International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6, August, 2019

Sentiment Analysis of Movies on Social Media using R Studio

Jaichandran R, Bagath Basha C, Shunmuganathan K.L, Rajaprakash S, Kanagasuba Raja S

Abstract*: This paper presents sentiment analysis of twitter data on movies using R-studio. Twitter is one of the largest social media that shares user opinion about a thing or event that happens all around the world. Recently social media analysis gained importance in digital marketing. User tweets about a product or event, person, movie, etc., are analyzed to know market trends and customer feedback. In this paper, first we have performed literature study on various methods used in twitter data analysis. Second, we have discussed about the steps involved in accessing twitter data. Finally, we have performed sentiment analysis on tweeter data for the movies titled kabali, Bharath Ane Nenu Mersal, and Dangal. User data for the movies are classified into positive, neutral and negative based on DBM and SVM. Sentiment scores are used as evaluation metrics. Results shows DBM is effective in classifying sentiments and produced better sentiment scores compared to SVM. Results are helpful in identifying popularity of the movies and audience feedback about the movies.

Keywords: Big Data Analytics, R Studio, Twitter, Sentiment Analysis.

I. INTRODUCTION

Today social media is one of the fastest medium that shares people opinion about an event or activity that happens all around the world. For example twitter, face book, linkedIn, youtube etc., are used by various people including celebrities, sport stars, politicians, actors, etc., all around the world. People post their opinion and share their views through social media1-3. This valuable information's are used for prediction, product review, popularity analysis, feedback, sentiment analysis, etc. Sentiment analysis in social media is a challenging task due to its raw unstructured nature. Peoples use different slangs, abbreviations and misspells.

In this paper, first we have performed literature study on various methods that are used in the analysis of twitter data. Second, we have discussed about the steps involved in accessing twitter data. Finally, we have performed sentiment analysis on tweets data for the movies titled kabali, Bharath Ane Nenu Mersal, and Dangal. User tweets for the movies are classified into positive, neutral and negative based on the keywords in Dictionary based Method (DBM). Sentiment scores for the movies are calculated and visualized. Results are helpful in identifying popularity of the movies and audience feedback about the movies. Performance of DBM is compared with Support Vector Machine (SVM) using sentiment score as evaluation metrics. Results show DBM performances better than SVM.

II. LITERATURE STUDY

Literature study on various sentiment analysis methods are discussed as follows. Arvind Singh Raghuwanshi et al., (2017) performed sentiment analysis in twitter data using three methods: SVM, naïve bayes, and logistic regression method. In SVM method, sentiments are classified into different classes and texts in twitted sentences are mapped to a particular class. In logistic regression method the sentiment classification are binary in nature. Naïve Bayesian method classifies sentiments based on prediction. Results showed SVM method performed in sentiment classification compared to other two methods. Ali Hasan et al, (2018) performed comparative study on three sentiment analyzers such as SentiWordNet, Word Sequence Disambiguation (WSD) and TextBlob for calculating sentiment scores. The results obtained are tested using two machine learning algorithms: Naïve Bayes, and SVM. Waikato Environment for Knowledge Analysis (Weka) is used for validation. Results showed WSD produced better results in SVM and TextBlob is produced relatively better results with Naïve Bayes. Ankita Gupta et al., (2017) proposed a hybrid approach for sentiment analysis using three stage models: preprocessing, feature generation, and classification. Stage one is preprocessing, which involves spell correction, stop words removal, stemming, etc. Step two is feature generation, which uses list of adjectives for sentiment scoring. In this stage, training data sets and testing data sets are transformed into list of adjectives.

Revised Manuscript Received on October 30, 2019.
* Correspondence Author
Jaichandran R*, Department of CSE, Aarupadai Veedu Institute of Technology, Vinayaka Mission's Research Foundation (Deemed to be University), Rajiv Gandhi Salai, Old Mamallapuram Road, Paiyanoor 603104, Kanchipuram (DT), Tamil Nadu,
C. Bagath Basha, Department of Computer Science and Engineering, Vinayaka Mission's Research Foundation, Salem, Tamil Nadu, India.
Shunmuganathan K.L, principal, Aarupadai Veedu Institute of Technology, Vinayaka Mission's Research Foundation (Deemed to be University), Rajiv Gandhi Salai, Old Mamallapuram Road, Paiyanoor 603104, Kanchipuram (DT), Tamil Nadu, India,
Rajaprakash, Department of CSE, Aarupadai Veedu Institute of Technology, Vinayaka Mission's Research Foundation (Deemed to be University), Rajiv Gandhi Salai, Old Mamallapuram Road, Paiyanoor 603104, Kanchipuram (DT), Tamil Nadu, India,
Kanagasuba Raja S, Department of IT, SRM Eswari Engineering College, Chenaai, Tamil Nadu, India.

