Novel Approach for Tweet Sentiment Analysis by Convex Optimization of features weight

Aamena Khanam, Akansha Dubey
RITM LUCKHNOW (UP), India

Abstract - Tweet sentiment analysis has been an effective and valuable technique in the sentiment analysis domain. Machine learning algorithms are used to perform sentiment analysis; however, data quality issues such as high dimensionality, class imbalance or noise may negatively impact classifier performance Based on our experiments, 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 shows 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.

I. INTRODUCTION

The most common definition describes characteristics of big data as volume, velocity and variety. Volume refers to the massive size of big datasets. Velocity refers to the rate at which data are generated and must be acted upon, such as filtered, reduced, transferred and analyzed, as opposed to stored for future processing [7]. Variety refers to the diverse data forms in big data, including structured (tabular such as in a spreadsheet or relational database), unstructured (such as text, imaging, video, and audio), and semi-structured (such as XML documents)[5]. Today, the textual data on the internet is growing rapidly. Several kinds of industries are trying to use this massive textual data for extracting the people's views towards their products. Social media is a crucial source of information in this case. It is not possible to manually investigate the heavy amount of data. This is where the requirement of automatic classification becomes clear. The popularity of micro blogging stems from its distinctive communication services such as portability, immediacy, and ease of use, which allow users to instantly respond and spread information with limited or no restrictions on content. Twitter is currently the most popular and fastest-growing microblogging service, with more than 140 million users producing over 400 million tweets per day—mostly mobile—as of June 2012.Twitter enables users to post status updates, or tweets, no longer than 140 characters to a network of followers using various communication services. Tweets have reported everything from daily life stories to latest local and worldwide events. Twitter content reflects real-time events in our life and contains rich social information and temporal attributes. Monitoring and analyzing this rich and continuous flow of user-generated content can yield unprecedentedly valuable information.

II. LITERATURE REVIEW

Arantxa Barrachina Arantxa Duque et.al. [1]: Technical Support call centres frequently receive several thousand customer queries on a daily basis. Traditionally, such organisations discard data related to customer enquiries within a relatively short period of time due to limited storage capacity. This paper proposes a Proof of Concept (PoC) end to end solution that utilises the Hadoop programming model, extended ecosystem and the Mahout Big Data Analytics library for categorising similar support calls for large technical support data sets. The proposed solution is evaluated on a VMware technical support dataset.

Chen Min, et.al. [2]: They review the background and state-of-the-art of big data. They first introduce the general background of big data and review related technologies, such as could computing, Internet of Things, data centers, and Hadoop. Then focus on the four phases of the value chain of big data, i.e., data generation, data acquisition, data storage, and data analysis. For each phase, they introduce the general background, discuss the technical challenges, and review the latest advances. Finally examine the several representative applications of big data, including enterprise management, Internet of Things, online social networks, medial applications, collective intelligence, and smart grid..

Hashem Ibrahim Abaker Targio et al[3]: Massive growth in the scale of data or big data generated through cloud computing has been observed. Addressing big data is a challenging and time-demanding task that requires a large computational infrastructure to ensure successful data processing and analysis. The rise of big data in cloud computing is reviewed in this study. The definition, characteristics, and classification of big data along with some discussions on cloud computing are introduced.

Ioannis Partalas et al[4]: This paper provides an overview of the workshop Web-Scale Classification: Web Classification in the Big Data Era which was held in New York City, on February 28th as a workshop of the seventh International Conference on Web Search and Data Mining. The goal of the workshop was to discuss and assess recent research focusing on classification and mining in Web-scale category systems.

Jonathan Stuart Ward and Adam Barker [5]: The term big data has become ubiquitous. Owing to a shared origin between academia, industry and the media there is no single unified definition, and various stakeholders provide diverse and often contradictory definitions. The lack of a consistent definition introduces ambiguity and hampers discourse relating to big data. This short paper attempts to collate the various definitions which have gained some degree of traction and to furnish a clear and concise definition of an otherwise ambiguous term.

