Detecting Spam Messages in Twitter Data by Machine learning Algorithms using Cross Validation

K Subba Reddy, E. Srinivasa Reddy

Abstract*— Now a day's human relations are maintained by social media networks. Traditional relationships now days are obsolete. To maintain in association, sharing ideas, exchange knowledge between we use social media networking sites. Social media networking sites like Twitter, Facebook, LinkedIn etc are available in the communication environment. Through Twitter media users share their opinions, interests, knowledge to others by messages. At the same time some of the user's misguide the genuine users. These genuine users are also called solicited users and the users who misguidance are called spammers. These spammers post unwanted information to the non spam users. The non spammers may retweet them to others and they follow the spammers. To avoid this spam messages we propose a methodology by us using machine learning algorithms. To develop our approach used a set of content based features. In spam detection model we used Support vector machine algorithm(SVM) and Naive bayes classification algorithm. To measure the performance of our model we used precision, recall and F measure metrics.

Keywords: Social media, Twitter, Spammer, SVM, Naive bayes.

I. INTRODUCTION

Social media networking sites such as facebook, Twitter etc allow users to share views, photos and videos, posts, and to inform others about online or real-world activities. The success of social networking services can be seen in dominance in today's society with Facebook having a massive 2.13 billion active monthly users and an average of 1.4 billion daily active users in 2017.Some social networking services require members to have a pre-existing connection to contact other members. Twitter is one of the most microblogging services in social media networking site with large number of users who interact to each other through this network. These users share their ideas, opinions, technical knowledge, views and their personal opinions on specific events in the society and are shared through the twitter messages.

Twitter is one of the most prominent social networking applications to form same community groups between technical professionals, friends and family members and business peoples. These community members express their views, news to other members in the community through twitter. Twitter messages are restricted to 280 characters.

Revised Manuscript Received on October 05, 2019.
K Subba Reddy, Research Scholar, Anucet, ANU, Guntur, AP, India.
Dr. E. Srinivasa Reddy, Dean, Anucet, ANU, Guntur, AP, India

Twitter networking site allows the users to follow their favourite scientists, business men and other eminent persons. A user can create his own account in network openly without any restrictions, simply providing his personnel details like name, personnel id and address. Due to this open accessing policy into twitter network many users misuse the network activities. They simply mislead the normal users through retweets, url links and hash tags. Users of the Twitter network have different levels of Knowledge with respect to security threats hidden in social media networks. Spammers are attracted by twitter network as a supporting tool to spread spam messages, advertisements to genuine users. The spammers also send urls and malicious links to the genuine users. Spam is one of the biggest problems in social media sites like faebook, linkedIn. Researchers shows that more than 3% of tweets are spam messages. The trending topics are also attacked by spammers.

To handle attacks from spammers, social media network like Twitter provides different ways to report the spams. A user can report a spam by clicking a link in their home page. The user given reports are analyzed by twitter and the spam accounts are being suspended. Another publically available method is to post a tweet in the '@spam @username' manner. Twitter network also puts its efforts in efficient manner to disclose the malicious tweets and suspicious user accounts. At the time of filtering malicious tweets and suspicious accounts, some of the legitimate user accounts are filtered out by twitter spam detection methods. So we need some of efficient methods to automatically detect spam messages and spammer accounts. At the same time these advanced methodologies are not affect the legitimate user tweets and accounts.

In this work study the classification of spam messages and harm messages. Our aim is to identify helpful features that can be used in machine learning algorithms to classify messages are spam or ham. The major work of this proposal is

- To propose using content based features to detect spam messages.
- To compare the performance of classifiers namely SVM and Naive Bayes classifiers.
- To develop a model to evaluate the spam detection approach on selected features.
- The results show that our spam detection model has good precision, recall and F measure.

In section 2, we gave some discussion about the Twitter network, and discussed related work. In section 3, we discuss the proposed methodology. In section 4, we describe the evaluation and experimental results of our proposed methodology. In section 5, we concluded our approach.

