Predicting Risk in Sentiment Analysis using Machine Learning

Rakhi Gupta, Nashrah Gowalker, S.D. Joshi, Suhas Patil

***Abstract*: The purpose of this research is to do risk modeling after a sentiment analysis of Twitter posts based on a particular or certain sentiment with the help of the PRISM model .The model is named PRISM as the results obtained are an amalgamation of seven different attributes used in the research for comparison and tabulation of quantitative scores. These attributes are Accuracy, Precision, Recall, F1-Score, Support, Confusion Matrix, and Tweets. PRISM model can serve the law enforcement agencies in many ways and help them maintain peace, law and order in society as it is a proactive model.**

The sub-modules which are part of the PRISM model help to give quantitative values to predict the risk level on the sentiment of interest. After analysis of obtained testing results, it is observed that Support Vector Machine gives better results in accuracy, precision, F1-Score, Support and Recall as compared to the other three classifier models i.e. Naive Bayes, Decision Tree, and K nearest neighbor. It is also observed that with an increase or decrease in data, regarding the number of tweets, the fluctuation in performance of SVM is most stable i.e. it shows the least deviation and variation. The other algorithms show a considerable deviation in their performance.

Keywords* : *emotions, machine learning algorithms, risk modeling , sentiment analysis.

I. INTRODUCTION

In this research, we analyze posts of several users or a particular user to check whether they can be a cause of concern to the society or not. [Ahlgren, 2016] Every sentiment like happy, sad, anger and other emotions are going to provide scaling of severity in the conclusion of the final table on which machine learning algorithm is applied. This scaling will be of polarity negative, positive or neutral where less negative means less harmful and higher the value on the scale means it's more harmful. Positive words are given scales accordingly.

Based on the overall sentiment value generated from the final table, machine learning algorithms are applied for a comparative study between classifier models. Then risk modeling and analytics graph are created, which provide better visualization and help in making sound judgments. The purpose of the Crime Mitigation System is to find a possible suspicious activity and inform the concerned authorities to help them evade a possible crime from happening. This is done by risk modeling and risk identification. [Jagadish, 2014]The research work is based on storing, examining and analyzing the data being generated on Twitter in the form of tweets. Twitter embeds a large amount of data posted by individuals all over the world. Twitter also serves to carry out keyword searching which retrieves the tweets and all types of relevant data with respect to that keyword. By using this platform one can collect data related to his or her topic of interest and then perform an action on it. Using this real-time data the research analyses the data in terms of positivity and negativity which is being expressed in the Twitter tweets. These types of tweets give an understanding of the emotional quotient of the society related to the subject for which the data is collected. So performing analytics over this relevant data can help the higher authorities to understand the potential risk which may arise in the future. This may help to mitigate crime, prevent them from happening and take proactive action against it. Analyzing such datasets aims to decrease crime rates. Additionally, using machine learning, the study uses training and testing the classifier models such as Naive Bayes, Support Vector Machine, K-nearest neighbor and Decision Tree to predict the accuracy of the supervised data. [Gupta, 2017].

II. RESEARCH APPROACH

The research work is carried to develop a way where a huge amount of data is bought into the desired format for analysis and tested using different machine learning algorithms for Sentiment Analysis. The proposed methodology aims to aid and ease out the work of higher authorities in order to prevent critical incidents from taking place in society. In this research, a large amount of data is collected from Twitter and analyzed. That category of data which gives information about the emotional quotient of people on a particular subject is sieved out. It calculates how many false positive, false negative, true positive and true negative tweets are there in the data sets generated by the system. The dataset is picked up from yearly data (for 2 years i.e. 2017 and 2018) under monitoring from twitter under different timelines. It is put through experiments and testing which yield the results. These results give the law enforcement department the insight into the pulse of society related to the topic of interest. This enables and helps them to take proactive and preventive measures. Different machine learning algorithms (classification and clustering)
work over the data to predict the accuracy of the results generated by the previous modules of the proposed methodology.

