing: Arrange the generation counter g = 1; Introduce the crowd of MP bats b instantly where individual bat is related to the solution for such problems like, define loudness D; Set frequency B and the initial velocities s ; Set pulse rate i and weight factor W.
Step 4: For individual bat in P which is evaluated from f(x) there is an evaluation of quality w.
Step 5: when an elimination criterion isn't fulfilled or g< maximum generation then we has to sort the population of bats P in the order of quality w for individual bat from best to worst.
r = 1:MP (all Bats) do
Select uniform randomly $i_1$≠$i_2$≠i3≠ r
$1_4$ = [ MP * rand]
$s_r^g = s_r^{g-1} + ( s_r^g - z_* ) * B$
$z_r^g = z_r^{g-1} + s_r^g$
 if (rand >i) then
$z_v^g = z_* + \alpha \mathcal{E}^g$
Else
$z_v^g = z_{i1}^g + W (z_{i2}^g - z_{i3}^g )$
end if
Evaluate the fitness for the offspring $z_v^g$ , $z_r^g$ ,$z_{i4}^g$
Select the offspring $z_h^g$ with the best fitness among the off springs
$z_v^g$ , $z_r^g$ ,
if (rand < D) then
$z_{i4}^g = z_h^g$
end if
end for r
g = g+1;
Step 6: end while
Step7: Post-preparing the outcomes and representation.
Step 8: Calculate accuracy, precision and recall.
```

Fig. 9: Algorithm 2.

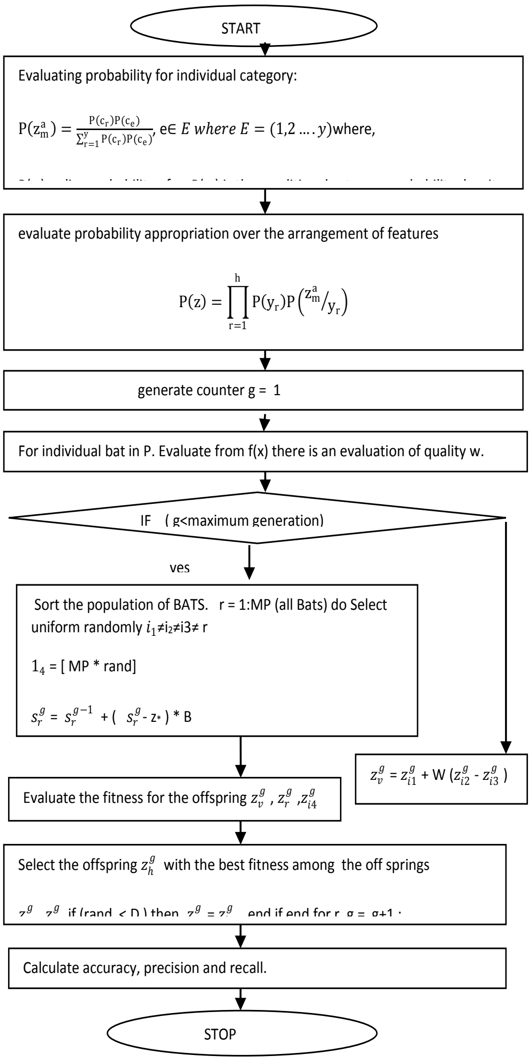
START
Evaluating probability for individual category:
$P(z_m^a) = \frac{P(c_r)P(c_e)}{\sum_{r=1}^{y} P(c_r)P(c_e)}$, e∈ E $where$ $E = (1,2 \ldots . y)$where,
evaluate probability appropriation over the arrangement of features
$P(z) = \prod_{r=1}^{h} P(y_r)P\left({z_m^a}/{y_r}\right)$
generate counter g = 1
For individual bat in P. Evaluate from f(x) there is an evaluation of quality w.
IF (g<maximum generation)
yes
Sort the population of BATS. r = 1:MP (all Bats) do Select uniform randomly i_1≠i₂≠i3≠ r
1_4 = [MP * rand]
$s_r^g = s_r^{g-1} + (s_r^g - z_*) * B$
$z_v^g = z_{i1}^g + W (z_{i2}^g - z_{i3}^g)$
Evaluate the fitness for the offspring z_v^g , z_r^g ,z_{i4}^g
Select the offspring z_h^g with the best fitness among the off springs
z^g z^g if (rand < D) then z^g = z^g end if end for r g = g+1 ;
Calculate accuracy, precision and recall.
STOP

Fig. 10: Flowchart 2.