II. RELATED WORK

Twitter is one of the very familiar social networking sites like facebook, Myspace and linkedIn. Twitter networking site is one of the short messaging service where users can send short messages to their friends. A user of the twitter network is identified by username and also identified by actual name. A user can follow another user, consequently that user receives messages on his own page. One user can follow another user who wants desires. Messages can be grouped using hashtags beginning with a "#" special character. Hashtags allow users to search tweets based on interest. When a user likes someone users tweet, he can also retweet that message. Consequently that forwarded message is shown to all his followers. A social media network user can protect his profile from other users. The strength of social media networks are based on network trust. To study spam detection in social networks many researchers proposed various methodologies. De Wang et al [1] proposed a spam detection frame work. In this frame work they define models namely profile model, message model and web page model. They discuss cross social corpora classification and associative classification. In [2] , authors identify and analyze six different features to detect spam. They use these features to distinguish spammers from geniune users. They used simple features like TagSpam, TagBlur, DomFp, NumAds, Plagiarism and ValidLinks. They used SVM and AdaBoost supervised learning algorithms to construct social spam detectors based on these features. Enhua Tan et al. [3] proposed an unsupervised social network spam detection methodology to detect spammers. The proposed UNIK scheme can detect spammers with false positive rate of 0.6% and false negative rate of 3.7%. UNIK scheme has the same level of detection performance as SD2..Xueying Zhang et al [4], the authors propose an extreme learning machine spammer detection methodology. The simulation of this algorithm is carried out in MAT lab environment. The proposed methodology has efficiently found 99.1% spammers and 99.9% non spammers correctly. The authors in [5] present a Markov Clustering based method to identify spam profiles. In this methodology they used facebook profiles and these profiles are modelled as weighted graph. For weight calculation of edge active friends, page links and URL features are used. The authors [6] propose a model to identify spam URLs in social networks with behavioural analysis. They discuss advanced approach instead of traditional process to detect spam URLs. They examine who posted the URLs and who click the URLs. This approach has the 0.86 precision, 0.86 recall. Hailu et al [7] discuss a spam identification approach using facebook and twitter datasets. They used nearly common features between two datasets to detect spam. In this method they combine twitter spam data with facebook data for training and combine facebook spam data with twitter data for training .Random forest classifier shows the best performance compare to other traditional classifiers. In [8] authors discuss one of the supervised machines learning approach to detect spam in social media networks. In this method used content based features and user based features to detect spammers. To implement this approach they used SVM classifier. This methodology has given highest precision, recall and F measure. In [9], the authors employ a model to detect spam using content based features and graph based features in Twitter data. In graph based features they include the number of followers, number of people you are following and ratio between the number of followers over the total sum of the number of followers and the number of people a user is following. The content based features contain content similarity, number of tweets that contain HTTP links, the number of tweets that contain "@" symbols and the number of tweets that contain the hashtag symbol. Using a Bayesian classifier, the result shows that the spam detection system can achieve 89% precision. Igor Santos et al. [10] proposed a spam detection schema to detect spam in twitter data based on content based features. Their approach used compression based classifier to detect spam . In [11], the authors analyze the sentiment data in twitter data. The sentimental analysis include movie reviews, product reviews, spam detection and consumer needs. They use hybrid algorithm that includes Naive bayes classifier algorithm to analyze sentiment. In [12], the authors proposed an hybrid model to detect spammers in twitter data. In this model they used metadata, message based and interaction based features. They used nineteen different features includes newly defined features and redefined features. Zakia Zaman et. al [13] presented to solve the problem of detecting the spam comments posted on social media. In this approach they have been implemented several classification algorithms. They have used 10 fold cross validation to evaluate the performance of algorithms. Out of all classifiers ensemble classifier gives better result.

III. PROPOSED METHODOLOGY

Classification is a methodology used to classify the datasets into categories. In this work, we used different classification approaches to classify social media messages. Fig. 1 presents the brief description of working procedure related to our approach.

1) Collection of sample dataset.
2) Preprocessing the data.
3) Feature selection from data.
4) Constructing the model.
5) Evaluating the model using performance metrics
6) Model testing with cross validation.
7) Comparison of models

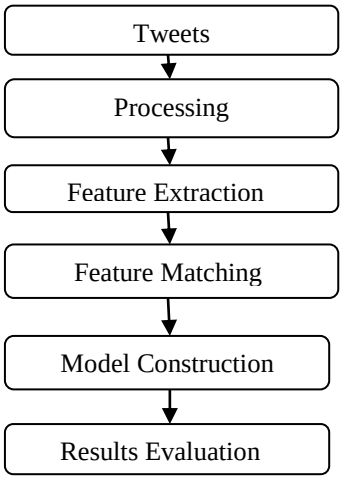

Fig. 1. Framework of Classification

In the evaluation of our proposed model, we use the various Twitter datasets, such as political dataset, entertainment information dataset and sports dataset. Each dataset contains some set of tweets, out of which some tweets are legitimate tweets and remaining are spam tweets. All these dataset tweets are manually collected from various users. Political dataset contains political tweets and retweets posted by various users. We have collected political tweets randomly in Twitter network. Entertainment dataset contain tweets relevant to movies, such tweets are randomly collected from various users for analysis of our model. Sports dataset has the tweets relevant to cricket. We have classified the tweets into spam and ham by manually.

