International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No.07, March 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

Novel Approach of Tweet Sentiment analysis by intelligent optimization with Machine learning Approach

Punita Bhardwaj[1], Aman Kumar[2], Ruchi Sngh[3]

Research Scholar[1], Assistant Professor[2],Assistant Professor[3]
Department of Computer Science and Engineering, IRIET Solan, Himachal Pradesh[1, 2 ,3]

ABSTRACT

In this paper analysis the sentiment classification by optimization with machine learning approaches. In the 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 Fig. 4.9 compare all experiment by ongraph 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 81.81% accuracy,82.55% precision and 82.5% recall.IN case of naïve Bayes NB_PSO 85.12% accuracy,79.61 precision and 79.93% recall, so experiments conclude that Naive Bayes improve Accuracy and SVM improve precision and recall when use as hybrid approach

Keywords:SVM(Support Vector Machine),PSO(Particle Swarm Optimization),ACO(Ant Colony Optimization),Naïve Bayes.

I. INTRODUCTION

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. 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 analysing this rich and continuous flow of user-generated content can yield unprecedentedly valuable information.

Sentiment analysis is the task of finding the opinions and affinity of people towards specific topics of interest. Be it a product or a movie, opinions of people matter, and it affects the decision-making process of people. The first thing a person does when he or she wants to buy a product online is to see the kind of reviews and opinions that people have written. Social media such as Facebook, blogs, twitter have become a place where people post their opinions on certain topics. The sentiment of the tweets of a particular subject has multiple usages, including stock market analysis of a company, movie reviews, in psychology to analyze the mood of people that

has a variety of applications, and so on. Sentiments of tweets can be categorized into many categories like positive, negative, neutral, extremely positive, extremely negative, and so on. The two types of sentiments considered in this classification experiment are positive and negative sentiments. The data, being labeled by humans, has a lot of noise, and it is hard to achieve good accuracy.

II. LITERATURE SURVEY

Arantxa et al. discussed technical support call centres frequently receive several thousand Customers 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 [1].

Chen et al. highlighted their views the background and state-of-the-art of big data. Firstly introduce the general background of big data and review related technologies, such as cloud computing, Internet of Things, data centres, 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, also introduced the general background, discusses 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, medical applications, collective intelligence, and smart grid [2].

Hashemite et al. presented 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 [3].

Ioannis et al. 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 [4].

Jonathan and Barker focused on 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 [5].

Lee et.al presented 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 [6].

Lu et.al evaluated 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 [7].

Min et al. reviewed 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 centres, 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 [8].

Ming et al.explored high-volume twitter data, also 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. Applied these techniques to a variety of twitter data, (e.g., movies, amusement parks, and hotels) to show their distribution and patterns, and too [9].

Le et al.introduced 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, 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[10].

III.RESEARCH METHODOLOGY

Step 1: Input the tweet text by continues streaming of tweets.

Step 2: Pre-processing the text by tokenization streaming and stop-word removal. It is a process of removing the noisy and inconsistent data from the tweets. This process removes the stop word from the text and also the symbols. It replaces the abbreviation by the full form and removes the hash tags from the tweets. It deletes the repeated words from the tweets and makes them meaningful because it is not easy to understand the proper meaning of this type of text. Repeated words can also change the meaning of the text. Sometimes it is not possible to understand the emotion is the happy or in the sad mood.

Step 3: Make vector space model with help of TF-IDF (Term Frequency-Inverse document frequency). IDF is calculated by using following equation.

Inverse document frequency = $\log\left[\frac{totalnumberofdocuments}{numberofdocumentscontaining-t}\right]$

Step 4: Clustering the document according to its TF-IDF and make a label with the help of PCA (Principle component analysis). PCA method is numerical method. This techniques is called are dimensionality reduction methods. PCA method is able to explore the intrinsic variability of data.

Step 5: Optimize the feature of TF-IDF with the help of meta-heuristic like PSO and ACO.

Step 6: Hybrid the meta-heuristic with classifier and make the classifier model.

Step 7: Check the performance of classifier model by precision, recall and accuracy.