Algorithm 3: Naive Bayes PSO Module

```
Step 1: Computing probability for each class: $P(x_n^d) = \frac{P(y_i)P(y_j)}{\sum_{i=1}^{c} P(y_i)P(y_j)}$,
j=1,2……..c
Where,
$P(y_i)$is the $y_i$prior probability,
$P(y_i)$is the conditional class probability density function.
Step 2: Calculate accuracy, precision and recall.

Step 3: In PSO model for each particle i in S do
Step4: for each dimension d in D do
Step5: //initialize each particle's position and velocity
Step6: $x_{i,d}$=Rnd($x_{max}$, $x_{min}$)
Step7: $v_{i,d}$=Rnd($-v_{max}$ /3,$v_{max}$/3)
Step8: end for
Step9: //initialize particle's best position and velocity
$v_i(k+1) = v_i(k) + \gamma 1_i (p_i - x_i(k)) + \gamma_{2i}(G - x_i(k))$
 New velocity
$x_i(k+1) = x_i(k) + v_i(k+1)$
 Where
i- particle index
 k- discrete time index
 $v_i$ –velocity of $i^{th}$ particle
 $x_i$ – position of $i^{th}$ particle
 $p_i$- best position found by $i^{th}$ particle (personal best)
 G- best position found by swarm (global best, best of personal bests)
 $G_{(1,2)i}$- random number on the interval[0,1]applied to the $i^{th}$ particle
Step10: $pb_i$=$x_i$
Step11: // update global best position
Step12: if f($pb_i$) <f(gb)
Step 13: gb = $pb_i$
Step14: end if
Step15: end for
```

Fig. 11: Algorithm 3.

Naïve Bayes Pso

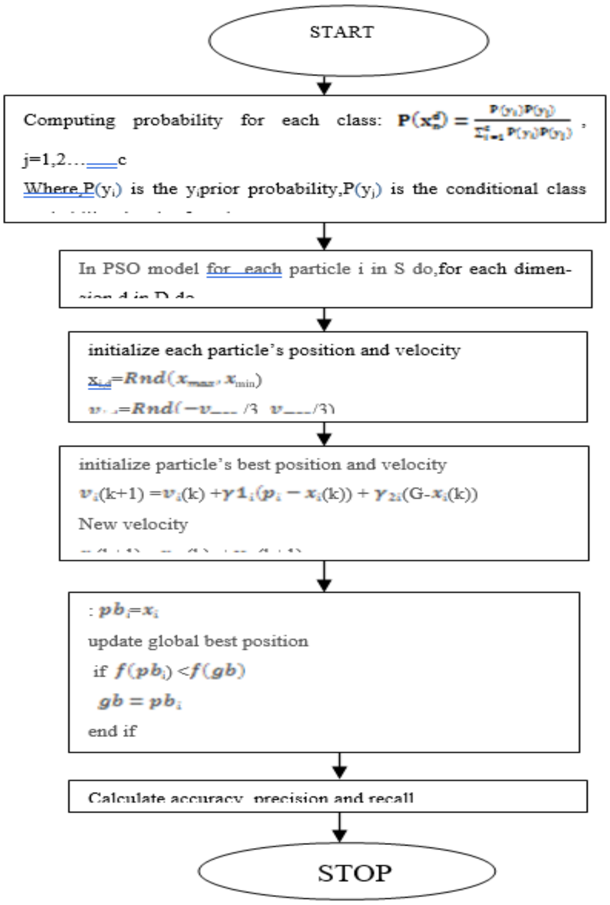
START
Computing probability for each class: $P(x_n^j) = \frac{P(y_i)P(y_j)}{\sum_{i=1}^{c} P(y_i)P(y_j)}$,
j=1,2…___c
Where P(y_i) is the y_i prior probability,P(y_j) is the conditional class
In PSO model for each particle i in S do,for each dimen-
initialize each particle's position and velocity
$x_{i,d}$=$Rnd(x_{max}, x_{min})$
initialize particle's best position and velocity
v_i(k+1) =v_i(k) +$\gamma 1_i(p_i - x_i$(k)) + γ_{2i}(G-x_i(k))
New velocity
: pb_i=x_i
update global best position
if $f(pb_i)$ <$f(gb)$
$gb = pb_i$
end if
Calculate accuracy, precision and recall
STOP

Fig. 12: Flowchart3.

