Sentiment Classification of Tweets, Supervised Learning Approach by Effective Feature Selection

Jatinder Singh[1], Dr. Kanwalvir Singh Dhindsa[2]

[1]*Research Scholar,*

[2]*Professor (CSE),*

[1-2]*Baba Banda Singh Bahadur Engineering College, Fatehgarh Sahib, I.K.G. Punjab Technical University, Jalandhar, PUNJAB.*

Abstract: Sentiment analysis of Tweet has been very useful and valuable technique in the field of sentiment analysis domain. This technique now the days used in the field of machine learning and data mining and is growing in popularity as a means of determining public opinion. In analysis, comparison with classifier and hybrid classifier has been done, for that SVM and Naive Bayes classifier is used, which is hybrid with PSO and ACO for effective feature weight. In fig 4.3 comparison of all experiment by one graph has been done in which it shows that SVM_ACO and SVM_PSO better perform than SVM. NB_ACO and NB_PSO perform better than NB but if comparision has is done between hybrid approaches then SVM_PSO show 81.80% accuracy, 85% precision and 80% recall. In case of naïve Bayes NB_PSO 76.93% accuracy, 76.24 precision and 82.55% recall, so experiments conclude that Naive Bayes improve recall and SVM improve precision and accuracy when use as hybrid approach.

Keywords: Optimization, Sentiment, Tweets, PSO, ACO, SVM.

I. INTRODUCTION

The most widely recognized definition depicts qualities of enormous information as volume, velocity and variety which represents the size of the data sets, rate of generation of data and different data respectively. This rate of generation acts upon filtered, reduced, transferred and used to analyzed to oppose the store for future processing [7]. Variety contains different form of data in big data which includes organized, unorganized and semi organized data.

Twitter is a famous micro-blogging plat form where users post there day to day activities related to their personal life and social life. Twitter contains messages in the limited number of words which show the users opinion on that time. By using these messages we analyze the users' behavior and opinion regarding the issues, products and their day to day situations.

II. LITERATURE REVIEW

Barrachina et al. [1] author proposed a proof of concept end to end solution that uses hadoop model. This examination is done on the VMware dataset. By using this method technical support requests are analysed .In improves the call resolution and increase the daily closure rate.

Min et al. [2] surveyed on the Big Data techniques and their background. It also includes the related technologies on cloud and Hadoop. Author discuss the four parts of Big data chain that data generation, acquisition, storage and analysis. It also focus on the challenges faced by the technologies. Data used for analysis is social media data and smart grid data.

Ibrahim et al. [3] reviewed by the Author that the latest technology that is cloud computing and its uses. It explains the current demand of the data used in the cloud that is Big Data. Cloud provides us the various services related to data scalability and availability. This Paper also explains the characteristic of Big data. This review show that cloud decrease the cost of hardware, space and software.

Partalas et al. [4] proposed web based classification in the field of Big Data. In this author mainly focus on the searching and data mining. In this paper author also discuss the fields in which future research is possible and the challenges in that field.

Jonathan [5] The term Big data is become really very popular in these days, author discussed this in this paper. This paper mainly focused on the concept of Big data because there is no single definition can define the meaning of this word. As the name it reflects the characteristics it contains a huge amount of data in it. This paper gives the detail on the big data technologies and their methods.

Lee et al. [6] explained the challenges and the advantages of the mobile communication on the Hadoop. It is also discussed by the author that a framework that collects the useful information from the internal and external sources. This framework performed the real time analytics on mobile devices and check the feasibility with the Hadoop platform.

