Prediction and Analysis of Sentiments on Twitter Data using Hybrid Naive Bayes Approach

Ch. Srinivasa Rao, G. Satyanarayana Prasad, Vedula Venkateswara Rao

Abstract—every single day, billions of users places their opinions on innumerable aspects of our life and politics with use of micro blogging above the internet. Microbloging websites are prosperous cradles of data for confidence mining and sentiment analysis. In our research work, we emphasis on expending Twitter for sentiment analysis for mining opinions about events, products, individuals and consume it for accepting the current tendencies. Twitter permits its customers a restriction of merely 140 letterings; this limitation powers the customer to be crisp as well as sensitive at the same time. This eventually creates twitter awater of sentiments. Twitter also offers designer responsive streaming. We scurry datasets above5 million tweets by a conventional intended crawler for sentiment analysis tenacity. We suggest a hybrid naive bayes classifier by means of assimilating an English lexical dictionary called SentiWordNet to the prevailing machine learning naïve bayes classifier algorithm. Hybrid naive bayes categorizes the tweets in negative and positive groups independently. Investigational results validated the dominance of hybrid naive bayes on multi sized datasets comprising of assortment of keywords over prevailing methods producing more than 90%correctness in common and 98.29%correctness in the best case. In our research work, we executed through English; nevertheless, the recommended method can be applied with any new language, as long as that language lexicon dictionary.

Index Terms: micro-blog, predictions, sentiment mining, twitter.

I. INTRODUCTION

With the propagation of web to applications like micro-blogging, social networks and forums, there can be assessments, observations, endorsements, rankings and criticisms produced by users. The user created content can be nearly virtually everything comprising people, products, politicians and events, etc. With the eruption of user produced content can be the essential by establishments, representatives, analysts, researches, service providers, social psychologists to mine and analyze the content for different uses. The majority of this user produced content necessary the use of computerized techniques for mining and analyzing.

Revised Manuscript Received on June 05, 2019

Ch Srinivasa Rao, Research Scholar, Dept of CSE Acharya Nagarjuna University, Guntur, Associate Professor, Dept of CS, SVKP & Dr K S RAJU A&Sc College, Penugonda, A.P, India,

Dr. G Satyanarayana Prasad, Professor, Dept of CSE, Dean, Training & Placements, RVR & JC College of Engineering Chowdavaram, Guntur, A.P, India,

Dr. Vedula Venkateswara Rao, Professor, Dept of CSE,Sri Vasavi Engineering College, Pedatadepalli, Tadepalligudem, A.P. India,

Bags of the wholesale user created content that have been premeditated are blogs [1] and product/movie [2] reviews. Micro-bloging has grown into a widespread communiqué tool amongst Internet users. Billions of messages are seeming daily in widely held web-sites that afford services for micro-blogging such as Twitter1, Tumblr 2, Facebook 3. Users of these amenities engrave about their life, stake opinions on assortment of topics and deliberate current concerns. Since the availability of unrestricted layout of messages and a tranquil approachability of micro-blogging platforms, Internet users tend to shift from traditional communication tools to microbloging services. Since large numbers of users place messages related to services and products they consume and express their political and religious views, microbloging web-sites become cherished cradles of people's beliefs and sentiments. Such information can be knowledgeably used for marketing and social studies. Sentiment analysis is a comprehensive research area that has been in study since eras. Its early use was prepared to investigate sentiment grounded on long texts like as emails, letters and emails. It is also deployed in the field of pre and post crime inquiry of criminal activities. Variants of approaches have been applied for the same. Applying this field with the micro-blogging community is an exciting job. This test turns into our motivation. Unnecessary to say we are not the first ones to work in this area. There has been significant investigation in equally machine learning and the lexical methods to sentimental analysis for social networks. We try to improve the existing approaches by diversifying variants to the research. In this paper, we recommend a Hybrid Naive Bayes classifier that is the amalgamation of a machine learning algorithm i.e. Naive Bayes and a special lexical dictionary called SentiWordNet6. We crawled multi size datasets consisting of approximately 4 million tweets. We test the proposed Hybrid Naive Bayes approach by means of Natural language Toolkit and notice that it overtakes the prevailing methods delivering modest results having 98.59 percent accuracy.