Lee Seungbae, et.al. [6]:. Significant innovations in mobile technologies are enabling mobile users to make real-time actionable decisions based on balancing opportunities and risks to take coordinated actions with other users in their workplace. This requires a new distributed analytic framework that collects relevant information from internal and external sources, performs real-time distributed analytics, and delivers a critical analysis to any user at any place in a given time frame through the use of mobile devices such as smart phones and tablets.

Lu Guofan, et.al. [7]: For call tracking system to adapt to the needs of large data processing, combined with a strong competitive advantage in recent years in large data processing Hadoop platform, designed and implemented a Hadoop-based call tracking data processing model, in order to verify its feasibility. The call tracking processing system model contains an analog data source module, data processing module, and a GUI interface.

Min Chen et al[8]: In this paper, we review the background and state-of-the-art of big data. We first introduce the general background of big data and review related technologies, such as could computing, Internet of Things, data centers, and Hadoop. We then focus on the four phases of the value chain of big data, i.e., data generation, data acquisition, data storage, and data analysis.

Ming Hao et al[9]: In this paper, to explore high-volume twitter data, they introduced three novel time-based visual sentiment analysis techniques: (1) topic-based sentiment analysis that extracts, maps, and measures customer opinions; (2) stream analysis that identifies interesting tweets based on their density, negativity, and influence characteristics; and (3) pixel cell-based sentiment calendars and high density geo maps that visualize large volumes of data in a single view. They applied these techniques to a variety of twitter data, (e.g., movies, amusement parks, and hotels) to show their distribution and patterns, and to

Le, Bac, et al. [10] In this paper, they introduce an approach to selection of a new feature set based on Information Gain, Bigram, Object-oriented extraction methods in sentiment analysis on social networking side. In addition, they also propose a sentiment analysis model based on Naive Bayes and Support Vector Machine. Their purpose is to analyse sentiment more effectively. This model proved to be highly effective and accurate on the analysis of feelings

Hao, Ming, et al.[11] Author introduce three novel time-based visual sentiment analysis techniques: (1) topic-based sentiment analysis that extracts, maps, and measures customer opinions; (2) stream analysis that identifies interesting tweets based on their density, negativity, and influence characteristics; and (3) pixel cell-based sentiment calendars and high density geo maps that visualize large volumes of data in a single view. Author applied these techniques to a variety of twitter data, (e.g., movies, amusement parks, and hotels) to show their distribution and patterns, and to identify influential opinions.

Saif, Hassan et. al. [12] In this paper, Author presented a novel approach of including semantics as extra highlights into the preparation set for conclusion investigation. For each extricated substance (e.g. iPhone) from tweets, we include its semantic idea (e.g. "Apple item") as an extra element and measure the relationship of the delegate idea with the negative/positive opinion. We apply this way to deal with foresee feeling for three diverse Twitter datasets. Our outcomes demonstrate a normal increment of F symphonious exactness score for distinguishing both negative and positive opinion of around 6.5% and 4.8% over the baselines of unigrams and grammatical form includes separately. We likewise look at against an approach in view of conclusion bearing subject investigation and locate that semantic highlights deliver better Recall and F score while characterizing negative supposition, and better Precision with bring down Recall and F score in positive sentiment classification.

