Spam Detection in Online Comments Based on Feature Weight Breakdown

Sherin MariamJohn, K. Kartheeban

Abstract: ***The user reviews posted online by the Internet users about a product plays a vital role in determining its success in the market. The reviews also influence the purchase decision of the consumers. The chances of getting cheated by fake reviews are very high because detecting spams in reviews is not an easy task either manually or automatically. Hence there is a need to evolve new techniques and methods to outperform the smartness of spammers. In this paper, we propose a Heterogeneous Feature Weight Analysis framework for extracting various features related to the review and certain parameters are calculated from these features to form a pattern for deceptive reviews. The features associated with the review are review content, review rating and user centric characteristics which are pulled out from the dataset retrieved from Amazon. This analysis has helped us to categorize reviews into normal and suspicious reviews. We have executed our algorithm in Python software and were able to achieve an accuracy of 71.6% inprediction.***

Keywords: Fake reviews, detect spam, sentiment analysis, feature analysis, online reviews

I. INTRODUCTION

Theusageofsocialmediabythepublicforexpressing their views on a topic or a product in market is on the rise with the increasing number of Internet users worldwide. It has become a practice for online consumers to post their opinions on Twitter, Facebook or blogs and these reviews have become very much important for individuals and marketers. These opinions are helpful for purchase decision making, innovating new products or increasing the demand for these products or services. Unfortunately, this has paved the way for making fake reviews in online media to increase the popularity of products/services or to downgrade the competitor's products/services. Many algorithms have been designed and evaluated for detecting spams in reviews and spammers are also intelligent to outwit these algorithms. It is logically difficult to find out whether a review is genuine or not by simply scanning the content, although there is a pattern of using words for the reviews by the same spammer.

In addition to evaluating the content, we can extract features associated witha review for fake review detection. Some of the prominent features that are extracted are overall rating, reviewer user id, geographic location, the time when the review was posted and the IP address and MAC address of reviewer's computer [16].

Revised Manuscript Received on December 16, 2019.

***** Correspondence Author

Sherin MariamJohn*,Computer Science EngineeringDepartment Kalasalingam Academy of Research and Education Krishnankoil,India sherinmjohn@gmail.com

K. Kartheeban,Computer Science Engineering Department, Kalasalingam Academy of Research and EducationKrishnankoil, India k.kartheeban@klu.ac.in

The online manipulations of reviews are now prevalent across many industries such as movie rating, e- commerce web sites, restaurants and hotels. Many studies on this topic are focused on distinguishing the behavioral pattern of spammers and non- spammers. For example, majority of honest reviewers write only one review for a movie, product/brand, or service. But if we can observe more number of reviews from the same user id during the same time period, that reviews become suspicious. This task of fake review detection is very challenging since spammers would intelligently frame their behaviors like honest people. So we have to look for more features for accuratespamdetection.Inthisresearch,wefoundthatthe spammers tend to write unusually long texts by highlighting each components of the product in review, while normal reviewers mention relatively one or two functions of the product in shortertext.

The features can be classified into three main categories as follows:

Features based on Review Content: The examples under this class are count of brand names mentioned in review, length of the review and the number of feedbacks which were helpful.

Features based on Reviewer: Examples include the averageratinggiven,negativeoutliersaremostlikelytobe spam, the mean and standard deviation in rating, ratio of first hand reviews to the total number of reviews to that product/service and the number of cases the reviewer was the only person to review the product orservice.

Features based on Product: Examples for product centric features are price of the product, sales volume, sales rank assigned based on sales volume and the mean andstanddeviationofreviewratingoftheproduct.The products with low sales rank are more likely to be spammed [16], [20].

The genuine reviewers generally give a mix of emotions both good and bad about different features in the product. This kind of information is very important for consumers as well as marketers to know better about the product. The fraudsters tend to use highly positive words in order to promote the business. So extracting thefeatures is very vital in sentiment analysis and it was found that lexicon based method is best suited for feature extraction [8].

To evaluate our proposed algorithm, we have downloaded the publicly available reviews dataset from Amazon. The dataset contains around 2 lakhs reviews but our computer processor cannot handle such volume of big data and hence we have reduced our data size to 25,000 records. The Python interpreter version used is 3.6.7.