Revised Manuscript Received on October 05, 2019
Rakhi Gupta, PHd Research Scholar , BVDUCOE , Pune
Nashrah Gowalker, Assistant Professor, K.C. College , Churchgate
Dr Suhas Patil , Professor, Comp, Engg ,BVDUCOE , Pune
Dr. S.D. Joshi, Professor, Comp, Engg ,BVDUCOE , Pune

III BLOCK DIAGRAM

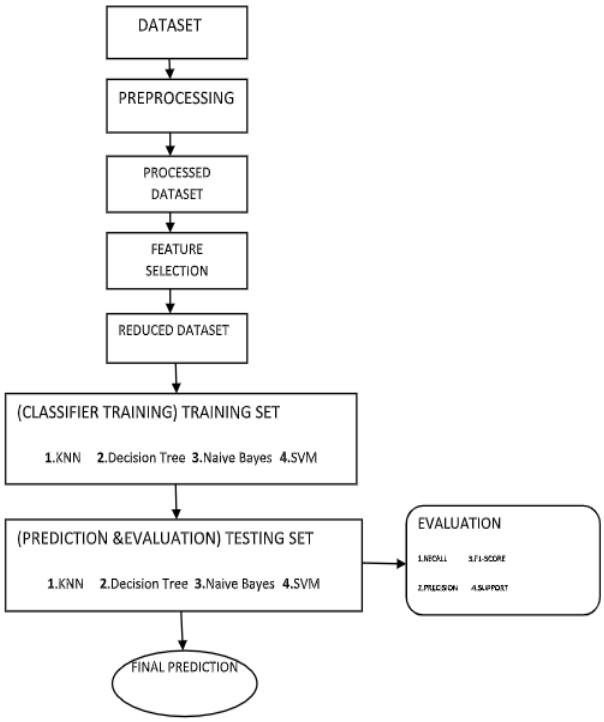

Fig : 1

Level 1

Retrieve data from Twitter

The data that is retrieved from Twitter is in raw format i.e. not processed. Processed data is free from expressions that occur regularly. So, initially, in order to access Twitter streaming API, we need to get four pieces of information from Twitter primarily: Access token, Access token secret, API secret, and API key. In this, we have use python library called tweepy to connect to Twitter streaming API and download the real-time data.

Level 2

Processing the Raw data

After receiving the raw data from Twitter, it needs to be processed in order to achieve the desired result. Raw data is then cleaned with the help of regular expression (regex) that is by removing the URL, username and special characters. All these URL, username and special characters are replaced by space.

Level 3

Textblob and Classifier

Data fetched from Twitter is processed and cleaned and brought into the desired format. Further, the data is passed through the textblob to attain the sentiment value of the tweet. Later the supervised data is divided into test data and training data. The classifier model is trained with the help of a training set which in return carries out the prediction. [Sarma, 2018]

IV OBTAINING DATA FOR HAND ROLLED NAIVE BAYES CLASSIFIER

Here the tweets are manually searched from the twitter and copied and pasted into two files. One for tweets about Naxal and other containing tweets that are not related to Naxal. For each class 150 tweets were collected and most of them contain the word Naxal or Naxalism.

For each class, the 150 tweets were shuffled and partitioned into 100 tweets in the training set and the remaining 50 tweets into a test set.

Hand rolled Naive Bayes classifier:

For each tweet, we first replace all occurrences of punctuation marks which are followed by the space character with a single space character. Then we split on whitespace and reject all tokens with 4 characters and less.

Later we train two classifiers: one for recognizing tweets about Naxal and others without Naxal.

$$P(\text{token}|\text{class}) = (xi+\alpha) / (N+\alpha*|V|)$$

where xi is the token count in the class, $\alpha=1$, N is the sum of all token counts in the given class, and $|V|$ is the size of the vocabulary in the entire training set (regardless of class). This is identical to the formula used for the Cross Validated question. The only difference is that it doesn't contain an additional +1 in the denominator.