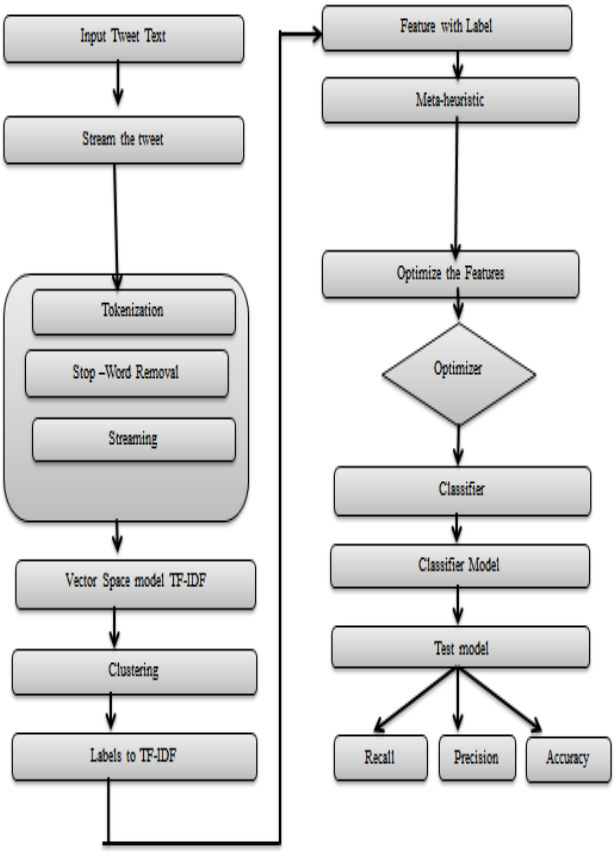
Kontopoulos, Efstratios, et al. [13] This paper proposes the arrangement of unique philosophy based systems towards a more effective sentiment analysis of Twitter posts. The curiosity of the proposed approach is that posts are not just described by a sentiment score, similar to the case with machine learning-based classifiers, however rather, get a sentiment review for each particular idea in the post. Generally speaking, our proposed design brings about a more itemized analysis of post sentiments with respect to a particular point.

Gokulakrishnan, Balakrishnan, et al. [14] This paper talks about an approach where an advertised stream of tweets from the Twitter microblogging website are preprocessed and arranged in light of their passionate substance as positive, negative and superfluous; and examinations the execution of different ordering calculations in view of their accuracy and

review in such cases. Further, the paper epitomizes the uses of this examination and its constraints.

Bifet, Albert et.al.[15] In this paper, the author talked about the difficulties that Twitter information streams pose, concentrating on order issues, and afterward consider these streams for supposition mining and sentiment analysis. To manage gushing unequal classes, we propose a sliding window Kappa measurement for assessment in time-changing information streams. Utilizing this measurement we play out an investigation on Twitter information utilizing learning calculations for information streams.

III. PROPOSED METHODOLOGY



Step1: Input the tweet text by continue 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 metaheuristic like PSO and ACO.
Step6: Hybrid the metaheuristic with classifier and make the classifier model.
Step7: Check the performance of classifier model by precision, recall and accuracy.

IV. ALGORITHM USED

Classification using Naïve Bayes

Naïve Bayes classifier is the one among the family of probabilistic classifiers in machine learning, between features which is generally based on naïve independence assumption with applied Bayes' theorem. Number of features is required to be parameter linear in learning problem where Naïve Bayes classifier is highly scalable.

The probability distribution over the set of features: $P(x) = P(c_i)P(X_1,X_2,X_3\ldots\ldots X_n/c_i)$

$$P(X_1,X_2,X_3\ldots\ldots X_n/c_i) = \prod_{i=1}^{n} P(^{X_n}/_{c_i})$$

$$P(x) = \prod_{i=1}^{k} P(c_i)P(^{x_n^d}/_{c_i})$$

Where
$X_{1,2,..n}$ features values to certain class label c,
k ←is the number of classes,
c_i ←is the i^{th} class

3.5 Classification using Support Vector Machine

Vapnik introduced for support vector machine, and is popular tool for supervised machines learning methods which are based on the minimization of the structural risk. The SVM basic characteristics is the original non-linear data into data class and the separation margin among itself is maximized and typing points nearer from the support vectors.

The training sample is

$$n = \{(u_i, v_i)|i = 1, 2, \ldots., m\}$$

Where

m←Sample no.