II. RELATED WORK

The work done by Kumar [1] and his team of researchers found that the features describing the reviews of spammers are heavily skewed. This means that the spammers have a tendency to distort the natural distribution of opinions. They have employed feature engineering to detect and predict opinion spammers using supervised learning algorithms. The various features taken for study are review gap, review count, rating entropy, rating deviation, time of review and user tenure. Ridhima Ghai et al. [2] assigned scores to each feature like caps score, reviewer count and rating variation. Reviews were termed as helpful if total score value is high and reviews corresponding to lower score values are termed as not helpful and probably spams. Yelp.com provides a review dataset which contains information on review related features,userrelatedfeaturesandbusinessrelatedfeatures. The Yelp dataset structure is explained well in the paper and the authors have proposed an algorithm based on location and businesses[5].

M.Saini [21] points out that feature centric opinion spam detection algorithms face two main issues. The first problem is identifying and constructing new features and second issue is evaluating these features effectively. The feature ranking is based on degree of dependency and concluded that multi-view learning is more powerful than single-view learning. Another work concentrates on singleton review spam detection by transforming the problem into a temporal pattern discovery problem [18]. The fake reviews can be divided into four types viz. untruthful review, off-topic review, brand only review and non-review [3]. Non-reviews contain a lot of advertisements, link to other websites, phone numbers, price etc. It contains practically no helpful opinion words.

Spammers can work in groups which are more harmful than individual spammer and thirteen features pertaining to group spammers and the methods of detection are defined in the paper by Rupesh et al. [4]. Another paper [14] constructed co-bursting network for group spammer detection. Spammers have adopted different ways to escape detection methods and a detailed account on compromised or fake user identities to spread spam is taken for survey in the paper [15]. Lin et al. [17] experimentally proved that unsupervised method of spam detection yielded more accuracy than supervised methods like Logistic Regression and SVM. Guan Wang et al. [19] proposes review graph for defining the relationship between stores, reviewers and reviews. From the graph, they have derived three conclusions like faithfulness of reviewers, honesty in reviews and trustworthiness of stores. Rohit et al. [13] considered sentiment score also along with three sets of features to train the classifiers and got improved accuracy.

The main evaluation criteria used in almost all works are Area Under Curve (AUC), Precision and accuracy metrics. Mostafa Salehi et al. [11] have differentiated review and user based features into behavioural and linguistic centred features. The network spam detection algorithm is evaluated on these parameters. Linguistic and psycholinguistic differences of honest and fake reviews have a crucial influence in spam detection [10].

In many papers, under supervised learning methods, SVM proved to be the best classifier. But in the analysis work done by Atefeh Heydari et al. [6] found that neural network model performance under deep learning is better than SVM. The more important features, whether it is content based or user based, should be given more weight in calculations for better performance [7]. Another work done by Somayeh Shojaee et al. [9], the writing style of a writer is deeply taken for study to identify fake reviews with various classifiers like SMO, SVM and Naive Bayes. Another method of feature extraction is TFIDF of word n grams. The selection of appropriate feature set is essential for proper training of classifiers [12].

III. METHODOLOGY

The dataset had the following features which were taken for analysis and detection of spam. The features were Reviewer Identity number, Product Identity number, an array which gives the rating values of helpful or non-helpful reviews, overall star rating, review content, date and time of review. The framework of methods carried out for determining suspicious reviews is given in figure 1.

The features from the dataset are extracted one by one for analysis. The rating feature is categorized into the number of reviews given a star rating of 5, 4, 3, 2 and 1 respectively. The average rating is calculated and the variation of each review rating with average rating is worked out. The values are normalized between 1 and 0. These values are pictorially represented for clarity in a chart. The number of reviews in which the usages of capital letters are greater is also found out programmatically. This is in accordance with the theory that spammers tend to use more capital letters in review content. A sparse matrix was formed with rows depicting reviewer Id, columns representing product id and cell values denoting the presence of reviewer rating against a product. This was done to check for a reviewer giving more than one review for the same product. The weights are normalized between 1 and 0 and assigned for each feature appropriately.

$$\text{Overallscore forareview} = \sum fw_i / n, \quad (1)$$

where i varies from 1 to n, n=total number offeatures, fw denotes featureweight.The threshold value is taken as 0.75 and all values above 0.75 were taken as normal helpful reviews and less than threshold value were termed as suspicious reviews. The programmatically predicted values were compared with actual helpful reviews specified in dataset tocompute the accuracy ofprediction.