To take care of tokens which are not present in the training set but are there in the test set, we default to using this probability:

$$P(\text{unseen token}|\text{class}) = 1 / (N+|V|)$$

When a tweet is passed, each classifier computes the sum of the log likelihood of every token in the tweet. We pass the tweet to both the classifiers and then compare the two log probabilities - the higher one wins and we conclude that the tweet belongs to that class.

V PROPOSED METHODOLOGY

The review of **"PRISM (PREDICTING RISK IN SENTIMENT ANALYSIS USING MACHINE LEARNING) "** is carried out to keep abreast with the latest technological advances in the field of information technology in general and involving machine learning modeling methodology in particular. Study of sentiments on social platform and risk associated with sentiments and related concepts are covered in the initial stage.

- Summary of this activity with the importance of modeling is represented.
- Necessary information is gathered from various sources at the information gathering phase. Data gathered from papers and websites is duly acknowledged in the bibliography and references.
- The problem is defined and System Engineering is carried out. How the defined problem can be a challenge for research and development is conveyed and also Risk management Tabulation is done.

The proposed methodology is a model called **PRISM (PREDICTING RISK IN SENTIMENT ANALYSIS USING MACHINE LEARNING)** which helps to collect relevant data from Twitter based on the keyword search method on a particular topic. On this data, a

dictionary of words that can define a sentiment of the tweet is stored. Textblob helps to calculate the value for the tweets which in return provides with sentiment values such as neutral, positive and negative with respect to the tweets. Further, multiple classifications and clustering machine learning algorithms that give quantitative analysis and comparison are run on this data. Based on the result of this quantitative analysis, the Confusion Matrix is developed. This effort will help to recognize the risk associated with the data that is being monitored and analyzed

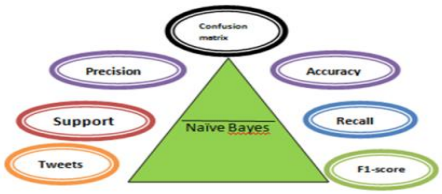

Fig. 2: PRISM

The confusion matrix is as follows on hand-rolled classifier:
positive , negative
Predictive positive a=67(TP) b=5(FN)
Predictive negative c=3(FP) d=65(TN)
Accuracy = 94.28%
Precision=0.95
Recall=0.93
F1-score=0.95

VI RESULTS AND DISCUSSION

Table 1 : Comparative Results Between The Algorithms

Data size (TWEETS)	K-NN	Decision Tree	Support Vector Machine	Naive Bayes
30	72.4	78.2	92.3	90.6
50	69.6	77.5	89.9	88.4
80	67.4	76.4	91.4	86.2
120	68.3	77.9	90.2	85.1
150	71.3	74.2	93.2	89.2

Description: The raw tweets are processed and classified into training data and test data. The research has trained the classifier by providing it with the training data to prepare it for quantitative analysis. The following points can further be elaborated based on the comparative results in Table 1.

From the study and analytics of table 1, it is inferred that Support Vector Machine gives better results compared to the other three classifier models i.e. Naive Bayes, Decision Tree, and K Nearest Neighbor. Also, it is observed, as the data size increases, the accuracy levels of SVM also increase in comparison to the other models. The data set provided to all the classifiers is real-time data which has been monitored over a period of time according to time series analysis.

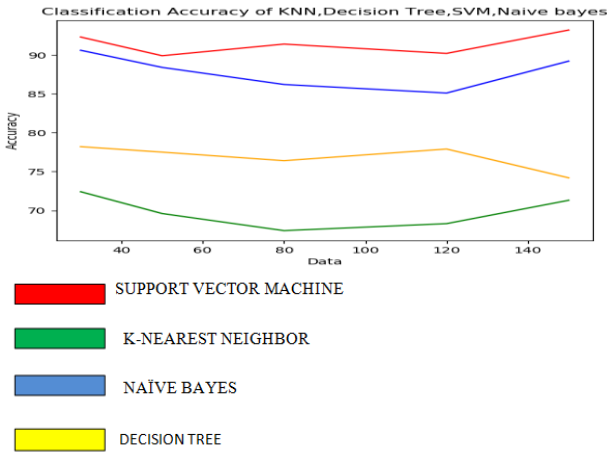

Fig. 3 : Statistical Comparison of Accuracy levels of various classifiers based on the data set.