Lu Guofan, et.al. [7] explained the call tracing system in the distributed network. Author explains the process of running job on Hadoop and call trace method. This system contains the analog data

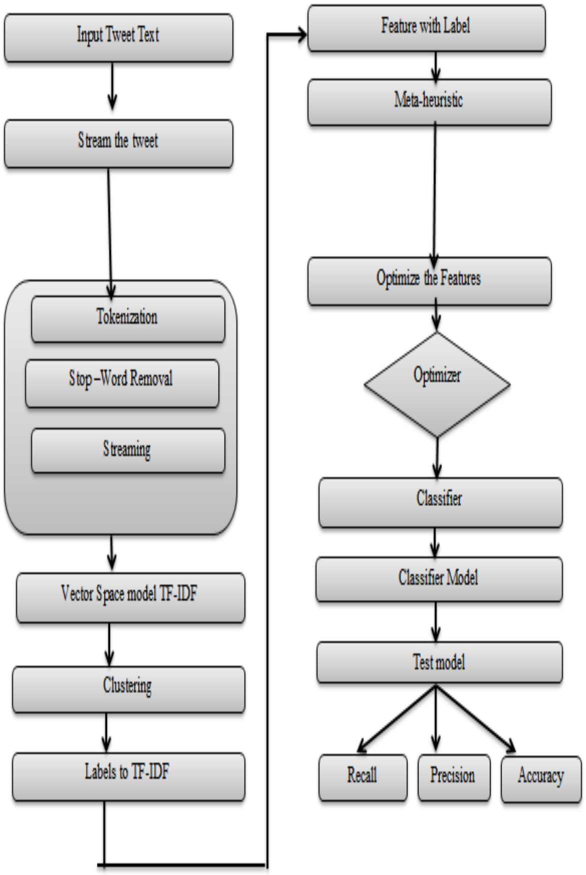
Source module and graphical user interface. Hadoop uses Map reduce method and oracle data base in it.

Chen et al [8] Min et al. [8] Illustrated the framework and State of the Art of big data. Introduced firstly, the general framework of big data and then experimented related technologies, such as could computing, Internet of Things, data centers, and Hadoop. The main focus was on the fourth phase of the value chain of big data, i.e., data acquisition, data generation, data data analysis and data storage.

Hao et al [9]: In this paper three novel time-based visual analyses technique is introduced to investigate data from social sites. The given techniques are 1) topic based analyses2) stream analyses3) pixel based analyses. Topic based analyses helps to extract the data from maps, measure customer opinion whereas stream analyses helps to identify interesting tweet on the basis of their density or polarization. Last but not least pixel-cell based analyses helps to analyse the sentiments having high density geo maps that visualize high amount of data in a single view. These three techniques are applied in several types of twitter data to show their sentiments.

Bac, et al. [10]: In this paper there is an introduction of new feature set on the basis of bigrams, data gain for the sentiment analysis from social sites. On the basis of Naïve Bayes and SVM this paper proposes a model for sentiment analysis. The main goal of this paper is to increase the analyses in effective manner. The result shows that this introduced model gives accuracy while analyses and highly effective for analyses.

III. PROPOSED WORK



Step1: Input the tweet text by continues streaming of tweets.
Step2: Pre-processing the text by tokenization streamming and stop-word removal.
Step3: Make vector space model with help of TF-IDF (Inverse document frequency).
Step4: Clustering the document according to its TF-IDF and make a label with the help of PCA (Principle component analysis).
Step5: Optimize the feature of TF-IDF with the help of meta-heuristic like PSO and ACO.
Step6: Hybrid the meta-heuristic with classifier and make the classifier model.
Step7: Check the performance of classifier model by precision, recall and accuracy.

Algorithm Used

Algorithm 1: SVM _PSO Module

Step 1: With the help of optimization model there is a description on classification of SVM model$\mathbf{min}_{\omega,\xi,Q}\mathbf{A}(\omega,\xi)$

$$\mathbf{min}_{\omega,\xi,Q}\mathbf{A}(\omega,\xi_r)=\frac{1}{2}\omega^t\omega+\frac{1}{2}\gamma\sum_{r=1}^{n}\xi_r{}^2$$

$$\mathbf{v}_l[\omega^t\phi(\mathbf{u}_r)+\mathbf{Q}=\mathbf{1}-\xi_r,\mathbf{l}=\mathbf{1},\mathbf{2},\ldots\ldots,\mathbf{n}$$

$$\xi=(\xi_1,\xi_2,\ldots\ldots.\xi_n)$$

Where
ξ_r←Slack variable
Q←Offset
ω←Support vector
γ← For balancing the model complexity and fitness error, classification of parameters.
Step2: With optimization model there is a description on the classification of SVM model. $\mathbf{min}_{\omega,\xi,Q}\mathbf{A}(\omega,\xi)$

$$\mathbf{min}_{\omega,\xi,B}\mathbf{A}(\omega,\xi_i)=\frac{1}{2}\omega^t\omega+\frac{1}{2}\gamma\sum_{r=1}^{n}\xi_r{}^2$$

$$\mathbf{s}_r[\omega^t\phi(\mathbf{u}_r)+\mathbf{Q}=\mathbf{1}-\xi_r,\mathbf{r}=\mathbf{1},\mathbf{2},\ldots\ldots,\mathbf{n}$$