Table 1. Used datasets in our approach

Dataset	Total Samples
Political dataset	2200
Entertainment	2300
Sports	2600

Data preprocessing includes the preparation of data in required format to implement our model. The data should be clean, no noise and consistency. The preprocessing is also used for reducing the large amount of data into useful data format. To analyze the quality of data, data preprocessing step plays major role. In this process metadata details are discarded and consider only text comments of each user. Initially to preprocess the data a set of words was implemented, where a message is represented as the multiset of words. Feature selection for spam detection model is a critical step. The goal of selecting best features is to improve classification efficiency, computational efficiency of the model. The dimensionality reduction of dataset shows little accuracy loss, but overall performance of model is increased [16]. Feature selection is critical process in text data classification due to high dimensionality of text features, irrelevant features and duplicate features. Stop word removal is common feature selection methodology in supervised and unsupervised applications. Stemming is another feature selection methodology in this approach; different forms of the same word are summarized into a single word.

Content based features- In our proposed approach we also used the content based features along with above methodology features. With using only stop word removal and stemming method feature selection, we are unable to detect spam messages in efficient manner. These features characterize the tweets send by the various users to their neighbours. In existing spam detection methodologies, quality of content is taken one of the measures for spam detection in social media. Spammers have various techniques to incorporate their information into social media to mislead traditional spam detection methodologies. Tweet quality is a measure to know the intention of user, based on intention of the user we can check out the tweet as spam or non spam. In our proposed approach we used a set of content based features. Some of the features are defined as:

Mention Ratio: Users of the Twitter social media network can be tagged by "@" symbol followed by twitter handler. Spammers of the network can misuse this feature. The spammers motivate and tempting the benign users to know the sender of the message. The mention ratio for the user is calculated as the ratio between number of mentions in tweets and number of tweets posted by user. Naturally the benign users mention ratio is low for compared to spammers.

Ratio of URL: In Twitter, users generally post their ideas, opinions about a specific topic and share articles through tweets. The tweets include URLs, these refer source pages that contain detailed information. Some of the users include much number of URLs into tweets continuously, so we can suspect them as spammers. The Ratio of URL for user is the ratio of number of URLs used in his tweets to total number of tweets posted by that user. Generally more number of URLs used by spammers in the tweets to share their intention to users. Spammers use the more number of URLs where as legitimate users use the less number of URLs in tweets. The spammers URL ratio is nearer to one or more than one where as for benign users the URL ratio is very small or closer to zero.

Unique mention ratio: Generally benign users contact with friends and colleagues and at the time of sending tweets they can use this group of the people or set of the people regularly but spammers tag the unknown persons randomly within their tweets. Generally the spammers unique mention ratio is very high and low for genuine users.

Unique URL ratio: More number of URLs used by the spammers in the tweets to fulfill their intention but at the same time some of the spammers use the same URL for many number of times for the same user. The genuine user seen the same URL many number of times and tempted then clicked and traverse to malicious site. The unique URL ratio is the ratio of number of unique URLs to number of URLs used in the tweets.

Hashtag and Content similarity: On users wall twitter lists the most frequent hashtags and trending topics. Benign users include these hashtags and trending topics into their malicious tweets to attract the genuine users. The trending hashtags are injected by benign users into the tweets, but there is no relation between hashtags and content.

Hashtag ratio: To group the tweets related to specific topic Hashtagr is used. A group is created by hashtag to discuss specific topic. Top trending hashtags regularly display on user's wall. These trending hashtags are hijacked by

spammers and inject them into their tweets. Whenever genuine users search for these trending hashtags, tweets by the spammers are also shown in the search result.

In our proposed spam detection methodology we used Naive bayes classifier and SVM classifier.