Time Series Analysis

Time series data are a collection of ordered observations that are recorded at a specific time, for instance, hours, days, months, or years.

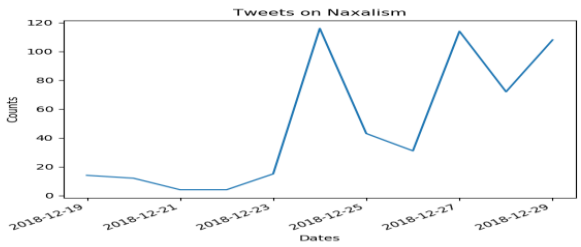

Fig. 4 : Time series analysis: This statistical graph is a collection of data monitored hourly, daily and on a weekly basis in the month of December 2018. (District Gadchiroli, Maharashtra).

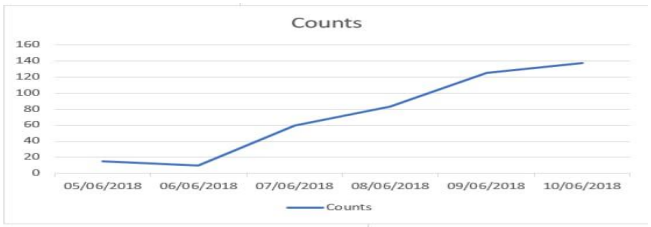

Fig. 5 : Time series analysis: This statistical graph is a collection of data monitored hourly, daily and on a weekly basis in the month of June 2018. (District Gadchiroli, Maharashtra).

Fig. 6 : Time series analysis: This statistical graph is a collection of data monitored hourly, daily and on a weekly basis in the month of January 2018. (District Gadchiroli, Maharashtra).

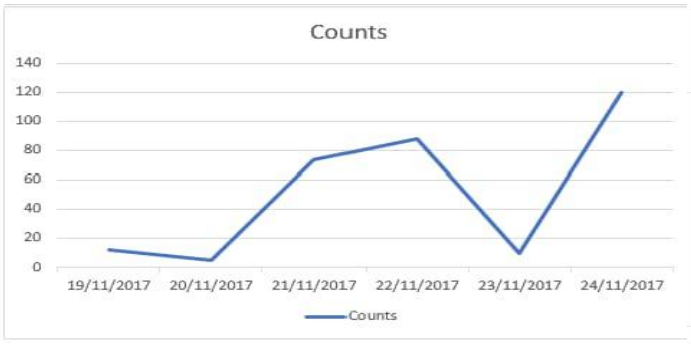

Fig. 7 : Time series analysis: This statistical graph is a collection of data monitored hourly, daily and on a weekly basis in the month of November 2017. (District Gadchiroli, Maharashtra).

Fig. 8 : Time series analysis: This statistical graph is a collection of data monitored hourly, daily and on a weekly basis in the month of May 2017. (District Gadchiroli, Maharashtra).

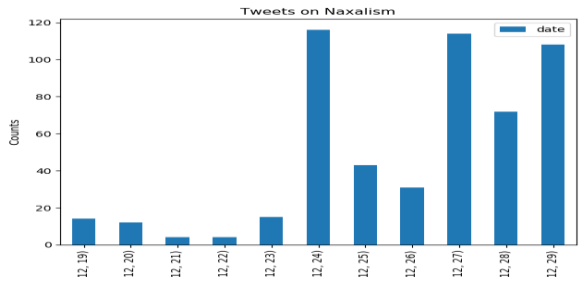

Fig. 9 : Bar graph

The above figure is a collection of data being represented in the bar chart format which lists down the number of tweets launched by the individuals on the topic Naxalism (District Gadchiroli, Maharashtra) in the month of December 2018.