$\{u_i\} \in r_k$←Input vector set

$v \in \{-1,1\}$←Desired corresponding input vector

Then, optimal classification of existing hyper-plane has following condition to meet:

$$\begin{cases} \omega^t u_i + B \le 1, v_i = 1 \\ \omega^t u_i + B \le -1, v_i = -1 \end{cases}$$

Where

ω^τ←Super plane omega vector,

B←offset quality

Then, the decision function is classified as:

$$F(u_i) = sgn(\omega^t u_i + B)$$

SVM classification model is described with optimization model $min_{\omega,\xi,B} P(\omega, \xi)$

$$min_{\omega,\xi,B} P(\omega, \xi_i) = \frac{1}{2}\omega^t\omega + \frac{1}{2}\gamma\sum_{i=1}^{m}{\xi_i}^2$$

$$v_i[\omega^t\phi(u_i) + B = 1 - \xi_i, i = 1,2, \dots\dots, m$$

$$\xi = (\xi_1, \xi_2, \dots\dots \xi_m)$$

Where

ξ_i←slack variable

B←offset

ω←support vector

γ←classification parameter for balancing the model complexity and fitness error.

Transforming the optimization problem into dual space and for solving it, Lagrange function is introduced:

$$l(B, \omega, \alpha, \xi) = \frac{1}{2}\omega^t\omega + \frac{1}{2}\gamma\sum_{i=1}^{m}\xi_i^2 - \sum_{A=1}^{m}\alpha_i\{v_i[\omega^t\phi(u_A) + B] - 1 + \xi_i\}$$

Where

α_i←Lagrange multiplier

Then, describing the classification decision function:

$$F(x_i) = sgn(\sum_{i=1}^{m}\alpha_i v_i A(u, u_i) + B)$$

Algorithm 3: SVM _PSO Module

Step 1: SVM classification model is described with optimization model $\boldsymbol{min_{\omega,\xi,B} P(\omega, \xi)}$

$$\boldsymbol{min_{\omega,\xi,B} P(\omega, \xi_i) = \frac{1}{2}\omega^t\omega + \frac{1}{2}\gamma\sum_{i=1}^{m}{\xi_i}^2}$$

$$\boldsymbol{v_i[\omega^t\phi(u_i) + B = 1 - \xi_i, i = 1,2, \dots\dots, m}$$

$$\boldsymbol{\xi = (\xi_1, \xi_2, \dots\dots \xi_m)}$$

Where

$\boldsymbol{\xi_i}$←Slack variable

B←Offset

$\boldsymbol{\omega}$←Support vector

$\boldsymbol{\gamma}$←Classification parameter for balancing the model complexity and fitness error.

Step2: SVM classification model is described with optimization model $\boldsymbol{min_{\omega,\xi,B} P(\omega, \xi)}$

$$\boldsymbol{min_{\omega,\xi,B} P(\omega, \xi_i) = \frac{1}{2}\omega^t\omega + \frac{1}{2}\gamma\sum_{i=1}^{m}{\xi_i}^2}$$

$$\boldsymbol{v_i[\omega^t\phi(u_i) + B = 1 - \xi_i, i = 1,2, \dots\dots, m}$$

Step 3:Then, describing the classification decision function:

$$\boldsymbol{F(x_i) = sgn(\sum_{i=1}^{m}\alpha_i v_i A(u, u_i) + B)}$$

Step 4: Calculate accuracy, precision and recall.

Step 5: In PSO model for each particle **i** in **S do**

Step6 : for each dimension **d** in **D do**

Step7: //initialize each particle's position and velocity

Step8: $\boldsymbol{x_{i,d} = Rnd(x_{max}, x_{min})}$

Step9: $\boldsymbol{v_{i,d} = Rnd(-v_{max}/3, v_{max}/3)}$

Step10: end for

Step11: //initialize particle's best position and velocity

$$\boldsymbol{v_i(k+1) = v_i(k) + \gamma 1_i(p_i - x_i(k)) + \gamma_{2i}(G - x_i(k))}$$