Accuracy = predicted values of helpful reviews (2)
actual helpful reviews in the dataset

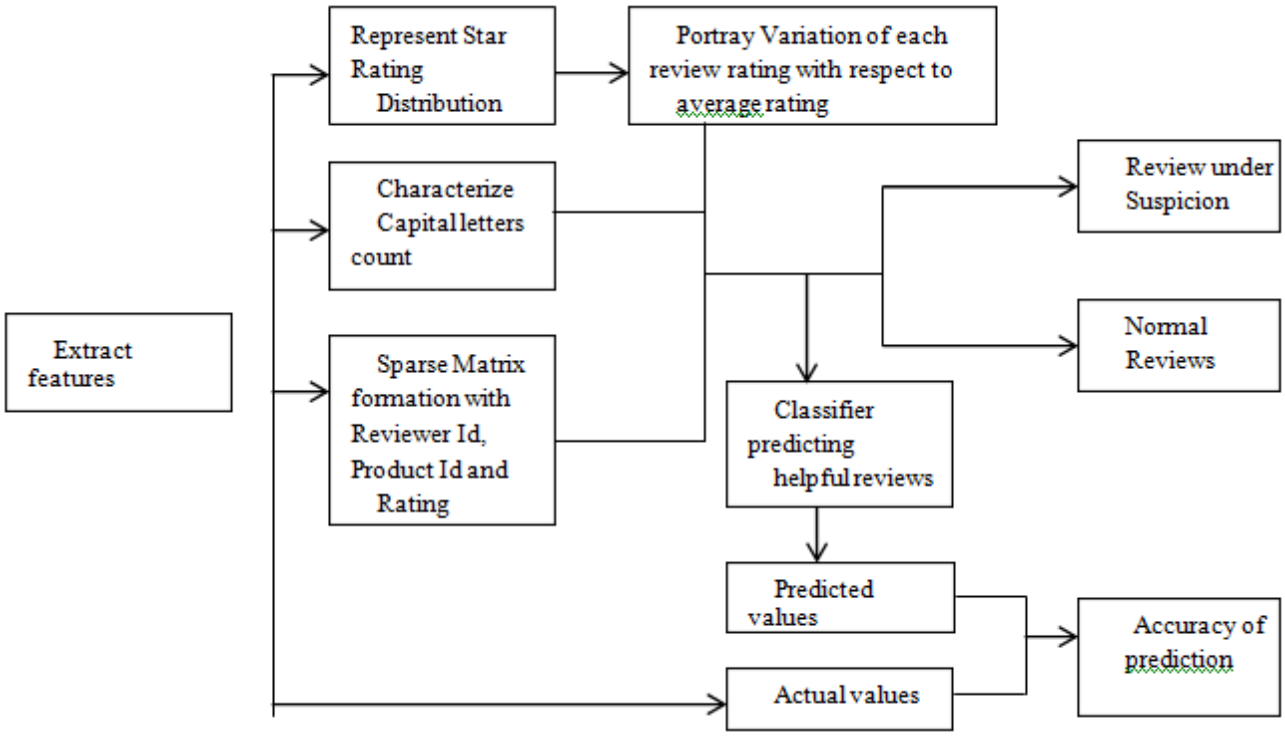

Fig. 1. Hetrogeneous Feature Analysis Framework

IV. RESULTS ANDDISCUSSION

Itiswidelyacceptedthatthereviewratinghasadirect link to the revenue of the product. Hence the spammers tend to give the highest value for rating. When we characterized the distribution of review rating, the graph shown in figure 4 is skewed to the highest value of 5. But we cannot come to a conclusion with this single feature analysis result. So we have calculated the average review rating and found out the variation of each review with the average value and it is presented in figure 2. The logic applied is that the smaller the value, the higher the deviation. But it is found that only a small percentage of reviews (9.6%) have the value of 0.25 and hence we understood that a small percentage of reviews only come under suspicion. A value of 0.5 means normal deviation and a value of 1 means no deviation at all. It is clear from the graph that the number of reviews with 0.25 is less than that of 0.5 and 1. The capital count graph given in figure 3 shows that only a very few reviews are in full capital letters since a value of 1 shows that the count of capital letters is zero or less than 2. The sparse matrix was also analyzed and we found only a few rows with high value. Overall, we can conclude that the dataset contained a maximum of normal or genuine reviews and a small percentage of reviews only came under suspicion. The weight assigned to each feature was normalized inthe range 0 to 1. All these weights were summed up and compared with an optimal threshold value of 0.75 which we have got from previous studies. All those records having score above 0.75 were marked as helpful and those below 0.75 were termed as non-helpful reviews. The null values were discarded from calculation.