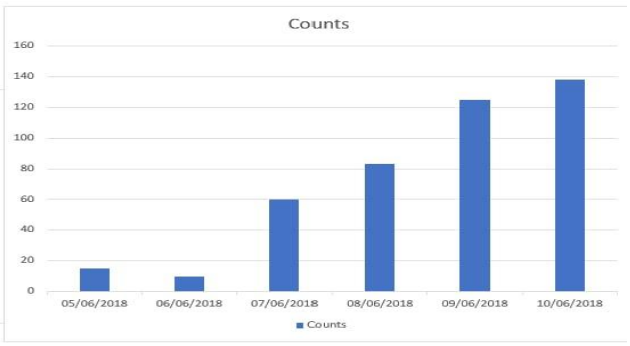

Fig. 10 : Bar graph

The above figure is a collection of data being represented in the bar chart format which lists down the number of tweets launched by the individuals on the topic Naxalism (District Gadchiroli, Maharashtra) in the month of June 2018.

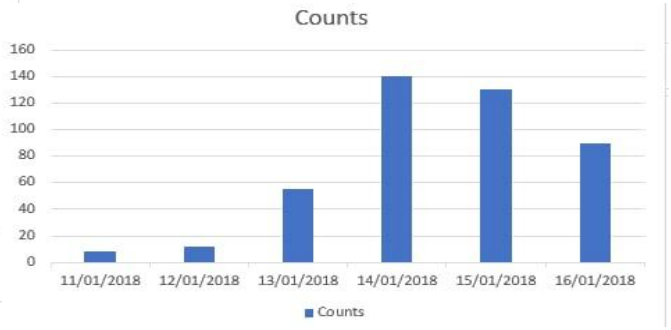

Fig. 11 : Bar graph

The above figure is a collection of data being represented in the bar chart format which lists down the number of tweets launched by the individuals on the topic Naxalism (District Gadchiroli, Maharashtra) in the month of January 2018.

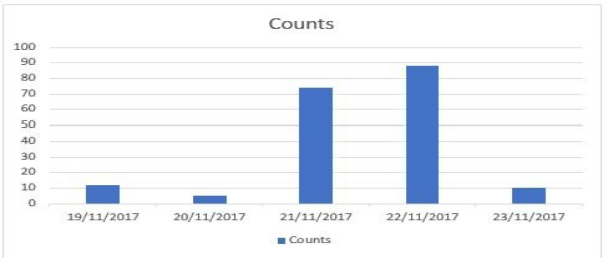

Fig. 12 : Bar graph

The above figure is a collection of data being represented in the bar chart format which lists down the number of tweets launched by the individuals on the topic Naxalism (District Gadchiroli, Maharashtra) in the month of November 2017.

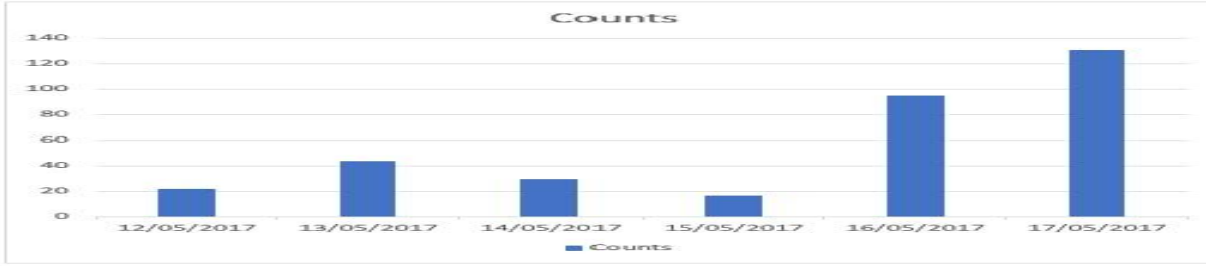

Fig. 13 : Bar graph

The above figure is a collection of data being represented in the bar chart format which lists down the number of tweets launched by the individuals on the topic Naxalism (District

Gadchiroli, Maharashtra) in the month of May 2017

VII COMPARISON OF CLASSIFIERS

Comparison of Precision, Recall, F1-score and Support values of Naive Bayes, K-NN, Support Vector Machine and Decision Tree.