Step 3:Then, describing the classification decision function:

$$\mathbf{F}(\mathbf{z}_r)=\mathbf{sgn}(\sum_{r=1}^{n}\alpha_r\mathbf{s}_r\mathbf{A}(\mathbf{u},\mathbf{u}_r)+\mathbf{Q})$$

```
Step 4: Calculate accuracy, precision and recall.
Step 5: In PSO model for  each particle r in S do
Step6 :     for each dimension a in D do
Step7:      //initialize each particle's position and velocity
Step8:      $z_{r,d}=\mathbf{Rnd}(\mathbf{z}_{max},\mathbf{z}_{min})$
Step9:      $\mathbf{s}_{r,d}=\mathbf{Rnd}(-\mathbf{s}_{max}/3,\mathbf{s}_{max}/3)$
Step10: end for
Step11: //initialize particle's best position and velocity
            $\mathbf{s}_r(h+1)=\mathbf{s}_r(h)+\gamma\mathbf{1}_r(\mathbf{f}_r-\mathbf{z}_r(h))+\gamma_{2r}(X-\mathbf{z}_r(h))$
        New velocity
            $\mathbf{z}_i(h+1)=\mathbf{z}_r(h)+\mathbf{s}_r(h+1)$
```

Where
r- particle index
h- discrete time index
s_r –velocity of rth particle
z_r – position of rth particle
f_r- great position found by rth particle (best, personal)
X- great position found by swarm (global best, best of

```
bests, personal)
 $X_{(1,2)r}$- random number on the interval[0,1] which is executed to the r^th particle
Step12: $\mathbf{ft_r}$=$\mathbf{z_r}$
Step13: // update global best position
Step14: if $\mathbf{f(ft_r)} < \mathbf{f(jt)}$
Step 15:    $\mathbf{jt = ft_r}$
Step16: end if
Step17: end for
```

```
Algorithm 2: NB_PSO Module
Step 1: Computing probability for each class: $\mathbf{P(z_m^a) = \frac{P(e_r)P(e_o)}{\sum_{r=1}^{r} P(e_r)P(e_o)}}$ , o=1,2……..v
Where,
P($e_r$) being the $e_r$ probability, prior.
P ($e_o$) being the conditional class probability density function.
Step 2: Calculate probability distribution over the set of features: $\mathbf{P(x) = \prod_{r=1}^{h} P(y_r)P({}^{z_m^a}/_{v_r})}$
Where
k is the no. of classes,
$v_r$ being the r^th class.
Step 3: Calculate Accuracy, precision and recall.
Step 4: In PSO model for each particle r in S do
Step 5:    for each dimension a in D do
Step6:    //initialize each particle's position and velocity
Step7:      $z_{r,d} = \mathbf{Rnd(z_{max}, z}_{min})$
Step8:    $\mathbf{s_{r,d}}=\mathbf{Rnd(-s_{max}}$ /3 ,$\mathbf{s_{max}}$/3)
Step9: end for
Step10: //initialize particle's best position and velocity
          $\mathbf{s_r}(h+1) = \mathbf{s_r}(h) + \boldsymbol{\gamma}\mathbf{1_r}(\mathbf{f_r} - \mathbf{z_r}(h)) + \boldsymbol{\gamma}_{\mathbf{2r}}(X - \mathbf{z_r}(h))$
       New velocity
          $\mathbf{z_i}(h+1) = \mathbf{z}_{\mathbf{r}}(h) + \mathbf{s_r}(h+1)$
 Where
  r- particle index
  h- discrete time index
  $s_r$ – velocity of r^th particle
  $z_r$ – position of r^th particle
  $f_r$- found best place by r^th particle [ best, personal]
 X- found best place by swarm (global best, best of bests, Personal)
 $X_{(1,2)r}$- random number on the interval [0,1] which is executed to the r^th particle
Step11: $\mathbf{ft_r}$=$\mathbf{z_r}$
Step12: // update global best position
Step13: if $\mathbf{f(ft_r)} < \mathbf{f(jt)}$
Step 14:    $\mathbf{jt = ft_r}$
Step15: end if
Step17: end for
```