4. Results

Tweet data is most significant data, which can be used for prediction, product analysis. This type of activities will be more beneficial if they are optimized and then classified. Optimization improves the efficiency of classifier. We have evaluated different algorithms on eclipse tool. Different optimizers like PSO, ACO, BAT are hybridized with Naive bayes classifier and then results are evaluated and compared with [3] in which classifiers are used for classification of tweet data. Graphs also show the comparison between algorithms. Classification is done on the basis of performance matrix features which tells the how accurate and precise are the extracted features and recall points to the learning capacity of our system. Table 1 shows the relative comparison between Naive Bayes PSO, NB ACO and NB BAT with the algorithms explained in [3]. Our work is more optimized because optimizer work on maximum use of resources results more accurate, precise output. The Fig 13 also tells us this relative comparison.

Table 1: Result of Naive Bayes PSO, NBACO, NB BAT, Random Forest, SVM, Naive Bayes, Logistic Regression and Ensemble Learning

	Naïve Bayes PSO	NB ACO	NB BAT	Random Forest	SVM	Naïve Bayes	Ensemble Learning	Logistic Regression
Acc(1k)	80.34	76.34	74.56	73.67	74.34	71.34	75.45	74.34
Pre(1k)	83.34	80.45	79.34	73.45	74.34	72.45	78.34	76.45
Pre(1k)	81.23	76.34	76.56	72.34	74.55	71.23	77.45	76.45
Acc (2k	79.98	74.34	73.45	75.34	72.34	71.23	72.45	76.45
Pre(2k)	78.45	71.34	74.34	73.23	74.56	71.34	75.45	76.56
Recall (2k	78.97	74.56	75.78	73.56	73.45	74.67	78.56	72.45
Acc(3k)	78.21	74.38	74.38	72.34	73.22	70.23	73.44	73.22
Pre(3k)	79.17	72.86	72.86	71.22	72.54	70.34	74.33	75.33
Pre(3k)	77.53	75.6	75.6	70.23	71.99	73.22	75.34	72.22

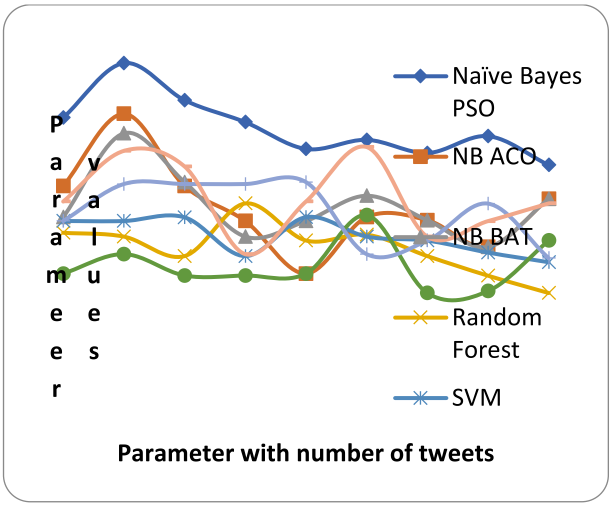

Fig. 13: Graph between Proposed Algorithms, Algorithms Explained in [3] With Number of Tweets and Parameter Values.

There are three features accuracy, precision and recall in performance matrix. Accuracy tells us how accurate classification is done by classifiers and the graph shows the relative comparison between algorithms. In the Fig 14, x-axis show the algorithms and y-axis give the parameteric values. Tweeter data which is evaluated is taken as 1K,2Kand 3K and it is observed that in case of Naive Bayes PSO (1MB data) there is 8.9%, 7.9% and 12% relative improvement as compare to Random Forest, SVM and Naive Bayes. Similarly in case of NBACO and NBBAT there is significant improvement 3.6%,1.2% respectively as we compare them with Random Forest algorithm for a sample of 1k data.

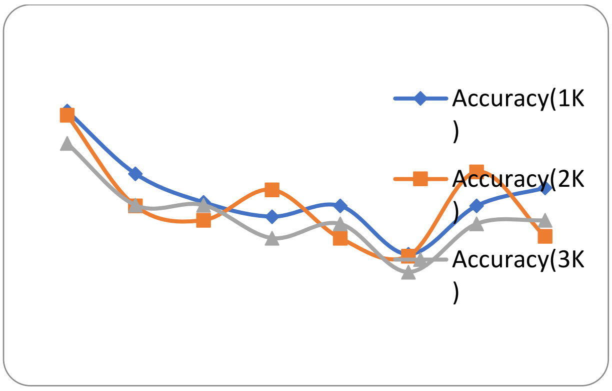

Fig. 14: Accuracy Measure between Proposed Algorithms, Algorithms Explained in [3] with Number of Tweets and Parameter Values.