Table 2: Comparison of Precision, Recall, F1-score and Support values of Naive Bayes, K-NN, Support Vector Machine and Decision Tree.

CLASSIFIERS	NAIVE BAYES	SVM	K-NN	DECISION TREE
PRECISION	0.92	0.94	0.53	0.85
RECALL	0.91	0.93	0.73	0.80
F1-SCORE	0.90	0.93	0.61	0.78
SUPPORT	11	14	11	5

VIII CONCLUSION

This research has focused on the comparison between various classification and clustering algorithms in sentiment analysis. The data used in this research has been collected from social media platform i.e. Twitter. It is monitored over a period of two years (2017 and 2018). Monitoring is based on Naxal activity in the Gadchiroli district in Maharashtra. An effort is made to find a pattern in increased activity on related sentiment on Twitter, as the date of incident comes closer. The data tested through machine learning algorithms is analyzed for a comparative study to compare the performances of different algorithms.Initially, we developed a hand rolled machine learning algorithm i.e. modified Hand Rolled Naive Bayes. In this algorithm, the data which was manually collected from Twitter and on which classification of features had been done is passed through the modified algorithm to predict the accuracy and also confusion matrix is derived. In confusion matrix, the parameters studied are true positive, true negative, false positive and false negative. Under standard metric evaluation, tabulation is also done for precision, recall, f1-score, and support.After analysis of testing results, it can be concluded that the performance and results of Modified Naive Bayes are better as compared to results from scikit Naive Bayes. When modified hand rolled