IV. RESULT

NB_PSO

Table 4.1 NB_PSO output table

Accuracy	79.48
Precision	76.24
Recall	82.55

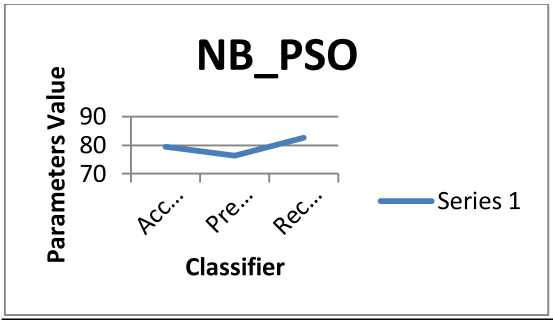


Fig. 1: Graph of NB_PSO outputs

In Table 4.1 and Fig. 4.2 analysis of tweet classification by naïve Bayes hybridization Particle swarm optimization which reduces the noise of features by given the relative weight to feature, in this learning process done by Naïve Bayes but PSO work on feature set. After Naïve bayes training test the model in which use tenfold cross validation. Which divide the training set into ten parts and one part given to test. Which every time change then calculate average accuracy precision and recall. In case of NB accuracy 79.48%, precision 76.24% which represent false positive error reduce. But recall 82.55% means false negative error more reduces false positive error.

SVM_PSO

Table 4.2

Accuracy	81.81
Precision	85
Recall	80

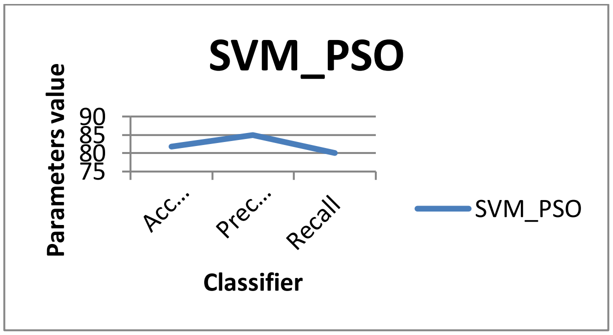


Fig. 2: Graph of SVM_PSO outputs

Table 4.3 Comparison of both SVM and NB

Classifier	Accuracy	Precision	Recall
SVM	71.42	76.38	69.44
SVM_ACO	75	79.36	73.01
SVM_PSO	81.81	85.0	80.0
NB	74.0	71.79	76.93

NB_ACO	77.79	74.85	79.0
NB_PSO	76.93	76.24	82.55

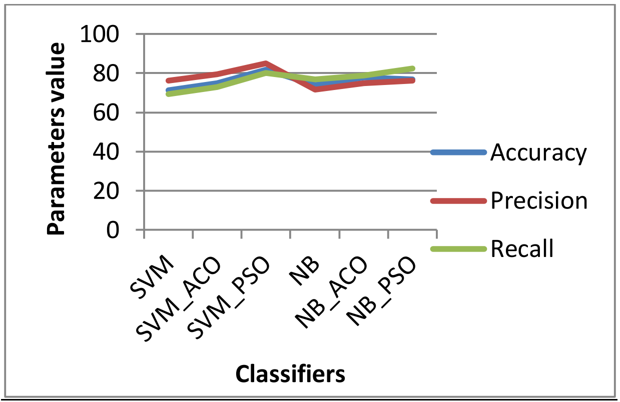


Fig. 3: Comparison Graph of SVM and Naïve Baiyes

In Fig: 4.3 compare all experiment by one graph which shows that SVM_ACO and SVM_PSO performs better than SVM. NB_ACO and NB_PSO performs better than NB but if compare between hybrid approaches then SVM_PSO it show 81.80% accuracy, 85% precision, and 80% recall. In case of naïve Bayes NB_PSO 76.93% accuracy, 76.24 precision, and 82.55% recall. Therefore experiments conclude that Naive Bayes improves Recall and SVM improve Precision and Accuracy when used as a hybrid approach