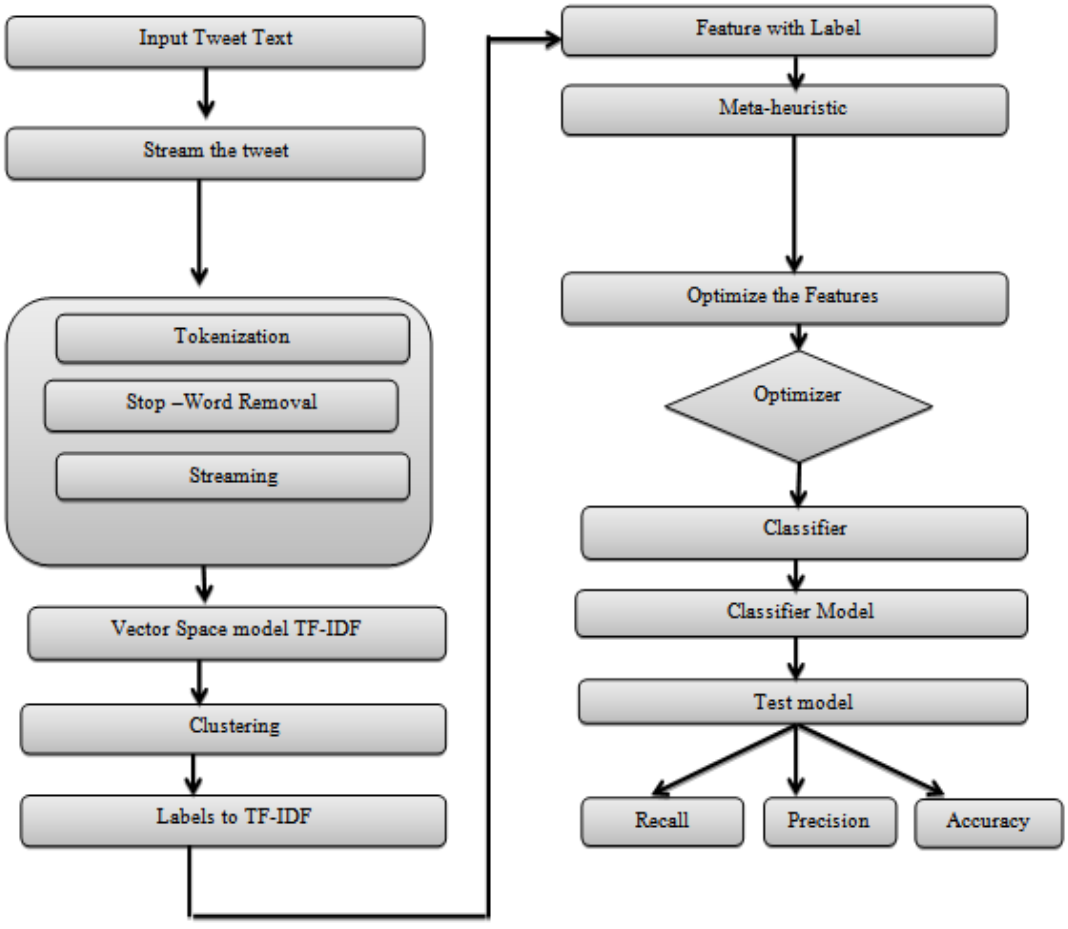

Figure 1: Flow Chart

IV. RESULTS

Table 1:Comparison of SVM, SVM_ACO, SVM_PSO

Classifier	Accuracy	Precision	Recall
SVM	71.42	72.91	72.96
SVM_ACO	75	76.19	76.19
SVM_PSO	81.81	82.5	82.5

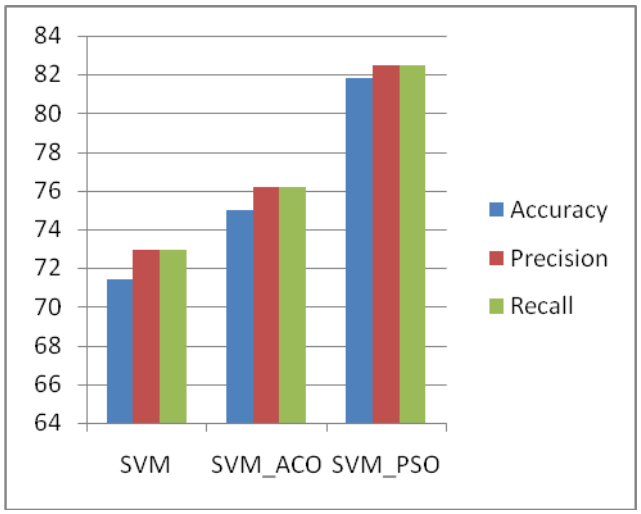

Figure2Comparison of results SVM, SVM-ACO, SVM_PSO

In Table 1 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 2: Comparison of NB, NB_ACO, NB_PSO

Classifier	Accuracy	Precision	Recall
NB	76.15	76.04	76.57
NB_ACO	78.05	75.04	79.30
NB_PSO	85.12	79.61	79.93