Table I. Variation In Review Rating.

Rating Variation (RV)	**Normalized value**	**Count of Reviews**	**%**
RV=0	1	7928	31.7
RV<2 and RV>=1	0.5	14673	58.7
RV>=2	0.25	2399	9.6
Total Reviews		25000	

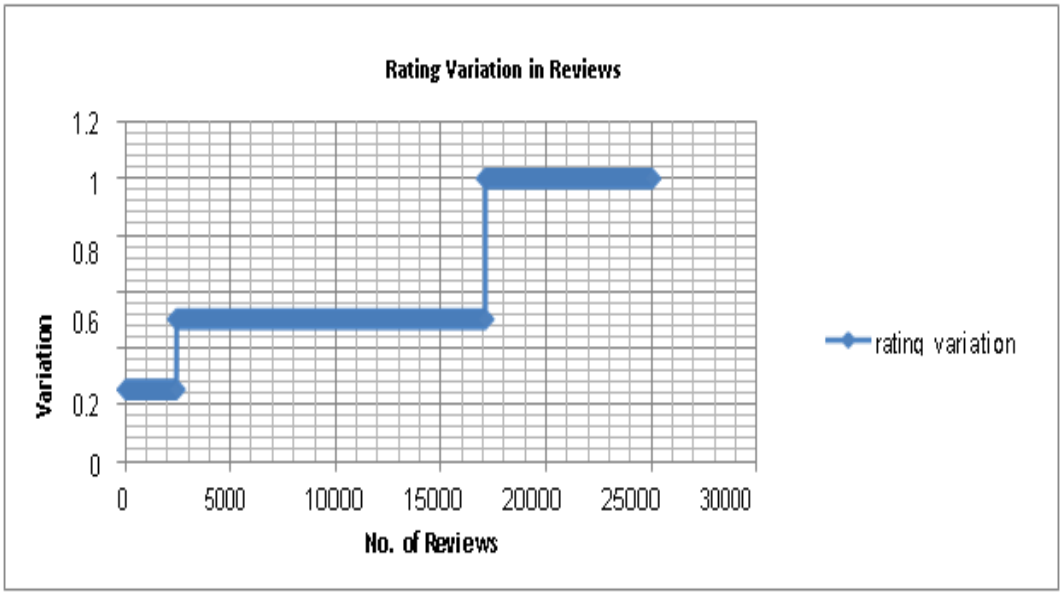

Fig. 2. shows the variation in star rating with respect to the average value. 1 means no deviation and a value of 0.5 means normal deviation. 0.25 shows the outlier values which are under suspicion.

V. CONCLUSION

The null values in the dataset mean that the usershave not given the answer to the feedback question of whether the review was helpful or non-helpful. Hence the null columns were ignored forcalculation.

Table Ii. Prediction Of Helpful Reviews.

Predicting Helpful reviews			
	Total No. Reviews	25000	
	Threshold value	0.75	
	Helpful	Not helpful	Null values
Actual Values in Dataset	3340	3851	17809
Predicted Values	2394	4797	17809
Accuracy	71.6 %		

earchfocusedondetectingfakereviewsfroma dataset which contained random reviews labeled with helpful or non-helpful feedback from Amazon website. The features were extracted and assigned normalized weights and the total score was computed and compared against an optimal threshold value for categorization into helpful normal reviews and suspicious spam reviews. The research was based on previous study on the topic and requires improvement in terms of accuracy of prediction. Our limitation was the capacity of processor power for handlinglarge records of data. Our future endeavors are set for improving accuracy of prediction by incorporating temporal features and linguisticfeatures.