II. RELATED WORK

Sentiment analysis trapped responsiveness as one of the greatest vigorous investigation areas with the detonation of social networks. The massive user-produced content resultant from these social mass media confined treasured information in the form of opinions, reviews about events, products and persons. Utmost sentiment analysis revision spractice machine learning methods, which require large amount of user generated content for training. The investigation on sentiment analysis so far has mostly concentrated on 2 possessions: ascertaining whether a specified written entity is subjective or objective, and ascertaining divergence of subjective texts [3].

A. *Sentimental Analysis Approaches*

Sentiment analysis has been accomplished *on* a collection of topics. For instance, there exist sentiment analysis readings for product assessments[4], product assessments [5], and news and blogs ([3], [6]). Sentiment analysis is expressed as a computational linguistics problem.

The taxonomy can be advanced from different viewpoints depending on the nature of the job at hand and standpoint of the person booming out sentiment analysis. The conversant methods are discourse-driven, relationship-driven, language-model-driven, or keyword-driven.

B. *Twitter specific approaches*

The main difference between sentiment analysis of twitter and documents is that, twitter based approaches are more specific towards determining the polarity of words which are adjectives; although the document centered methods are specific towards

job of formative features in the text. There are three foremost methods for twitter precise sentiment analysis.

- Lexical analysis approach
- Machine learning approach
- Hybrid approach
-

Using one or a combination of the different approaches, one can employ one ora combination of lexical and machine learning methods. Specifically, one can use unsupervised methods, supervised techniques or a combination of them.

C. *Lexical Analysis Approach*

Normally exploits a thesaurus or lexicon of pre-tagged words. Every word that is contemporary in a text is matched beside the thesaurus. If a word is existing in the dictionary, then its polarization value is added to the \total polarity score" of the text. Consider an example, if a match has been established with the word \excellent", that is interpreted in the thesaurus as positive, and then the total polarity score of the blog is augmented. If the total polarity score of a text is positive, then that text is categorized as positive, otherwise it is categorized as negative.

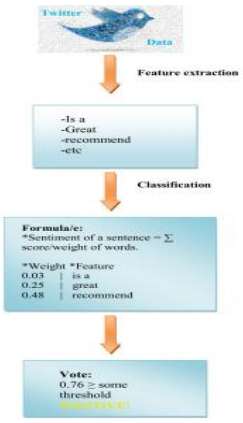

Figure1 Architecture of Lexical Analysis approach

D. *Machine Learning approach*

The focal way of research inside this area has employed supervised machine learning methods. Inside the machine learning approach, a sequence of feature vectors are selected and a group of tagged masses are provided for training a classifier, which can then be pragmatic to an un-tagged amount of text. In case of machine learning method, the selection of features is crucial to the success rate of the classification. Furthermost usually, a selection of unigrams orn-grams from a document in sequential order is preferred as feature vectors. Other projected landscapes consist of the number of positive words, number of negating words, and the length of a document. Support Vector Machines [14] and the Naive Bayes algorithm [29] are the utmost regularly employed classification practices.