Naive Bayes Classifier : This is one of the efficient classifier and is used to classify the text message as spam message or ham message . This classifier use the probabilistic learning method based on Bayesian theorem. The features of a dataset used in this theorem are mutually independent. This classifier is also used for sentiment analysis, document analysis, spam detection and text data categorization [18], [19]. We have to build a classification model based on content based features in different classes. This classifier is used to classify the tweet based on posterior probability of the tweets belonging to different classes. For a tweet d and a class c Bayes theorem is as

$$P(c|d) = \frac{\mathrm{P(d|c)P(c)}}{P(d)}$$

Support Vector Machine (SVM): This uses statistical theory to classify the dataset samples. This methodology is different from other methods which aim to minimize the risk occurs in other methods. SVM uses Kernel functions which reduces the complexity and computation. This provides efficient classification for problems based on Linear separation and non linear separation. Kernel methods are used for pattern analysis and these methods are enabled to work on high dimensional features. There are three types of kernels in SVM, namely linear kernel, polynomial kernel and Radial basic function. In our proposed model we used linear kernel

In our proposed model, Naive bayes classifier and SVM classifiers are constructed with above defined set of features. Then the classifiers are trained with Twitter dataset. Once the training of the models is completed then the models are tested with test tweets. In the test process of model, the test tweets are given as input to the classifier, then classifier classify the tweet as spam or ham tweet. Due to inefficiency of the classifier some of the tweets are misclassified. Further to improve the classification efficiency, the two classifiers performance is compared.

IV. EXPERIMENTS AND RESULT

In this section we present the experimental details and evaluation results of our proposed model for detecting spam in tweets. Classification, Association and Clustering machine learning algorithms are used for mining the hidden patterns in large amount of data. In our approach we have to build a models for Naive Bayes and Support Vector Machines (SVM) algorithms. In our approach we have measure the performance of Naive Bayes algorithm and Support vector Machine algorithm. Later we compare the some existing methodologies performance with our proposed approach. To evaluate our approach, we used standard metrics called precision, recall and F measure.

a) *Precision*: Precision is the number of True Positives divided by the number of true positives and false positives.

$$Precision = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

Precision is a measure of classifiers exactness. A low precision indicates a large number of false positives.

b) *Recall*: This is the number of true positives divided by the number of true positives and the number of false negatives.

$$Recall = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

Recall is a measure of classifier completeness. A low recall means many false negatives.

c) *F measure*: This metric measure the association between precision and recall.

$$F\ Measure = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

The theme of this work is to group the tweets into two classes such as spam or ham tweets. Here we have implemented Support Vector Machine and Naive Bayes classifiers.10-fold cross validation is used to perform the experiments. To validate the performance of these two classifiers, we used 10 fold cross validation techniques. Cross validation is an important step to evaluate the performance of classifiers when limited numbers of samples are used in dataset. In 10 fold cross validation; the given dataset is divided into 10 equal sized subsets in random fashion. Out of 10 subsets, 9 subsets are used for training the classifiers and remaining one subset is used for testing the trained classifiers. This process is repeated 10 times by taking each of the subset for testing once. All the 10 evaluation results are aggregated to get the final result of classifiers. In this section we evaluate the results obtained when applying SVM and Naive bayes classifiers on datasets without cross validation and with cross validation. In order to show the efficiency of the proposed approaches, here we have conducted experiments using Political dataset, entertainment dataset and sports datasets. In the first experiment, we have performed classification of different dataset samples into spam or ham with SVM classifier. In every dataset, 90% of tweets are used for training the SVM classifier and remaining tweets are used for testing the classifier. Table 2 describes the performance of SVM classifier on different datasets.

Table 2. Performance of SVM classifier on datasets without cross validation.

Dataset	Precision	Recall	F measure
Political tweets	0.93	0.92	0.924
Entertainment	0.93	0.93	0.93
Sports	0.94	0.92	0.929

Next, we have performed classification of different dataset samples into spam or ham with Naive bayes classifier. In every dataset, 90% of tweets are used for training the classifier and remaining tweets are used for testing the classifier. Table 3 describes the performance of Naive bayes classifier on different datasets

Table 3. Performance of Naive bayes classifier on datasets without cross validation.