Step three is classification, in this stage probabilistic predictive decision is applied for SVM. Authors compared the performance of the proposed hybrid method with two machine learning algorithms K-Nearest Neighbors (KNN) and SVM. Accuracy and f-measure is used as metrics in evaluation. Results showed the proposed hybrid method performed better than other two methods. Vishal A.Kharde et al., (2016) done a performance study on following sentiment analysis method: Machine learning method, Lexical based method, cross lingual method, Cross domain method. Accuracy, precision, recall, and F1 are used as metrics in evaluating the performance the above methods. Where Accuracy = ((True Positive + True Negative) / (True Positive + True Negative + False Positive + False Negative)). Precision = (True Positive / (True Positive + False Positive)). Recall = (True Positive / (True Positive + False Negative)). F1 = (2* Precision*Recall) /(Precision + Recall). Tree machine learning algorithms: SVM, Co-Training SVM, and Deep Learning are tested for accuracy. Results show SVM is accurate than Co-Training SVM, and Deep Learning method. Two lexical based methods: corpus and Dictionary methods are tested for accuracy. Results show corpus method is better than dictionary method. Four cross-lingual methods: Ensemble, co-Train, EWGA, and CLMM are tested for accuracy. Results show EWGA is better than Ensemble, co-Train, and CLMM. Three cross-domain methods: Active Learning, Thesaurus, and SFA are tested for accuracy. Results show all three methods produced similar results. Overall results shows machine learning method SVM has highest accuracy.

N.M.Dhanya et al., (2018) performed sentiment analysis on the implementation of demonetization policy in India. Authors performed sentiment analysis on five thousand tweets using machine algorithms such as SVM, Naïve Bayes Classifier, and Decision tree. Similarly Arun K, et al., (2017) also performed sentiment analysis on demonetization policy in india. In addition authors also performed sentiment analysis on digital payments, operation clean money, and income tax payments. Sentiment analysis results are demonstrated using word cloud, bar charts and pie charts.

Pappu Rajan A, et al., (2014) carried out sentiment analysis on consumer service provided by five companies: Airtel, TCS, Titan, Colgate, and Bosch. Users text crawled from twitter are compared with list of positive words and negative words. The frequency of positive or negative words used by a user in the twitted text is considered in scoring the sentiments. Authors used very positive, positive, slightly positive, neutral, slightly negative, negative, and very negative as 7 point scale in scoring the sentiments.

Smailovic J et al., (2013) predicted stack market precise in advance using sentiment analysis in twitter data. In this work, user twits about a company and their products are categorized into three sentiments (positive, negative and neutral) using SVM method. Finally authors predicted the stock market price in advance using Granger causality test in sentiment scores. Rupawari Jadhav et al., (2017) also predicted stock market price in advance using sentiment scores. Authors discussed about various techniques used in sentiment scoring and stock market prediction. The sentiment scoring methods discussed are Naïve Bayes, SVM, Maximum Entrophy, and Random Forest. Stock market prediction methods discussed includes: random walk, moving average, regression method, and Auto Regressive Integrated Moving Average (ARIMA) model.

Christian Nwankwo et al., (2017) predicted movie rating using sentiment analysis of twitter data. Authors analyzed the opinion expressed by public about the movies The Magnificent Seven, Sully, Strokes, Masterminds, and Deepwater Horizon. Authors ranked the movies based on the sentiment score. Abhishek Kesharwani et al., (2017) also ranked movies based on sentiment score on twitter data. Authors analyzed the public opinion of about the nine movies (Ae Dil hai mushkil, Shivaay, Force 2, Pink, Raees, Kaabil,Dangal, Kahaani, Bahubali) in twitter and compared it with other public voting sites like Internet Movie Database (IMDb) and Rotten Tomatoes. Madan A., et al, (2018) performed sentiment analysis of Goods and Service Tax (GST) in india using lexicon based approach in Twitter data. In this paper author's categorized public opinion on GST in india into positive, neutral and neutral sentiments. Ten thousand data sets are used for evaluation. Results show public have positive opinion on GST with thirty three percentage positive sentiments, thirteen percentage negative sentiments, and fifty four percentage neutral sentiments. Similarly jayamalini K et al., (2019) used dictionary based approach to perform sentiment analysis of GST in india.