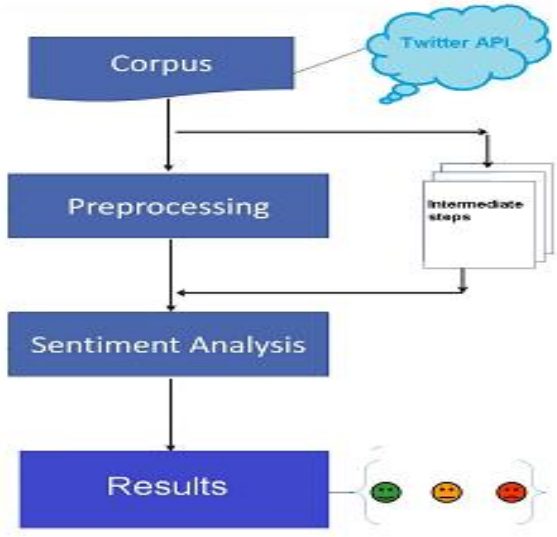

Figure2 Machine learning analysis approach

E. *Hybrid approach*

In Literature there exist specific approaches which use amalgamation of further approaches. One shared approach is followed [17]. They start with two word lexicons and unlabeled data. With the two biased word lexicons i.e. negative and positive, they craft mock documents comprising all the words of the selected lexicon. Afterward, they calculate the cosine correspondence among these mock documents and the un-labeled documents. Based on the cosine correspondence, a document is allocated either positive or negative sentiment. Then they use these to train a Naive Bayes classifier.

III. PROPOSED APPROACH – IMPROVED HYBRID NAÏVE BAYES

The superiority of both lexical approach (for its speed) and machine learning approach (for its accuracy) are not unknown to the world. A lexical approach is fast because of the predefined features (e.g. dictionary) it employs for extracting sentiments. Having a dictionary to refer at runtime reduces the time consumption almost exponentially. To improve performance of lexical approaches the feature set has to be increased drastically, i.e. a very large dictionary of variety of words with their frequencies has to be provided at runtime. This increases the overhead of the system and hence the performance suffers. Thus, there is a constant trade-of between Performance vs Time. On the other hand, machine learning approaches employ recursively learning and tuning of their features, given large input datasets, improve its performance way beyond any lexical approach can achieve. However, due to this

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

runtime performance tuning and learning the system undergoes drastic fall in time constraints.

Our goal is to propose an approach that is a combination of both lexical and machine learning, hence exploit the best features of both in one. For this purpose we choose to employ a Naive Bayes classifier and empower it with an English lexical dictionary SentiWordNet. Our hybrid naive bayes follows the ritual four steps namely: Data collection, Preprocessing, Training the classifier and Classification shown in the figure 3. Through the following sections we shall discuss each step in detail, one at a time. The following figure 3 shows the system architecture of our proposed approach. The labelsPhase I and Phase II shown in figure 3 are the deployment phases.