Dataset	Precision	Recall	F measure

Political tweets	0.92	0.90	0.909
Entertainment	0.93	0.91	0.919
Sports	0.92	0.90	0.909

In second part of our experiment, we have performed classification of different dataset samples into spam or ham with SVM classifier. In this experiment we have used 10 fold cross validation on every dataset to evaluate the performance of classifiers. In 10 fold cross validation every dataset is split into 10 partitions, out of 10 partitions 9 partitions are used for training the model and remaining partition is used for testing the classifier. In this validation of classifier, every partition participated in training and in testing part of the classifier. The above validation process is implemented on all datasets. Table 4 describes the performance of SVM classifier on different datasets

Table 4. Performance of SVM classifier on datasets with cross validation.

Dataset	Precision	Recall	F measure
Political tweets	0.95	0.93	0.939
Entertainment	0.94	0.94	0.940
Sports	0.94	0.93	0.930

Next, we have performed classification of different dataset samples into spam or ham with Naive bayes classifier by cross validation. Table 5 describes the performance of Naive bayes classifier on different datasets.

Table 5. Performance of Naive bayes classifier on datasets with cross validation.

Dataset	Precision	Recall	F measure
Political tweets	0.93	0.91	0.919
Entertainment	0.93	0.92	0.924
Sports	0.93	0.91	0.919

In third part of our experiment, we have compared the performance of SVM and Naive bayes classifiers with cross validation and without cross validation. Table 6 describes the performance of SVM classifier and Naive bayes classifier on different datasets without cross validation and also Table 7 describes the performance of SVM classifier and Naive bayes classifier on different datasets with cross validation.

Table 6. Comparison of SVM and Naive bayes classifiers performance on datasets without cross validation.

Dataset	Precision		Recall		F measure	
	SVM	Naïve	SVM	Naive	SVM	Naive
Political tweets	0.93	0.92	0.92	0.90	0.924	0.909
Entertainment	0.93	0.93	0.93	0.91	0.93	0.919
Sports	0.94	0.92	0.92	0.90	0.929	0.90

Table7. Comparison of SVM and Naive bayes classifiers performance on datasets with cross validation.

Dataset	Precision		Recall		F measure	
	SVM	Naïve	SVM	Naive	SVM	Naive
Political tweets	0.95	0.93	0.93	0.91	0.939	0.919
Entertainment	0.94	0.93	0.94	0.92	0.94	0.924
Sports	0.94	0.93	0.93	0.91	0.934	0.919

We have been observed that, for most of the datasets performance measure is above 92% for the SVM classifier compared to Naive bayes classifier without cross validation. In many cases SVM gives high performance. It has also been observed that SVM classifier has the highest performance measure is above 93% compared to Naive bayes classifier when cross validation is used. In our experiment, we have been observed that SVM classifier has the highest performance compared to Naive bayes classifier when applied cross validation. Here we present a comparative study of our approach with other classifiers studied by [20]. They have studied different spam detection algorithms on Twitter data,. For every classifier the same evaluation metrics are calculated for spam detection. Fig. 6 presents the performance comparison of the proposed SVM classifier approach with other classifiers, It can be observed from the figure that the proposed method outperforms in terms of precision, recall and F measure.

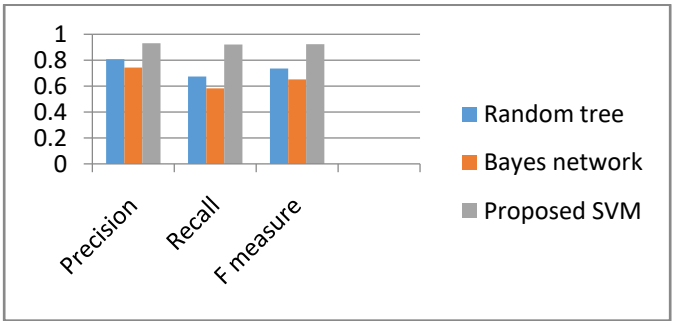

Fig 6. Performance comparison results

V. CONCLUSION

Social media networks are open source communication media to share views and ideas. Due to illegal and suspicious activities of social media users this media is now in bad condition. In this study, we have been working on Twitter messages, in order to find out the spam tweets. Performance measure of various spam detection algorithms are measured, such as Naive bayes and SVM classifiers. Performance of SVM and Naive bayes with different types of datasets are evaluated with cross validation and without cross validation. A comparative study has been made by applying these classifiers on content based features. We have observed that SVM classification model gives better performance compared to Naive bayes classifier in most cases with cross validation. In future work, we want to extend our work using more number of tweets, features and also on other social media datasets like facebook and Youtube comments etc.