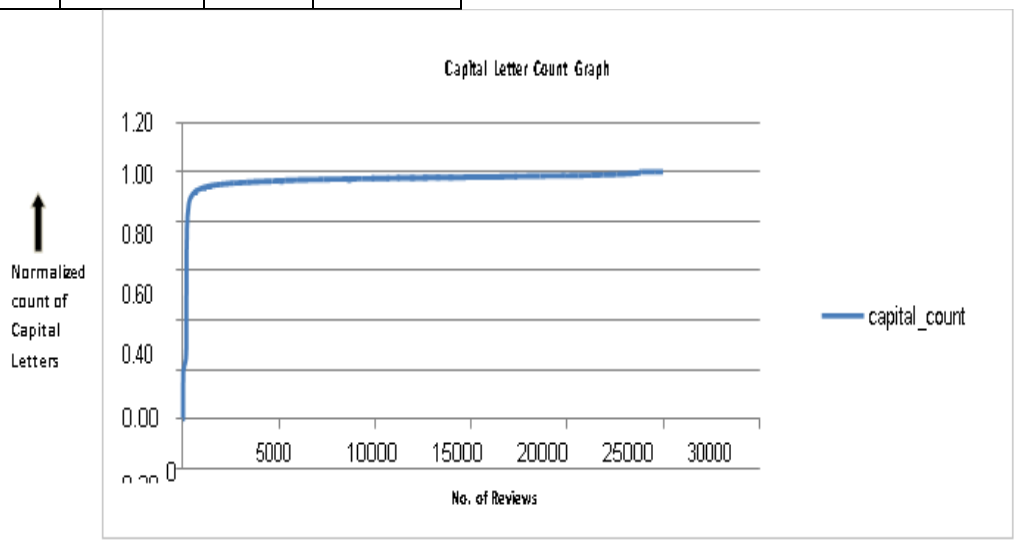

Fig. 3. shows the normalized value of count of capital letters in each review. Almost all reviews are touching the value 1 which means the count of capital letters is 0 or less than 2.

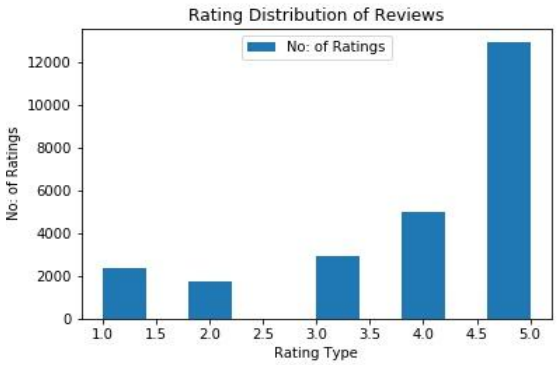

Fig. 4. shows the normalized distribution of Rating in Reviews. The number of reviews with rating 5 is the highest which means majority of reviewers have given a high rating of value 5.