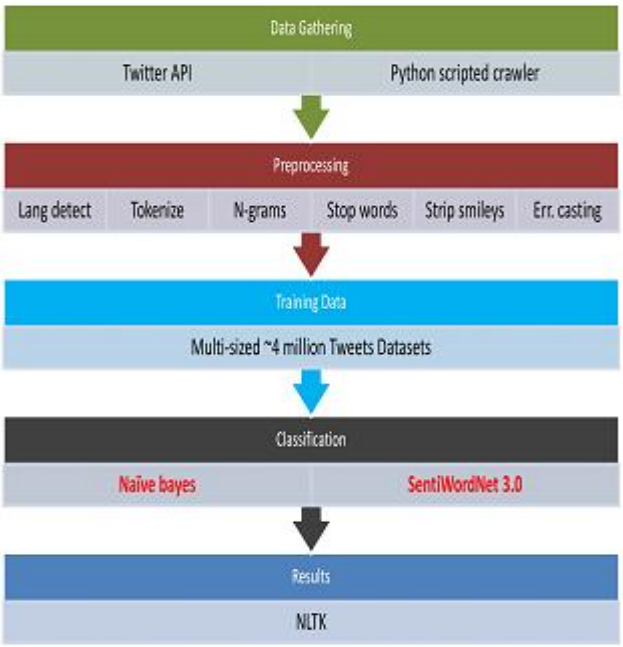

Figure 3 Architecture of Hybrid Naïve Bayes approach

The following Figure shows System architecture for implementation

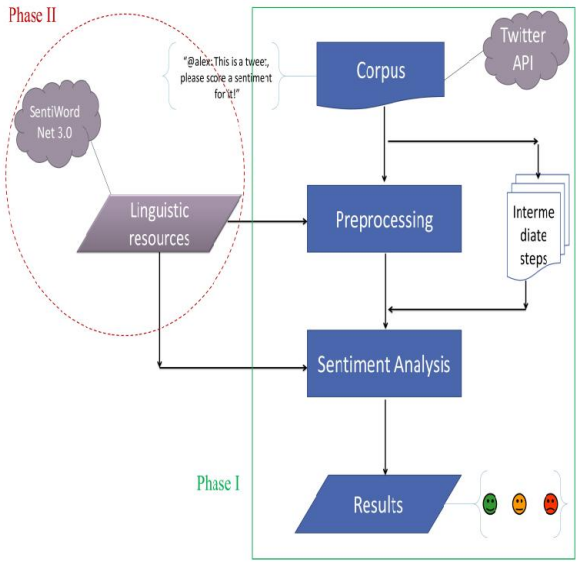

Figure 4 System Architecture of Hybrid Naïve Bayes

A. Data Collection

Twitter data is required for classifier to implement classification and training.. For this tenacity we plan to use API's that twitter provides. Twitter provides two API's; Stream API1 and REST API2. The dissimilarity between Streaming API and REST APIs are: Streaming API provides long lived association and delivers data in nearly real time. The REST APIs provides short lived association and are rate restricted. REST API sallow access to twitter data such as status updates and user info regardless of time. However, Twitter does not make data older than a week or so available. Thus REST entree is limited to data Twittered not before more than a week. The search API offers customers the capability to contact twitter search working process. It uses GET requests and returns results formatted using ATOM or JSON, JSON is recommended due to compactness.

B. Preprocessing

The tweets collected from twitter are a combination of urls, and few non-sentimental data like hash-tags \#", annotation \@" and re tweets \RT". To obtain n-gram landscapes, we first have to tokenize the text input. Tweets posture a delinquent for standard tokenizers aimed for proper and regular text. The below mentioned diagram presents many intermediate processing feature steps. The intermediate steps are the list of features to be taken account of by the classifier.

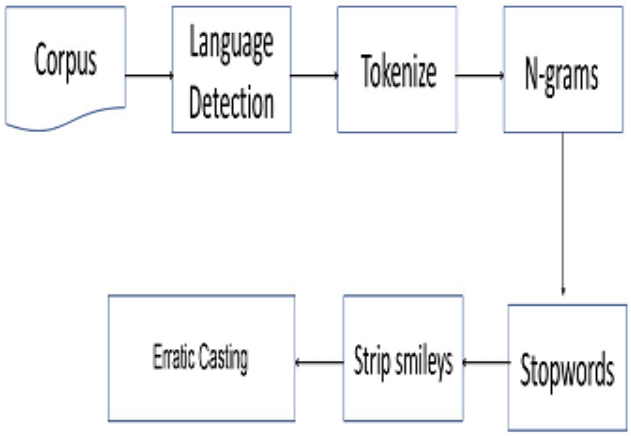

Figure 5 preprocessing steps

C. Sentiment Analysis - The classifier

Naive Bayes was our first choice based on the inference from literature review carried out. Naive bayes is Bayesian probability distribution model based algorithm. In general all Bayesian models are derivatives of the well-known Bayes Rule, which suggests that the probability of a hypothesis given a certain evidence, i.e. the subsequent probability of a proposition, can be acquired in terms of the prior probability of the confirmation, the prior probability of the hypothesis and the conditional probability of the confirmation given the hypothesis. Mathematically,

P(H/E) =P(H) * P(E/H) / P(E)(1)

where, P(H/E)- posterior probability of the hypothesis. P(H)- prior probability of hypothesis.