REFERENCES

1. Chirag Kansara , Rakhi Gupta , Dr. S.D. Joshi , Dr. S.H. Patil, “Crime Mitigation At Twitter Using Big Data Analytics and Risk Modelling “ , IEEE International conference on Recent Advances and innovations in Engineering (ICRAIE -2016) December 23-25 , 2016 , Jaipur , India.
2. Anuja P Jain , Padma Dandannavar , “Application of Machine Learning Techniques to Sentiment Analysis “ for 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT) , 978-1-5090-2399-8/16/$31.00_c 2016 IEEE
3. Geetika Gautam , Divakar yadav , “Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis ” for 978-1-4799-5173-4/14/$31.00 ©2014 IEEE
4. Mr. S. M. Vohra, 2 Prof. J. B. Teraiya,” A Comparative Study Of Sentiment Analysis Techniques”, Journal Of Information, Knowledge And Research In Computer Engineering Issn: 0975 – 6760| Nov 12 To Oct 13 | Volume – 02, Issue – 02 Pg 313-317
5. S Shayaa, NI Jaafar ,S Bahri , “ Sentiment Analysis of Big Data : Methods , Applications , and Open challenges, Vol 6 2018, DOI : 10.1109/ACCESS.2018.2851311, 2169-3536 , 2018 IEEE
6. B. Liu, E. Blasch, Y. Chen, D. Shen, G. Chen,“Scalable SentimentClassification for Big Data Analysis Using Naïve Bayes Classifier.”Scalable IEEE Intl Conf. on Big Data, Oct. 2013
7. Sentiment Analysis: A Comparative Study On Different Approaches: Devika M D, Sunitha C, Amal Ganesha: Fourth International Conference on Recent Trends in Computer Science & Engineering, Chennai, Tamil Nadu, India ,1877-0509, Procedia Computer Science 87 (2016) 44 –49, DOI:10.1016/j.procs.2016.05.124
8. Techniques for sentiment analysis of Twitter data: A comprehensive survey**:** Mitali Desai, Mayuri Mehta**:** DOI:10.1109/CCAA.2016.7813707
9. Big Data Analysis**:** Challenges and Solutions: Puneet Singh Duggal, Sanchita Paul : International Conference on Cloud, Big Data and Trust 2013, Nov 13-15, RGPV, pp 269-276.
10. Large-Scale Sentiment Analysis for News and Blogs :Namrata Godbole, Manjunath Srinivasaiah, Steven Skiena : Google Inc.
11. Study of Twitter Sentiment Analysis using Machine Learning Algorithms on Python: Bhumika Gupta, Ph.D., Monika Negi, Kanika Vishwakarma, Goldi Rawat, Priyanka Badhani : International Journal of Computer Applications (0975 8887) Volume 165 – No.9, May 2017
12. Comparative Study of Classification Algorithms used inSentiment Analysis : Amit Gupte, Sourabh Joshi, Pratik Gadgul , Akshay Kadam: ISSN: 0975-9646, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (5), 2014, 6261-6264
13. Machine Learning-Based Sentiment Analysis for Twitter Accounts: Ali Hasan, Sana Moin, Ahmad Karim, and Shahaboddin Shamshirb TonDucThang: Math. Computer Appl. 2018, DOI :10.3390
14. Emotion Detection Of Tweets Using Naïve Bayes Classifier: Hema Krishnan, M. Sudheep Elayidom, T. Santhanakrishnan: International Journal of Engineering Technology Science and Research, IJETSR ISSN 2394 – 3386 Volume 4, Issue 11 November 2017
15. Sentiment Analysis on Twitter: Akshi Kumar, Teeja Mary Sebastian: IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 3, July2012,, ISSN (Online): 1694-0814
16. Sentiment Analysis Techniques in Recent Works: Zohreh Madhoushi, Abdul Razak Hamdan, Suhaila Zainudin, Science and Information Conference 2015 July 28-30,2015, London, UK, DOI: 10.1109/SAI.2015.72371.
17. Emotion Analysis of Social Media Data using Machine Learning Techniques: Sonia Xylina Mashal, Kavita Asnani : IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p-ISSN: 2278-8727 PP 17-20, National Conference On Advances In Computational Biology, Communication, And Data Analytics (ACBCDA 2017)
18. Research On Sentiment Analysis: The First Decade: Oskar Ahlgren: 16th International Conference on Data Mining Workshops, 2016 IEEE, 2375-9259/16 © 2016 IEEE, DOI 10.1109/ICDMW.2016.94
19. Big Data And Its Technical Challenges: H.V. Jagadish, Johannes Gehrke, Alexandros Labrinidis, Yannis Papakonstantinou, Jignesh M. Patel, Raghu Ramakrishnan, Cyrus Shahabi : DOI 10.1145/2611567, Communications Of The ACM | July 2014 | Vol. 57 | No. 7
20. A Survey on Sentiment Analysis by using Machine Learning Methods: PengYang, Yunfang Chen: Department of the Internet of Things, Nanjing. 978-1-5090-6414-4/17/$31.00, 2017 IEEE 2nd Information Technology,Networking, Electronic and Automation Control Conference(ITNEC)

AUTHORS PROFILE

Rakhi Gupta , B.E. degree in computer engineering from ,IET , Lucknow University in 1993. M.E. Computer Engineering at Bharati Vidyapeeth Deemed University Pune in 2007 . She is currently pursuing her PHd in Computer Enggineering from Bharati Vidyapeeth Deemed University Pune . She is working in KC College , Churchgate , Mumbai as HOD , IT Department

Nashrah Gowalker received her B.Sc degree in Information Technology from Elphinstone College Mumbai in 2015, the M.Sc. Degree in Information Technology from K.C. College Mumbai, 2017 and currently is working in KC College , Churchgate , Mumbai as Assistant Professor ,IT Department.

Dr. Shashank Joshi received his B.E.degree in Electronics and Telecommunication fromGovt. College of Engineering, Pune in 1988, the M.E. and Ph. D. Degree in Computer Engineering from Bharati Vidyapeeth Deemed University Pune. He is currently working as a Professor in Computer Engineering Department Bharati Vidyapeeth Deemed University College of Engineering, Pune research work, membership, achievements, with photo that will be maximum 200-400 words.

Fourth Author Dr. Suhas Patil . He is currently working as a Professor in Computer Engineering Department Bharati Vidyapeeth Deemed University College ofEngineering, Pune . He is innovative , dynamic teacher devoted to Education and Learning with an experience of over 29 years.