New velocity

$$\boldsymbol{x_i(k+1) = x_i(k) + v_i(k+1)}$$

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
Step12: $pb_i = x_i$
Step13: // update global best position
Step14: if $f(pb_i) < f(gb)$
Step 15: $gb = pb_i$
Step16: end if
Step17: end for

Algorithm 4: NB_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,**

$P(y_j)$ **is the conditional class probability density function.**

Step 10: **Calculate probability distribution over the set of features:** $P(x) = \prod_{i=1}^{k} P(c_i) P(x_n^d / c_i)$

Where

k is the number of classes,

c_i **is the i[th] class.**

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

V. RESULT

Table 1.1 Comparison Of SVM, SVM_ACO, SVM_PSO

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

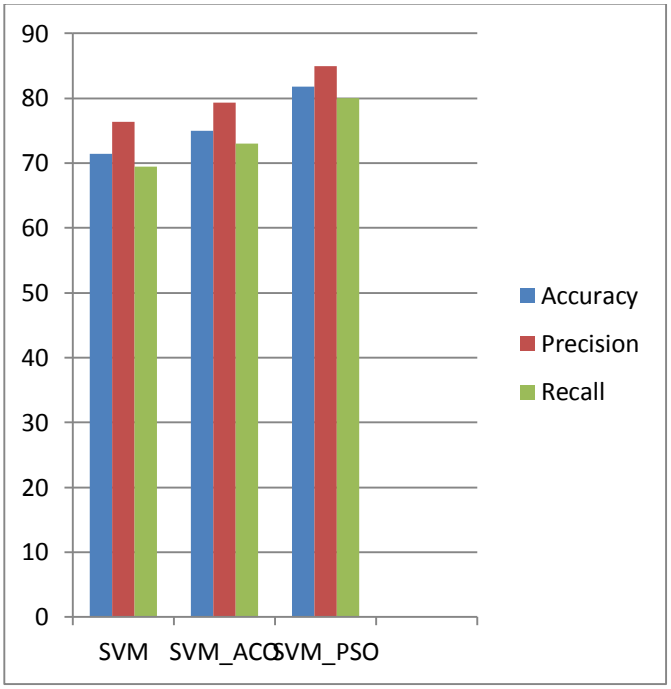


Fig.1: Comparison of results SVM, SVM-ACO, SVM_PSO

In above given table and graph analysis of tweet classification by SVM, SVM_ACO and SVM_PSO comparison on the basis of accuracy, precision and recall. IN graph clear represent SVM hybrid with optimization (PSO and ACO) perform significantly well compare to SVM. If compare SVM_PSO and SVM_ACO. In SVM_PSO perform well because of features local and global optimization.

Table 1.2 Comparison of NB, NB_ACO , NB_PSO

Classifier	Accuracy	Precision	Recall
NB	74.0	71.79	76.93
NB_ACO	77.79	74.85	79.0
NB_PSO	79.48	76.24	82.55

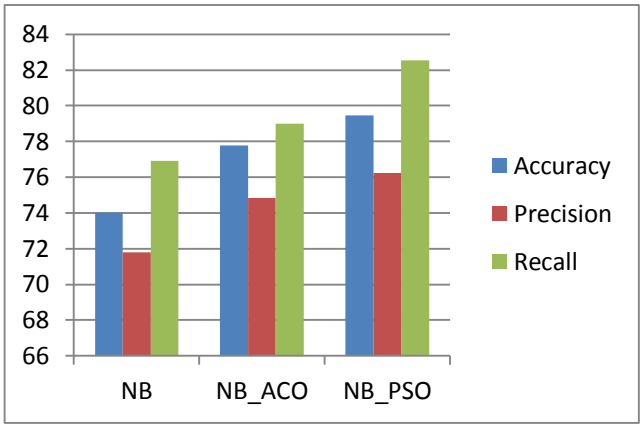


Fig.2: Comparison graph of NB, NB_ACO, NB_PSO

In above given table and graph analysis of tweet classification by NB, NB_ACO and NB_PSO comparison on the basis of accuracy, precision and recall. IN graph clear represent SVM hybrid with optimization (PSO and ACO) perform significantly well compare to SVM. If compare NB, NB_ACO and NB_PSO. In NB_PSO perform well because of features local and global optimization.