P(E)- prior probability of evidence. P(E/H)- conditional probability of evidence of given hypothesis. Or in a simpler form:

Posterior =(Prior) *(Likelihood) / Evidence2

To explain the concept, let's take an example. For

instance, we have a new tweet to be classified in to one of the positive or negative classes. Given that in the previously classified tweets, positive tweets are twice the number of negative tweets. Since the new tweet's class is not known, the problem is estimating correctly the class that the tweet is to be categorized in. This can be found out by Bayes rule calculating
the probabilities of the likelihood of the tweet to be positive or negative. Hence, from eq. 1 we have:
P(n/p) =P(n) * P(p/n) / P(p) 3
Since there are twice as many positive tweets as negative, it is reasonable to believe that a new case (which hasn't been observed yet) is twice as likely to have membership positive rather than negative. observed yet) is twice as likely to have membership positive rather than negative. In the Bayesian analysis, this belief is known as the prior probability. Prior probabilities are based on previous experience, in this case the percentage of positive tweets and negative tweets, and often used to predict outcomes before they actually happen. Thus, we can write:
Prior Probability of positive tweetP(p) =No. of positive tweets / Total no. of tweets
Prior Probability of negative tweetP(n) =No. of negative tweets / Total no. of tweets

Let there be say a total of 6k tweets, 4k of which are positive and 2k negative, our prior probabilities for class membership are(where k = 103):
Prior Probability for positive tweet P(p) =4k / 6k
=4 / 6
=2 / 3
Prior Probability for negative tweet P(R)=2k / 6k
=2 / 6
=1 / 3
The likelihood of the tweet falling into either of the classes is equal, since we have only two classes. So likelihood of X = 0.5. So now calculating the posterior probability of the new tweet say X, being positive or negative, will be:
Posterior probability of X being positive = (Prior probability of positive) * (Likelihood of X being positive)
=2 / 3 *1 / 2=1 / 3= 33.34% chances of X being positive.
Posterior probability of X being negative = (Prior probability of negative) * (Likelihood of X being negative)
=1 / 3 * 1 / 2=1 / 6= 16.67% chances of X being negative.
Thus this tweet will fall in to the positive class.
In our case we would have two hypothesis and many other features on basis of which the one that has the highest probability would be chosen as a class of the tweet whose sentiment is being predicted. After every classification step all the probabilities are again calculated and updated accordingly.

IV. EXPERIMENTAL RESULS AND ANALYSIS

To test the proposed approach, we created a setup with the following system requirements. We tested our approach on both Linux and Windows platforms. We used a Dell Optiplex 980 Windows-7 core-i5 (64 bit) machine equipped with 4 GB of RAM and a Linux server system with Quardcore processor equipped with 8 GB of RAM. The tools and technologies used are Python 2.7, NLTK, LMF, Senti Word Net 6.0.
The results of both of the existing classifier and proposed hybrid classifier are presented and compared. Tests were carried out using multiple twitter datasets consisting of a mixture of new and old keywords.
The performance of base naive bayes classifieris shown in the table below with respect to dataset size.

Table1 Performance of hybrid naïve bayes classifier

Data Set Size	Accuracy
1K	28.34
10K	29.37
50K	50.66
100K	59.86
1M	70.22

After integrating SentiWordNet lexicon dictionary, same procedure was carried on the hybrid naive bayes classifier and the following results were harvested.

Table2 Performance of hybrid naïve bayes classifier with SentiWordNet

Data Set Size	Accuracy
1K	63.75
10K	97.3
50K	98.4
100K	95.28
1M	94.3

The following figure shows comparison of hybrid Naïve Bayes classifier performance