Prabhsimran Singh., et al (2018) performed state wise sentiment analysis on demonetization of five hundred rupees and thousand rupees by government of india. Sentiments of peoples in thirthy states in india is categorized as very happy, happy, neutral, sad, very sad, and no data.Out of thirthy states in india, in seven states majority of people are very happy, in five states majority of people are happy,in two states majority of people are neutral,in four states majority of people are sad, in five states majority of people are very sad, in seven states no data about demonetization.

III. METHODOLOGY

Figure 1 illustrates the methodology for twitter data analysis. Steps in the methodology are summarized as follows.

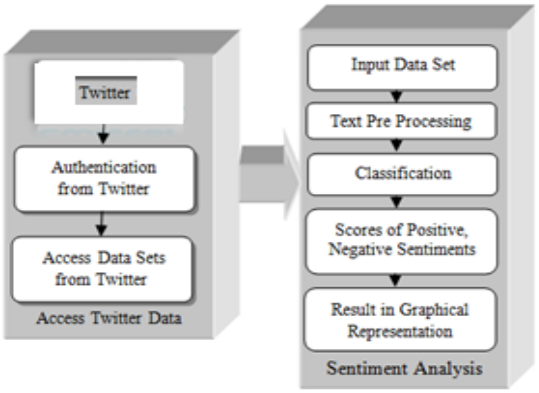

Fig. 1 Methodology

A. Access Twitter Data

It involves two steps described as follows.

Twitter Authentication: It is the first step in mining any data from Twitter. Twitter provides a Twitter Application Program Interface (API) to authenticate user using twitter application. User has to create an application to connect with Twitter and the data can be streamed. Twitter API provides consumer_key, consumer_secret, access_token, and access_secret which can be used to establish the connection with Twitter and R-studio. The above data are unique to the user and it is passed as a argument in setup_twitter_oauth() function.setup_twitter_oauth(consumer_key,consumer_secret,access_token,access_secret) Three-way handshake is accomplished between user and twitter. Now any data from the twitter can be streamed.

Accessing Datasets from Twitter: Twitter provides 6 digit pin for authentication. This pin is generated for every transaction between twitter and R-studio application. Using this pin, twitter hash tags can be fetched and ready to mining the knowledge from the data. In this regard, we use searchTwitter () function to access the data and stored in the text format which can be shown as -searchTwitter('GST',n=max,lang = 'en')

B. Sentiment Analysis

Steps involved in sentiment analysis of data sets of Twitter.

Input Data Set: The data set accessed from twitter need to be given as input to data analysis software for sentiment analysis. Here R studio is used access and analyze data sets from twitter. Apache Hadoop, NodeXL are the other tools that can be used to access and analyze twitter data.

Text Pre Processing: The data collected from the twitter are need to be filtered by removing unnecessary details of hash tag which is called as noise reduction from tweets such as punctuation, HTML tag, white space, numbers etc. It can be done step by step as follows:

- **Replacing Emoticons:** An emotion is a pictorial representation of facial expression characters. Since, Twitter allows only limited numbers of words so user used emoticons to express the feeling and emotions. In this process, all the emoticons are replaced by the respective words in the tweets.
- **URL and Hashtags:** The tweets generally contain hashtag to highlight the words. User used this because of limitation of words in the post. It must be processed to analyze the tweets in more appropriate manner.
- **Lowercase:** Tweets contain both the cases of letters that is uppercase and lowercase. It gives the uneven meaning of the words such as TwITtEr DaTA. It is desirable case during mining of tweets so it must be converted into lowercase to reduce the anomaly.
- **Tokenization:** It is needed in feature selection of the sentence. So all the tweets is converted into small tokens which is further processed by sentiment analyzer to rate the score of tweets.
- **Stemming:** It removes the prefix of the words containing etc. It is further reduces the complexity of analyzer to analyze in a proficient manner.
- **Stop Words:** Stop words are the words which contains no meaning in the tweets such as like, is, are, the etc related to conjunction and preposition. So it is necessary to remove the stop words from the tweets to reduce complexity of the sentence and analyze in smooth manner.