REFERENCES

1. Naveen Kumar, Deepak Venugopal, Liangfei Qiu & Subodha Kumar, Detecting Review Manipulation on Online Platforms with Hierarchical Supervised Learning, Journal of Management Information Systems, Volume 35, Issue 1, pp 350-380,2018
2. Ghai R., Kumar S., Pandey A.C, Spam Detection Using Rating and Review Processing Method, Smart Innovations in Communication and Computational Sciences, Advances in Intelligent Systems and Computing, Springer, vol 670, pp 189-198,2019
3. Hoang Long, Nguyen, Opinion Spam Recognition Method For OnlineReviewsUsingOntologicalFeatures,JournalofScience, Volume 61, pp. 44-59, 2014
4. Dewang, R.K. & Singh, A.K., State-of-art approaches for review spammer detection: a survey, Journal of Intelligent Information Systems, Volume 50, Issue 2, pp 231–264,2018
5. Amani Karumanchi, Lixin Fu, Jing Deng, Prediction of Review Sentiment and Detection of Fake Reviews in Social Media, Int'l Conf. Information and Knowledge Engineering, pp 181 – 186, 2018
6. Chih-Chien Wang, Min-Yuh Day, Detecting Spamming Reviews Using Long Short-term Memory Recurrent Neural Network Framework, Proceedings of the 2nd International Conference on E-commerce, E-Business and E-Government, Association for Computing Machinery, pp 16–20,2018
7. Arpita kunne, Roopalakshmi, Spam Reviews Detection Framework Based on Heterogeneous Information Network (HIN), Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies, IEEE Xplore, pp 1791 – 1795, 2018
8. Nagwa M. K. Saeed,Nivin A. Helal , The Impact of Spam Reviews on Feature-based Sentiment Analysis, IEEE explore, pp 633–639,2018
9. Somayeh Shojaee, Masrah Azrifah Azmi Murad, Detecting Deceptive Reviews Using Lexical and Syntactic Features , IEEE, pp 53-58,2013
10. Atefeh Heydaria,Mohammad ali Tavakoli, Detection of review spam: A survey,Expert Systems with Applications,Volume42,Issue 7, Elseveir, pp 3634-3642, 2015
11. Saeedreza Shehnepoor, Mostafa Salehi, NetSpam: a Network- based Spam Detection Framework for Reviews in Online Social Media, IEEE Transactions on Information Forensics and Security, Volume 12 , Issue 7, pp 1585 - 1595,2017
12. Muhammad Hassan Arif, Jianxin Li, Sentiment analysis and spam detection in short informal text using learning classifier systems, Soft Computing, Springer, Volume 22, Issue 21, pp 7281–7291, 2018
13. Rohit Narayan, Jitendra Kumar Rout, Sanjay Kumar Jena, Review Spam Detection Using Opinion Mining, Progress in Intelligent Computing Techniques: Theory, Practice, andApplications,
14. Springer, Volume 719, pp 273-279, 2018
15. Huayi Li, Geli Fei, Shuai Wang, Bimodal Distribution and Co-Bursting in Review Spam Detection, Proceedings of the 26th International Conference on World Wide Web, ACM DL, pp 1063-1072 ,2017
16. Ravneet Kaur, Sarbjeet Singh, Harish Kumar, Rise of Spam and Compromised Accounts in Online Social Networks: A State-of- the-Art Review of Different Combating Approaches, Journal of Network and Computer Applications, Elseveir, Volume 112, pp 53-88,2018
17. Bing Liu, Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers,2012
18. Yuming Lin, Tao Zhu, Towards Online Review Spam Detection, Proceedings of the 23rd International Conference on World Wide Web, ACM DL, pp 341-342, 2014
19. Sihong Xie, Guan Wang, Review Spam Detection via Temporal Pattern Discovery, Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and datamining,
20. ACM DL, pp 823-831, 2012
21. Guan Wang, Sihong Xie, Review Graph based Online Store Review Spammer Detection, IEEE International Conference on Data Mining, pp. 1242–1247,2011
22. Shwet Mani, Sneha Kumari, Spam Review Detection Using Ensemble Machine Learning, Springer, Volume 10935, pp 198–209,2018
23. Mayank Saini, Sharad Verma, Aditi Sharan, Multi-view Ensemble Learning Using Rough Set Based Feature Ranking for Opinion Spam Detection, Advances in Computer Communication and Computational Sciences , Springer, Volume 759, pp 3-12,2018

AUTHORS PROFILE

Sherin Mariam John is currently engaged as a PhD scholar in the Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Krishnankoil, Tamilnadu, India. She has completed her masters in Computer Science Engineering from Anna University,Chennai, India. She has 7 years of Industry experience and 12 years of academic experience. Her area of interest is in data mining and text mining and has published articles in various International Journals.

Dr. K. Kartheeban working as Associate professor, Department of computer Science and Engineering, Kalasalingam Academy of Research and Education, Krishnankoil, Tamilnadu, India. Also he was worked as a Deputy Director Academic. He received his M.E degree in computer science and Engineering from College of Engineering, Anna University, and Chennai in 2007 and Completed PhD in the title of "Development of Efficient Algorithms for Secure Communication in Distributed Computing Environment" in the Department of Computer Science and Engineering at Kalasalingam Academy of Research and Education, Anandnagar, and TAMILNADU, INDIA in 2014. He worked as a faculty from Adhiyamaan College of Engineering, Hosur, India between 1996-1998. Currently 1 scholar completed his PhD and 6 students are doing their PhD under him with the topics such as Internet of Things, Medical Image Processing, Video Analytics, Cyber Forensics , Sentimental Analysis and Scheduling in Cloud computing. He has published many papers in SCI journals and Scopus indexed conferences. Also he has submitted proposals toDeIT, SERB and DRDO. His areas of interests include IoT, Medical image processing, cryptography and bioinformatics and grid and cloud computing.