Table 4.9 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

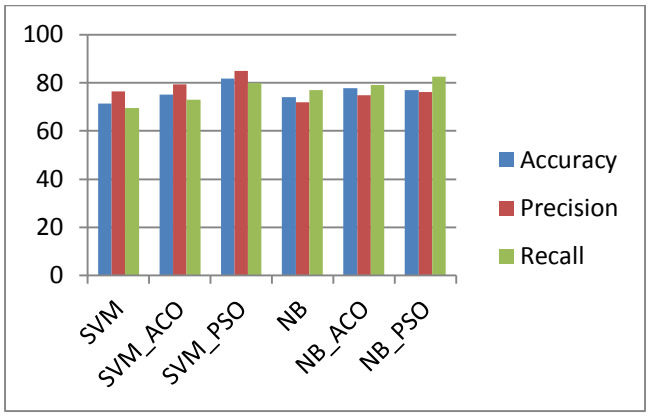


Fig.3: Comparison Graph of SVM and Naïve Baiyes

VI. CONCLUSION AND FUTURE SCOPE

In this paper analyzes sentiment on Twitter using Machine Leaning Techniques. Another consideration is that we applied Bigram, Unigram, Object-oriented features as an effective feature set for sentiment analysis. Used a good memory for resolving features better. However, we chose an effective feature set to enhance the effectiveness and the accuracy of the classifiers shows 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. Feature Extraction:

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

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

VII. REFERENCES

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

[2]. Chen, Min, Shiwen Mao, and Yunhao Liu. "Big data: a survey." *Mobile Networks and Applications* 19.2 (2014): 171-209.

[3]. Hashem, Ibrahim Abaker Targio, et al. "The rise of "big data" on cloud computing: Review and open research issues." *Information Systems* 47 (2015): 98-115.

[4]. Ioannis Partalas,, et al. "Web-scale classification: web classification in the big data era." *Proceedings of the 7th ACM international conference on Web search and data mining*. ACM, 2014.

[5]. Jonathan Stuart, and Adam Barker. "Undefined by data: a survey of big data definitions." *arXiv preprint arXiv:1309.5821* (2013).

[6]. Lee Seungbae , Kanika Grover, Alvin Lim. Enabling actionable analytics for mobile devices: performance issues of distributed analytics on Hadoop mobile clusters.USA Journal of Cloud Computing: Advances, Systems and Applications 2013, 2:15: doi: 10.1186/2192-113X-2-15.

[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) 658-664.

[8]. Min Chen, Shiwen Mao, Yunhao Liu. Big Data: A Survey: Science+Business Media New York 2014. Springer Mobile Netw Appl (2014) 19:171–209 .DOI 10.1007/s11036-013-0489-0.

[9]. Hao, Ming, et al. "Visual sentiment analysis on twitter data streams." *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*. IEEE, 2011

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

[11]. Hao, Ming, et al. "Visual sentiment analysis on twitter data streams." *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*. IEEE, 2011.

[12]. Saif, Hassan, Yulan He, and Harith Alani. "Semantic sentiment analysis of twitter." *The Semantic Web–ISWC 2012* (2012): 508-524.

[13]. Kontopoulos, Efstratios, et al. "Ontology-based sentiment analysis of twitter posts." *Expert systems with applications* 40.10 (2013): 4065-4074.

[14]. Gokulakrishnan, Balakrishnan, et al. "Opinion mining and sentiment analysis on a twitter data stream." *Advances in ICT for emerging regions (ICTer), 2012 International Conference on*. IEEE, 2012.

[15]. Bifet, Albert, and Eibe Frank. "Sentiment knowledge discovery in twitter streaming data." *International conference on discovery science*. Springer, Berlin, Heidelberg, 2010.