C. Classification

In DBM, people sentiments about a topic in social media can to be classified into 'positive', 'negative', 'neutral' based on set of key words. For example following polarity of words is used in classification of sentiments. Positive = {'good', 'comfortable', 'nice', 'interesting', 'awesome' 'amaze' etc}, negative = {'not', 'no', 'bad', 'poor', 'never' etc}. But SVM is a binary classifier that classifies sentiments into positives and negatives. R studio uses dictionary based approach and SVM for sentiment classification.

D. Scores of Positive and Negative Sentiments

Sentiment scores of a posted message are calculated based on number of positive words and negative words. R studio calculates sentiment score based on Breen's approach given as follows:

Sentiment Score = (Total number of positive words) – (Total number of negative words). If Sentiment score > 0 then the posted message has positive opinion else negative opinion.

E. Graphical Representation

The result of Sentient Analyzer can be visualized using various packages of R-studio such as word clouds, ggplot2, bar Chart, histogram etc. These tools can be used to visualize the sentiment score of a message posted in Twitter.

IV. EXPERIMENTS AND RESULTS

Experiments are conducted by extracting user twittes on following movies: Kabali, Bharath Ane Nenu, Mersal, and Dangal. Number of tweets used for analysis is one thousand. Table 1 shows sample tweets for the movies under negative, neutral and positive categories. Table 2 shows sentimental score for five movies titled kabali, Bharath Ane Nenu Mersal, and Dangal using DBM and SVM. In DBM out of thousand tweets kabala movie has 519 positive tweets, 456 neutral tweets and 25 negative tweets. Bharath Ane Nenu movie has 783 positive tweets, 216 neutral tweets and 10 negative tweets. Mersal movie has 323 positive tweets, 665 neutral tweets and 12 negative tweets. Dangal movie has 480 positive tweets, 452 neutral tweets and 28 negative tweets. In SVM out of thousand tweets kabala movie has 720 positive tweets, 0 neutral tweets and 280 negative tweets. Bharath Ane Nenu movie has 750 positive tweets, 0 neutral tweets and 250 negative tweets. Mersal movie has 650 positive tweets, 0 neutral tweets and 350 negative tweets. Dangal movie has 600 positive tweets, 0 neutral tweets and 200 negative tweets.

Table 1: Sample Tweets for Movies

Movie Name	Sentiment Category		
	Negative	Neutral	Positive
Kabali	RT @prjaishankar: #SemmaWeightu after #kabali SaNa didn t give any hits but @beemji still believed him but he disappointed him.	#SemmaWeightu the jazz feel like veera thurandara from kabali and the rap style Chennai vadachannai from Madras	@farhaz98 @Rajini_Gowtham1 @rajinifans @KABALI__FC @naveenrajini Yes true.. a very nice movie. Great concept.
Bharath Ane Nenu	@Advani_Kiara: Okay guys got to go now!! Im sorry I couldnt answer everyones questions but we chat again super soon.. lots of love	When the #BharathAneNenu Team is officially going to release 200crs gross poster https://t.co/b7OhPlqMwd	RT @hiphoptamizha: @sivakoratala amazing movie sir. #BharathAneNenu is a must watch irrespective of which state u r in. Superb political movie
Mersal	RT @YogiBabu_offl: So much problems for #Mersal and its not for the movie Its for the one man #VijayThadaigalaiudaithu sarithiram padaip	RT @TMoffcial: Director #Atlee receives the Mass director of the year Award for #Mersal in #galattanaksatra award Thanks for the precious movie	RT @otvfofficial: #mersal was the most toughest film ,among the other films I did... and #mersal is the film I loved n enjoyed the most too
Dangal	Damn it #bollywood! Damn you @aamir_khan!!! #Dangal #DangalMovie is melting my heart! I miss my dad	RT @BollyNumbers: #Dangal has Collected $374,035 in South Korea till Tuesday. In Japan it Grossed more then $500k.	@iamdwarf_ @PrinceRKA631314 @prince8715108 @Prashant_RKF @itsjat32 Yaah dangal has made record for leggiest mo movie in china