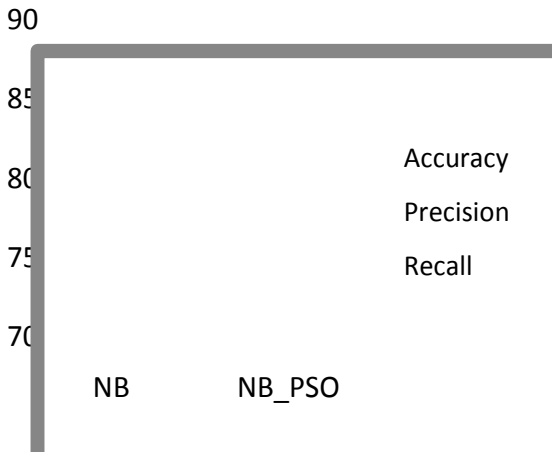

Figure3Comparison Graph of NB,NB_ACO, NB_PSO

In Table 2 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, NB_PSO perform well because of features local and global optimization.

Table 3: Comparison of both SVM and NB

Classifier	**Accuracy**	**Precision**	**Recall**
SVM	71.42	72.91	72.96
SVM_ACO	75	76.19	76.19
SVM_PSO	81.81	82.5	82.5
NB	76.15	76.04	76.57
NB_ACO	78.05	75.04	79.30
NB_PSO	85.12	79.61	79.93

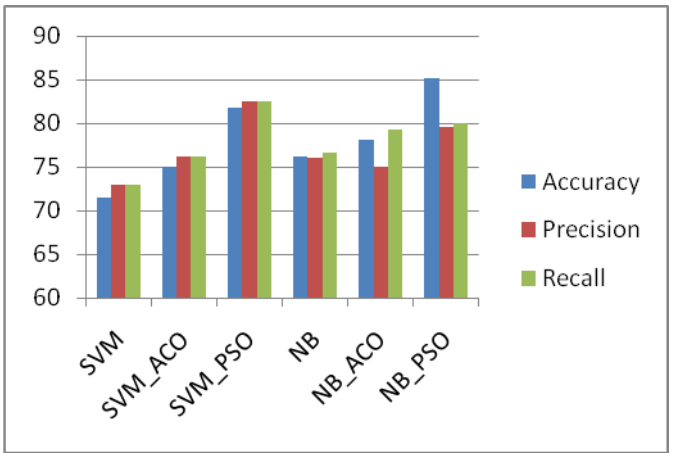

Figure4 Comparison Graph of SVM and Naïve Bayes

In Figure3 compares all experiment by one graph which shows that SVM_ACO and SVM_PSO perform better than SVM. NB_ACO and NB_PSO performs better than NB but if compare between hybrid approaches then SVM_PSO it show 81.81% accuracy,82.5% precision, and 82.5% recall. In case of naïve Bayes NB_PSO 85.12% accuracy,79.61 precision, and 79.93% recall. Therefore experiments conclude that Naive Bayes improves accuracy and SVM improve Precision and Recall when used as a hybrid approach.

IV. CONCLUSION

In this research, work sentiment analyzes of Twitter using Machine Leaning Techniques has been done. While consideration Bigram, Unigram, Object-oriented features has been applied, as an effective feature set for sentiment analysis.For this used a good memory for resolving features better. However, 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.

V FUTURE SCOPE

In future this work can be enhanced on two parameters:

- Firstly Enhance Feature Extraction: Improve the features set by reducing sparsely in features by n-gram approach or NLP natural language related features which reduce the information loss and improve the accuracy.
- Secondly Optimization features selection: Improve the feature selection by hybrid approach of optimization as in this improve the accuracy.

REFERENCES

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

[2] Chen, Min, Shiwen Mao, and Yunhao Liu. "Big data: a survey." *Mobile Networks and Applications* 19.2, Vol. 19, Page171-209. 2014.

[3] Ibrahim Hassan and TargioAbaker. "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] IoannisPartalas."Web-scale classification: web classification in the big data era." *Proceedings of the 7th ACM international conference on Web search and data mining*.ACM, New York, USA,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 GuofanQingnian 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+Business Media New York 2014.Springer, Vol.19 pp. 171–209, 2014.

[9] Hao, Ming. "Visual sentiment analysis on twitter data streams." *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference held at Providence, RI, USA, 2011*.

[10] AbinashTripathy, AnkitAgarwal, Santanu Kumar Rath. "Classification of Sentiment reviews using N-gram machine learning approach". Expert System With Applications. Elsevier, Page 117-126,2016

[11] BacLe, Huy Nguyen. "Twitter sentiment analysis using machine learning techniques." *Advanced Computational Methods for Knowledge Engineering*.Springer, Cham, 2015.pp. 279-289, 2015.