Precision feature points out the relevant data from the extracted sample. Precision tells us how precise is our classifier to extract the required data from a given sample of data. In Fig 15 graph provide us the relative comparison between different algorithms on different set of data set where x-axis show the algorithms and y-axis give the parametric values. It is observed that Naive Bayes PSO, NBACO, NBBAT has significant improvement from previous algorithms explained in[3] for eg. there is 13%,9.53%,8.01% improvement respectively when Naive Bayes PSO, NBACO, NBBAT compared with Random Forest for a sample of 1k data.

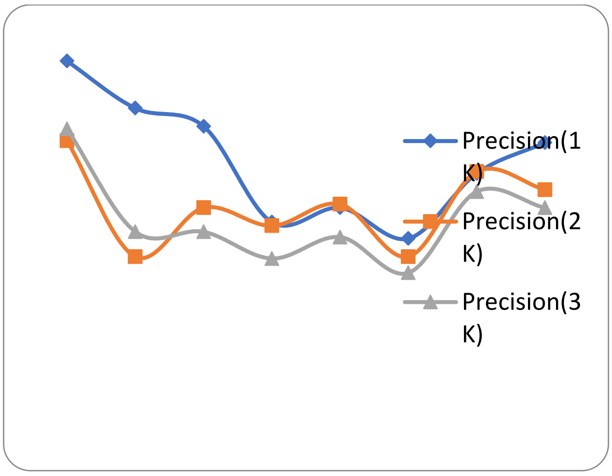

Fig. 15: Precision Measure between Proposed Algorithms, Algorithms Explained in [3] with Number of Tweets and Parameter Values.

Recall feature points out the relevant data which is used by user from the relevant data. This feature points out the learning capability of our algorithm. Fig 16 shows the relative comparison between algorithms having different size of data set where x-axis shows the algorithms and y-axis represent parametric value and in this case also there is remarkable improvement for eg. Naive Bayes PSO has 12.01/% improved recall features as compared to Random Forest algorithm for a sample of 1k data set.

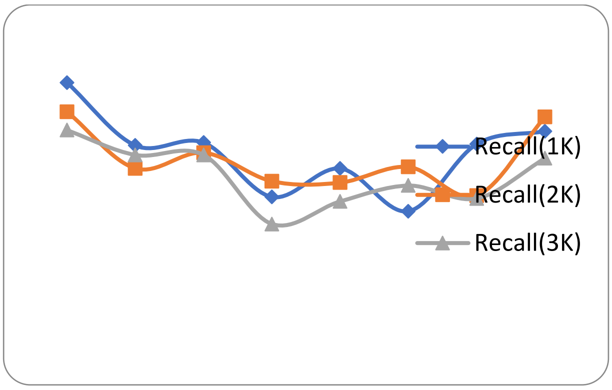

Fig. 16: Recall Measure between Proposed Algorithms, Algorithms Explained in [3] with Number of Tweets and Parameter Values.

5. Conclusion

In today's era technology is booming. Different social sites have come into existence in which user post their experiences. These experiences can be used for prediction and analysis by companies and political parties. Reducing the sparse value in feature vector from the feature set we developed an Optimized Feature Sentiment Classifier system (OFSC) which optimize the features and replace sparse value by optimized weight and then classify them. Proposed approach focused to enhance the performance matrix of our classifier system and it is achieved by hybridizing Naive Bayes Classifier with optimization techniques for enhancing the value of features weight. Evaluate the three of parameters of performance matrix. In future work we can enhance the classification to work on neutral data and further to do more refine categorization of positive and negative tweets

References

[1] https://en.wikipedia.org/wiki/Big_data.
[2] https://en.wikipedia.org/wiki/Twitter.

[3] Nadia F. F. da Silva, Eduardo R. Hruschka,Estevam R. et.al. “Tweet Sentiment Analysis with Classifier Ensembles” july16-2014. https://doi.org/10.1016/j.dss.2014.07.003.

[4] Al-Jarrah, Omar Y., et al. "Efficient machine learning for big data: A review." *Big Data Research* 2.3 (2015): 87-93. https://doi.org/10.1016/j.bdr.2015.04.001.

[5] Yu, Yang, and Xiao Wang. "World Cup 2014 in the Twitter World: A big data analysis of sentiments in US sports fans’ tweets." *Computers in Human Behavior* 48 (2015): 392-400. https://doi.org/10.1016/j.chb.2015.01.075.