REFERENCES

1. De Wang, Danesh Irani and Calton Pu. A social spam detection framework. In proceedings of the Eight annual collaboration,

Electronic messaging, AntiAbuse and Spam Conference (CEAS 2011), 2011.
2. Benjamin Markines, Ciro Cattuto and Filippo Menczer. Social Spam Detection. ACM-2009.
3. Enhua Tan, Lei Guo, SongQing, Xiaodong Zhang and Yihong Zhan. UNIK: Unsuprvised Social Network Spam Detection. ACM-2013.
4. Xueying Zhang, Xianghan Zheng. A Novel Method for Spammer Detection in Social Networks. IEEE-2015.
5. Faraz Ahmed, Muhammad Abulaish, An MCL based approach for spam profile detection in online social networks, 11th international conference on trust, security and privacy in computing and communications, IEEE-2012.
6. cheng cao and James Caverlee, Detecting Spam URLs in Social Media via Behavioral Analysis. In springer 2015
7. Hailu Xu, Weiqing Sun, Ahmad Javaid, Efficient Spam Detection across online social networks, IEEE-2015
8. Xianghan Zheng, Zhipeng Zeng, Zheyi chen, Yuanlong Yu, Chunming Rong, Detecting spammers on social networks, Elsevier-2015
9. Alex Hai Wang, machine learning for the Detection of Spam in Twitter Networks, springer-2012.
10. Igor Santos, Igor Minambres Macrcos, carlos Laorden, patxi Galan Garcia, Aitor Santamaria Ibirika and Pablo Garcia Bringas, international joint conference, advances in intelligent systems and computing, springer-2014
11. Sharvil Shah, K Kumar, Ra.k. Saravanaguru, Sentimental Analysis of Twitter data using classifier algorithms, vol. 6, no. 1, ijece-2016.
12. Mohd Fazil, Muhammad Abulaish, AHybrid approach for Detecting Automated Spammers in Twitter, IEEE-2018
13. Zakia Zaman, Sadia Sharmin, Spam Detection in Social media employing Machine learning Tool for text mining., 13th international conference on signal image technology & internet based systems, IEEE-2017
14. Sumaiya Pathan, R. H. Goudar, detection of spam Messages in social networks Based on SVM., international journal of computer applications, vol 145- No. 10, July 2016.
15. Zahra Mashayekhi, Ali HarounAbadi, A Hybrid approach for Spam Detection Based on Decision tree algorithm and Neural Network., International journal of Mechatronics, Electrical and Computer Technology. Vol. 7- July-17.
16. Rogati, Monica, Yiming Yang. “High performance feature selection for text classification.” Proceedings of the eleventh international conference on information and knowledge management. ACM, 2002.
17. Aggarwal C. Charu, Zhai Chengxiang, “Mining Text Data”. Springer-verlag New York, 2012.
18. Aslandogan, Y. Alip, and Gauri A. Mahajani. “Evidence combination in medical data mining.” Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004. International conference on. Vol. 2.IEEE, 2004.
19. Alizadehsani, Roohallah, et al. “Diagnosis of coronary artery disease using cost sensitive algorithms.” Data mining workshops (ICDMW), 2012 IEEE 12th international conference on IEEE, 2012.
20. Xianchao Zhang, Haijun Bai, Wenxin Liang. “ A social Spam Detection framework via Semi supervised learning”, springer international publishing, pp. 214-226, 2016

BIOGRAPHIES OF AUTHORS

Mr K Subba Reddy is PhD scholar in Department of Computer Science and Engineering, Acharya Nagarjuna University, Guntur, India. He has published papers in international conferences and journals. His area of interest is Big Data, Data mining and machine learning. He may be contacted at: kurapatisr80@gmail.com

Dr E Srinivasa Reddy, PhD., is currently serving as dean in University College of Engineering and Technology and also serving as Head of Department in Computer Science and Engineering, Acharya Nagarjuna University, Guntur, India. He has more than 25 years teaching experience. He is guiding PhD to 8 scholars and 15 has completed his PhD. Dissertations and contributed 5 articles in conferences and 130 papers in Research Journals. His area of interest is Image Processing and Data mining. He may be contacted at: esreddy67@gmail.com