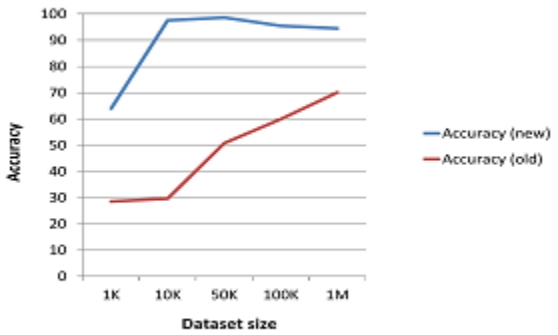

Figure 5 performance of Hybrid Naïve bayes classifier

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

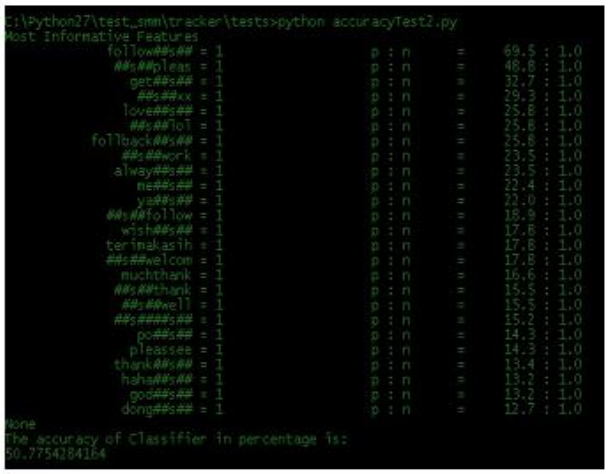

Figure 6 accuracy of hybrid Naïve Bayes classifier

V. CONCLUSION

The natural hybridization of real life inter species, it has conspicuously known that boundaries of both species can be subjugated. Having pragmatic the same assertiveness in our proposal, the investigational studies completed successfully show that hybridizing the existing machine learning analysis and lexical analysis techniques for sentiment classification yield comparatively outperforming accurate results. For all the datasets used, we recorded consistent accuracy of 90%.Clearly from the success of Hybrid Naive Bayes, it can positively be applied over other related sentiment analysis applications like financial sentiment analysis (stock market opinion mining), customer feedback services, and etc. In future we can modify our algorithm to implement Interpreting Sarcasm and multi linguistic support.