[6] Colleoni, Elanor, Alessandro Rozza, and Adam Arvidsson. "Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data." *Journal of Communication* 64.2 (2014): 317-332. https://doi.org/10.1111/jcom.12084.

[7] Wang, Wenbo et al. "Harnessing twitter" big data" for automatic emotion identification." *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*. IEEE, 2012. https://doi.org/10.1109/SocialCom-PASSAT.2012.119.

[8] Lin, Jimmy, and DmitriyRyaboy. "Scaling big data mining infrastructure: the twitter experience." *ACM SIGKDD Explorations Newsletter* 14.2 (2013): 6-19. https://doi.org/10.1145/2481244.2481247.

[9] Mishne, Gilad, et al. "Fast data in the era of big data: Twitter's real-time related query suggestion architecture." *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*. ACM, 2013. https://doi.org/10.1145/2463676.2465290.

[10] Sumbaly, Roshan, Jay Kreps, and Sam Shah. "The big data ecosystem at linkedin." *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*. ACM, 2013. https://doi.org/10.1145/2463676.2463707.

[11] Compton, Ryan, et al. "Detecting future social unrest in unprocessed twitter data:“emerging phenomena and big data”." *Intelligence and Security Informatics (ISI), 2013 IEEE International Conference On*. IEEE, 2013. https://doi.org/10.1109/ISI.2013.6578786.

[12] Goonetilleke, Oshini, et al. "Twitter analytics: a big data management perspective." *ACM SIGKDD Explorations Newsletter* 16.1 (2014): 11-20. https://doi.org/10.1145/2674026.2674029.

[13] David Zimbra et.al. “Brand-Related Twitter Sentiment Analysis using Feature Engineering and the Dynamic Architecture for Artificial Neural Networks”1530-1605/16 $31.00 © 2016 IEEEDOI 10.1109/HICSS.2016.244.

[14] RabiNarayanBehera et.al. “Ensemble based Hybrid Machine LearningApproach for Sentiment Classification- AReview”International Journal of Computer Applications (0975–8887)Volume 146 –No.6, July 2016 https://doi.org/10.5120/ijca2016910813.

[15] Alexander Pak,Patrick Proubek “Twitter as a Corpus for Sentiment Analysis and Opinion Mining”.

[16] Andres Montoyo,Patricio Martienz“Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments” https://doi.org/10.1016/j.dss.2012.05.022.

[17] Yang Yu,Wenjing Duan et.al.“The impact of social and conventional media on firm equity value: A sentiment analysis approach” https://doi.org/10.1016/j.dss.2012.12.028.

[18] Walaa Medhat,Ahmed Hassan“Sentiment analysis algorithms and applications: A survey” https://doi.org/10.1016/j.asej.2014.04.011.

[19] Stuti Mehla, Saurabh Upadhyay “Performance Comparison of Statistical Techniques with Big Data Analysis” International Journal of Computer Applications (0975 – 8887) Volume 169 – No.6, July 2017 https://doi.org/10.5120/ijca2017914770.

[20] Mohammed Hossein, Barkhordaril,Mahdi Nimanesh1“ScaDiPaSi: An Effective Scalable and Distributable MapReduce - Based Method to Find Patient Similarityon Huge Healthcare Networks Copyright © 2015 Elsevier Inc. https://doi.org/10.1016/j.bdr.2015.02.004.

[21] Tao Huangb, Liang Lanc et.al.“Promises and Challenges of Big Data Computing in Health Sciences” Copyright © 2015 Elsevier Inc.

[22] Xiaolong Jina, Benjamin W.Waha et.al.“Significance and Challenges of Big Data Research”Copy right © 2015 Elsevier Inc.

[23] Raymond YK Laub,J Leon Zhaob et.al.“Demystifying Big Data Analytic for Business Intelligence” Copyright © 2015Elsevier Inc.

[24] Ibrahim Abaker Targio Hashema, Ibrar Yaqooba et.al.“The rise of “big data” on cloud computing: Review and open research issues”Copyright © 2014Elsevier Ltd.

[25] Wullianallur Raghupathi and Viju Raghupathi “Big data analytics in healthcare: promise and potential” Raghupathi and Raghupathi Health Information Science and Systems 2014, 2:3 http://www.hissjournal.com/content/2/1/3. https://doi.org/10.1186/2047-2501-2-3.