V. CONCLUSION AND FUTURE WORK

In this paper by using Machine learning Techniques there is an analysis on the sentiments of twitter. Here we use some features like Bigram, Unigram, etc. for the analyses of sentiments. Here we are using set of features to improve the effectiveness and precision of the classifier which helps to show the comparative analysis of accuracy and precision between four algorithms showing the effect of features optimization. In case of the second classifier, Naïve Bayes with ACO shows effective precision as compare to only naive Bayes, but both are not effective in comparison to SVM and SVM with ACO is more effective among all of them. In case of Naïve Bayes, precision increases by when used with ACO and in case of SVM, increase is obtained when used with ACO. In experiment analysis compare with classifier and hybrid classifier for that use SVM and naïve Bayes classifier which hybrid with PSO and ACO for effective feature weight. In figure 4.9 compare all experiment by on graph which show that SVM_ACO and SVM_PSO better perform than SVM. NB_ACO and NB_PSO perform better than NB but if compare between hybrid approaches then SVM_PSO show 81.80% accuracy,85% precision and 80% recall. IN case of naïve Bayes NB_PSO 76.93% accuracy,76.24 precision and 82.55% recall, so experiments conclude that Naive Bayes improve recall and SVM improve precision and accuracy when use as hybrid approach.

In future this work enhance on two parameters

1. Feature Extraction:

Enhance features: Improve the features set by reducing sparsely in features by ngram approach or NLP9natural language related features which reduce the information loss and improve the accuracy.

2. Optimization feature selection: Improve the feature selection by hybrid approach of optimization as in this improve the accuracy.

VI. REFERNCES

[1]. Duque Barrachina and O' Driscoll. "A big data methodology for categorising technical support requests using Hadoop and Mahout" Journal of Big data, Springer, doi: 10.1186/2196-1115-1, 2014

[2]. Chen, Min, Shiwen Mao, and Yunhao Liu. "Big data: a survey." Mobile Networks and Applications Vol. 19, pp.171-209, 2014.

[3]. Ibrahim Hassan and Targio Abaker. "The Rise of "Big Data" on cloud computing. Review and open research issues." Information Systems" Vol. 47, New York, USA, pp. 98-115, 2015.

[4]. Ioannis Partalas. "Web-scale classification: web classification in the big data era." 7th International conference on Web Search and Data Mining. ACM, New York, US, 2014.

[5]. Jonathan Stuart, and Adam Barker. "Undefined by data: a survey of big data definitions." Ar Xivpreprint arXiv: 1309.5821 2013.

[6]. Khan, Farhan Hassan, Saba Bashir, and Qamar Usman. "TOM: Twitter opinion mining framework using hybrid classification scheme." Decision Support Systems" Amsterdam, Netherlands, Vol. 57, pp. 245-257, 2014.

[7]. Lu Guofan, Qingnian Zhang, Zhao Chen. Telecom Data processing and analysis based on Hadoop. Received 1 October 2014: Computer Modeling & New Technologies 2014 18(12B) pp. 658-664, 2014.

[8]. Min Chen . Shiwen Mao, Yunhao Liu " Big Data : A Survey: Science plus Business Media New York 2014. Springer Mobile Netw Appl, Springer, Vol.19, pp.171–209, 2014.

[9]. Hao, Ming, et al. "Visual sentiment analysis on twitter data streams." Visual Analytics Science and Technology (VAST),IEEE Conference, Providence, RI, USA, pp. 23-28, 2011.

[10]. Le, Bac, and Huy Nguyen. "Twitter sentiment analysis using machine learning techniques. " Advanced Computational Methods for Knowledge Engineering. Springer, pp. 279-289, 2015.