REFERENCES

1. J. Serrano-Guerrero, J. A. Olivas, F. P. Romero, and E. Herrera-Viedma, Sentiment analysis: a review and comparative analysis of web services, Information Sciences, 2015, Vol.311, pp.18–38.
2. Walaa Medhat , Ahmed Hassan , Hoda Korashy , Sentiment analysis algorithms and applications: A survey, Ain Shams Engineering Journal (2014) 5, 1093–1113
3. Marouane Birjalia, Abderrahim Beni-Hssanea , Mohammed Errital, Machine Learning and Semantic Sentiment Analysis based Algorithms for Suicide Sentiment Prediction in Social Networks, ScienceDirect Available online at www.sciencedirect.com Procedia Computer Science 113 (2017) 65–72
4. Taochen, Rruifengxu, Yyulanhe, Xxuanwang, Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN, Expert Systems with Applications, Volume 72, 15 April 2017, Pages 221-230
5. Ali Hasan , Sana Moin , Ahmad Karim and Shahaboddin Shamshirband, Machine Learning-Based Sentiment Analysis for Twitter Accounts, Math. Comput. Appl. 2018, 23, 11; doi:10.3390/mca23010011
6. Medhat, W.; Hassan, A.; Korashy, H. Sentiment analysis algorithms and applications: A survey. Ain Shams Eng. J. 2014, 5, 1093–1113. [CrossRef]
7. Sebastiani, F. Machine learning in automated text categorization. ACM Comput.Surv. 2002, 34, 1–47. [CrossRef]
8. Taboada, M.; Brooke, J.; Tofiloski, M.; Voll, K.; Stede, M. Lexicon-based methods for sentiment analysis.Comput.Linguist. 2011, 37, 267–307. [CrossRef]
9. Prabowo, R.; Thelwall, M. Sentiment analysis: A combined approach. J. Informetr. 2009, 3, 143–157. [CrossRef]
10. Dang, Y.; Zhang, Y.; Chen, H. A lexicon-enhanced method for sentiment classification: An experiment on online product reviews. IEEE Intell. Syst. 2010, 25, 46–53. [CrossRef]
11. Cambria, E. Affective computing and sentiment analysis. IEEE Intell. Syst. 2016, 31, 102–107. [CrossRef]
12. Jagdale, O.; Harmalkar, V.; Chavan, S.; Sharma, N. Twitter mining using R. Int. J. Eng. Res. Adv. Tech. 2017, 3, 252–256.
13. Anjaria, M.; Guddeti, R.M.R. Influence factor based opinion mining of twitter data using supervised learning. In Proceedings of the 2014 Sixth International Conference on Communication Systems and Networks (COMSNETS), Bangalore, India, 6–10 January 2014; pp. 1–8.
14. Dubey, G.; Chawla, S.; Kaur, K. Social media opinion analysis for indian political diplomats. In Proceedings of the 2017 7th International Conference on Cloud Computing, Data Science & Engineering, Noida, India, 12–13 January 2017; pp. 681–686.
15. Liu, B.; Hu, M.; Cheng, J. Opinion observer: Analyzing and comparing opinions on the web. In Proceedings of the 14th International Conference on World Wide Web, Chiba, Japan, 10–14 May 2005; pp. 342–351.
16. Razzaq, M.A.; Qamar, A.M.; Bilal, H.S.M. Prediction and analysis of pakistan election 2013 based on sentiment analysis. In Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014), Beijing, China, 17–20 August 2014; pp. 700–703.
17. Soucy, P.; Mineau, G.W.A simple knn algorithm for text categorization. In Proceedings of the 2001 IEEE International Conference on Data Mining, San Jose, CA, USA, 29 November–2 December 2001; pp. 647–648.
18. Lewis, D.D. Naive (bayes) at forty: The independence assumption in information retrieval. In Proceedings of the 10th European Conference on Machine Learning, Chemnitz, Germany, 21–23 April 1998; pp. 4–15.
19. Segnini, A.; Motchoffo, J.J.T. Random Forests and Text Mining. Available online: http://www.academia.edu/11059601/Random_Forest_and_Text_Mining (accessed on 26 February 2018).
20. Raschka, S. Naive bayes and text classification i-introduction and theory.arXiv 20
21. H. Kang, SJ. Yoo and D. Han, " Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews", Expert Systems with Applications, Elsevier, vol. 39, no. 5, pp. 6000—6010, 2012.

AUTHORS PROFILE

Ch Srinivasa Rao is a Research Scholar in the Department of Computer Science & Engineering at Acharya Nagarnuna University, Guntur, A.P, India. He is working as Associate Professor in SVKP & Dr K S Raju A&Sc College, Penugonda, A.P. He received Masters Degree in Computer Applications from Andhra University and Computer Science Engineering from Jawaharlal Nehru Technological University, Kakinada, India. His research interests include Data Mining, Big Data Analytics.

Dr.G Satyanarayana Prasad is Professor in the Department of Computer Science Engineering and Dean, Training & Placements at RVR & JC College of Engineering, Chowdavaram, Guntur, India. He received M.S in Computer Science from A&M University, ALABAMA, USA and PhD from Andhra University, Visakhapatnam, India. His research interests include Image Processing, Data Mining, Big Data Analytics. He guided two research scholars for the award of their PhD. He published books, several papers in International conferences and journals.

Dr.Vedula Venkateswara Rao is Professor in the Department of Computer Science Engineering at Srivasavi Engineering College, tadepalligudem, India. He received Matsers Degree in Computer Science Engineering from JawaharLalNehru Technological University Kakinada, Masters Degree In Information Technology from Punjabi University, Patiayala, India and PhD from Gitam University. His research interests include Cloud Computing and Distributed Systems, Data Mining, Big Data Analytics and Image Processing.He published several papers in International conferences and journals.