Table 2: Sentiment Score for Movies using DBM & SVM

Method	Movie Name	Number of Tweets			Sentiment Score
		Negative	Neutral	Positive	
DBM	Kabali	25	456	519	494
SVM		280	0	720	440
DBM	Bharath Ane Nenu	10	207	783	773
SVM		250	0	750	500
DBM	Mersal	12	665	323	311
SVM		350	0	650	300
DBM	Dangal	28	452	480	452
SVM		200	0	600	400

Performance of DBM is compared with SVM in classification of sentiments using sentiment score as evaluation metrics. Results in table 2 shows DBM produces better sentiment score compared to SVM. SVM is better in classifying positives sentiments compared to DBM. But the DBM if effective in classifying negative and neutral sentiments compared to SVM. Figure 2 is based on sentiment scores in table 2. Figure 2 shows sentiment scores of the movies using DBM and SVM. Using DBM, Bharath Ane Nenu movie has highest sentiment core of 773, followed by Kabali with 494, Dangal with 452 and Mersal with 311. Similarly using SVM, Bharath Ane Nenu movie has highest sentiment core of 500, followed by Kabali with 440, Dangal with 400 and Mersal with 300. Results show sentiment scores for all five movies in DBM is better than SVM.

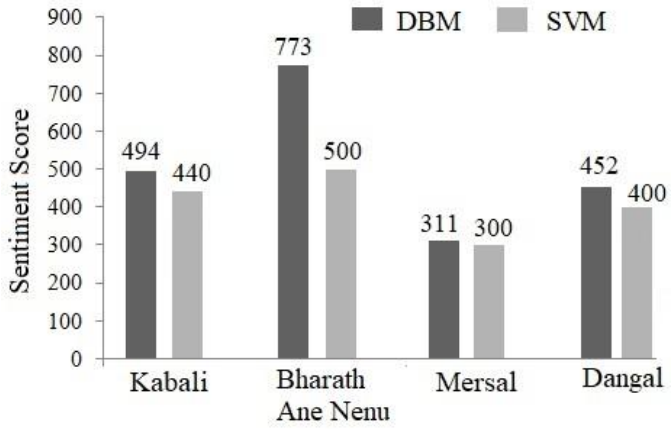

Fig. 2. Sentiment Score of the Movies

V. CONCLUSION

Sentiment analysis on twitter data for the movies titled kabali, Bharath Ane Nenu Mersal, and Dangal are performed using DBM and SVM. The user tweets for the movies are classified into negative, neutral and positive sentiments. Sentiment scores are used as evaluation metric in identifying effective method to classifying sentiments. Results shows DBM is effective in classifying sentiments and produced better sentiment scores compared to SVM.

REFERENCES

1. Abhishek Kesharwani, and Rakesh Bharti, "Movie Rating Prediction Based on Twitter Sentiment Analysis," Journal of Advanced Computing and Communication Techniques, vol. 5, Issue 1, pp. 6-10, 2017.
2. Rupawari Jadhav, and M.S. Wakode, "Survey: Sentiment Analysis of Twitter Data for Stock Market Prediction," International Journal of Advanced Research in Computer Communication Engineering, vol. 6, Issue 3, pp. 558-562, 2017.
3. Arun K, Srinagesh A, Ramesh M, "Twitter Sentiment Analysis on Demonetization Tweets in India using R Language," International Journal of Computer Engineering in Research Trends, vol. 4, Issue 6, pp. 252-258, 2017.
4. Arvind Singh Raghuwanshi, Satish Kumar Pawar, "Polarity Classification of Twitter Data Using Sentiment Analysis," International Journal on Recent and Innovation Trends in Computing and Communication, vol. 5, Issue 6, pp.434-439, 2017.
5. Ali Hasan, Sana Moin, Ahmad Karim, and Shahaboddin Shamshirband, "Machine Learning-Based Sentiment Analysis for Twitter Accounts," Mathematical and Computational Applications, vol. 23, Issue 11, pp. 1-15, 2018.
6. Ankit Gupta, Jyotika Pruthi, and Neha Sahu, "Sentiment Analysis of Tweets using Machine Learning Approach," International Journal of Computer Science and Mobile Computing, vol. 6, Issue. 4, pg. 444-458, 2017.
7. Vishal A.Kharde, Sonawane S.S, "Sentiment Analysis of Twitter Data: A Survey of Techniques," International Journal of Computer Applications, vol. 139, No. 11, pp. 5-15, 2016.
8. N.M.Dhanya and Harish, U.C., "Sentiment Analysis of Twitter Data on Demonetization using Machine Learning Techniques," Lecture Notes in Computational Vision and Biomechanics, vol. 28, pp. 227-237,2018.
9. Pappu Rajan A, and Victor S.P, "Web Sentiment Analysis for Scoring Positive or Negative Words using Tweeter Data," International Journal of Computer Applications, vol. 96, no. 6, pp. 33-37, 2014.

10. Smailović J., Grčar M., Lavrač N., Žnidaršič M. , "Predictive Sentiment Analysis of Tweets: A Stock Market Application," In: Holzinger A., Pasi G. (eds) Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data. Lecture Notes in Computer Science, vol. 7947, 2013.
11. Christian Nwankwo, Hayden Wimmer, Jie Du, "Predicting Movie Rating using Sentiment Analysis of Tweets," Midwest DSI Annual Conference Grand Rapids, Michigan, pp. 68-79, 2017.
12. Madan A., Arora R., Roy N.R, "Sentiment Analysis of Indians on GST," In: Panda B., Sharma S., Roy N. (eds) Data Science and Analytics. REDSET 2017. Communications in Computer and Information Science, vol 799. Springer, Singapore, 2018.
13. Jayamalini K., and Ponnavaikko M, "Social Media Mining: Analysis of Twitter Data to Find Opinions about GST," Journal of Engineering and Applied Sciences, vol. 14, Issue 12, 4167-4175,2019.
14. Prabhsimran Singh, Ravinder Singh Sawhney, Karanjeet Singh Kahlon, "Sentiment analysis of demonetization of 500 & 1000 rupee banknotes by Indian government," ICT Express, vol. 4, Issue 3, Pages 124-129, September 2018.

AUTHORS PROFILE

Dr.R.Jaichandran is currently working as Head of the department CSE, Aarupadai Veedu Institute of Technology an ambit institution of Vinayaka Missions Research Foundation (Deemed to be University), Tamil Nadu, India. He has 13 years of experience in academics, industry, research, anddevelopment activities. Published 33 research papers in referred Journals and Conferences. His area of Interest includes Wireless Sensor Networks, Internet of Things (IoT), Ethical Hacking, Big data Analytics, and Embedded systems.

C. Bagath Basha is having teaching experience about 6 years and 6 months. He served in various positions in Teaching. He is currently doing as Research Scholar, Department of Computer Science and Engineering, Vinayaka Mission's Research Foundation, Salem, Tamil Nadu, India. His area of interest includes Big Data and Data Analytics, Security.

Dr. K. L. SHUNMUGANATHAN is principal of AVIT & illustrious Professor in the Department of Computer Science and Engineering, AVIT. He has a vast experience in teaching, research & administration for more than 29 years at different levels. He earned his Ph.D from Anna University Chennai,in the area of AI& Networks, Master Engineering degree in of Computer Science and Engineering, Madurai Kamaraj University,and Bachelors degree in of Computer Science and Engineering from Bharathidasan University, Trichy. He has published over 147 papers in refereed International Journals and presented more than 42 papers in International and National conferences in India .

Dr.S.Rajaprakashis M.sc, M.Phil M.E Ph.D. currently working as Associate professor of CSE in Aarupadai Veedu Institute of Technology an ambit institution of Vinayaka Missions Research Foundation (Deemed to be University), Tamil Nadu, India. He has 17 years of experience in academics, research, and development activities. Published 17 research papers in referred Journals and Conferences. His area of Interest Artificial Intelligence, Computational Intelligence, Discrete Mathematics and Automata theory. .Received grants from Tamil Nadu State Council for Science and Technology .He has peer Reviewed Manuscripts in reputed international Journals and Conferences. He is a member in following professional societies: CSI and ISTE and Ramajunam Mathematicsl Society**.**

Dr. S.Kanaga Suba Raja, is working as an Associate Professor in Easwari Engineering College. He obtained his Ph.D degree in Computer Science in 2013 from Manonmaniam Sundaranar University, Tirunelveli. He has about 14 years of teaching and research experience. He has successfully guided many graduate and under-graduate students for their research projects. He has several papers published in International journals and conferences to his credit. His current research interests are Wireless Body Area Networks, Ad hoc and Sensor Networks and Mobile and Ubiquitous Computing**..**