International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-4, February 2020

An Optimal Aggregation of Product Data using Vector Space Model

Susmitha Gunti, Krishnamoorthi Makkithaya, Deepthi S.

Abstract: ***In the current scenario there exists many versions of a particular product. A product might be laptop, mobile or any other gadget. With increase in number of versions there is a need to analyze the reason for release of the new version of the product. This can be done by the study of reviews and ratings provided by consumers. To get a more accurate output we first relate the rating and review usingSentiment Analysis (SA). SA is a form of text mining that helps us to understand the attitude and behavior of a customer towards a product/service. The ratings given by the customer may not be in the same level of agreement as in the review text. Customer may have issues with the product and has explained in the review but can be generous and give decent rating, such circumstances often depends on the emotional quotient of the customer. Therefore, there is a need for a system which can elicit the polarity among the reviews and check if there is proper agreement between the ratings and reviews given by user till the product become obsolete.In order to provide the correlation between the ratings and reviews lexicon method of sentiment analysis is used to generate the sentiment score for each review. Based on the sentiment score obtained the reviews are further classified into extreme negative, negative, neutral, positive, and extreme positive and compared to the ratings given by the customer. With this reviews as input, feature selection is done using vector space model. The output obtained depicts the success factors and failures of a product which helps to build a better version.***

Keywords: Sentiment Analysis, Text mining, Lexicon approach, Feature selection, Vector space model

I. INTRODUCTION

Text mining is a method of eliciting the content based on context and meaning of text. It gathers information from unstructured and amorphous text and extracts the non-trivial patterns [1]. It is used in social media data analytics, spam filtering, customer care service etc. Though, the results produced by text mining are partial it is riveting in modern culture. As text is the easiest and common method for savingthe information,text mining embodiesgreater potential in the field of data mining. Text mining is used by organizations to fetch the answers for business questions and individualsto sensethefacts from data and draw the conclusions. Text mining can be done using entity extraction, categorization and sentiment analysis.

Revised Manuscript Received on February 28, 2020.
***** Correspondence Author
Susmitha Gunti*, Department of Computer Science and Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Karnataka, India.
Krishnamoorthi Makkithaya, Professor, Department of Computer Science and Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Karnataka, India.
Deepthi S, Assistant Professor, Department of Computer Science and Engineering at Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Karnataka, India.

Sentiment Analysis is one the popular and useful applications of text mining.
The primary goal of sentiment analysis comprises data analysis ontext for understanding the opinion and other key factors like modality and mood. It describes the opinion that comes from a review, rating, or comment. With the help of opinion rich resources such as online review sites opinion mining and sentiment analysis is carried out to find out what people think [2]. These opinions can be referred to as Positive, Negative, Neutral or have no sentiment at all. For example: “It is a very good product” is a positive text. Sentiment Analysis broadcasts the voice of customers through social media which is used predominantly by business organizations to understand customer views on their product or brand. Is someone for or against a product? Do they think a service was useful or not? Did they like or dislike something?
Sentiment Analysis is usually done in three methods or models. Document level, Sentence level and Aspect level.
Document Level: At document level the main task is to determine sentiment assuming that each document expresses opinion on a single entity [3].
Sentence Level: In sentence level sentiment analysis is performed sentence wise rather than whole document. In this levelthesubjective sentences expressing factual information and subjective sentences expressing opinions are distinguished [4].
Aspect Level: Aspect level is also called as feature level analysis and gives more precise results using natural language processing. The opinions are identified by polarity and a target of opinion [5].

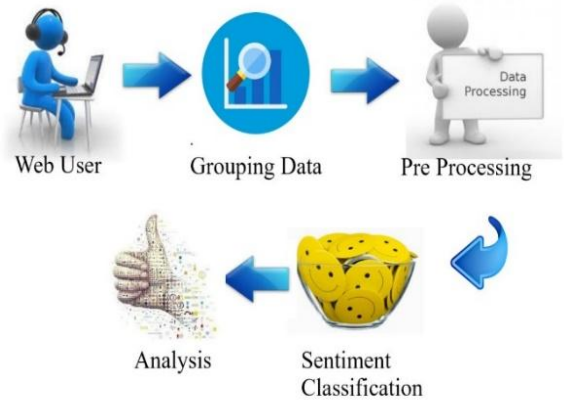

Fig. 1. Process of Sentiment Analysis

The Fig. 1 provides an overview on how sentiment analysis is carried out. SA starts with collection of data given by users in the form of reviews, discussions, and comments in social media. Since the data collected is unstructured, pre-processing is needed to clean the dataset by reducing the complexity and make the data structured.

This is an important task to prepare the data noise free for sentiment classification. The various phases in classification include removal of stop wordsand punctuations, stemming, tokenization etc. From the structured data sentiment analysis can be done using any one of the methods or classifiers. Classification of text is done using lexicon-based approach or machine leaning techniques shown in Fig. 2.

Lexicon or rule based sentiment analysis [6] is based on sentiment lexicons, which is a collection of precompiled terms. Thisapproach uses a sentiment dictionary with opinion words to map them with the data to deduce positive or negative sentiments, whereas machine learning techniques uses the linguistic approachfor analysis. This approach uses classifiers like Naive Bayes,Maximum entropy and Support vector machines. The results are analysed according to the need.

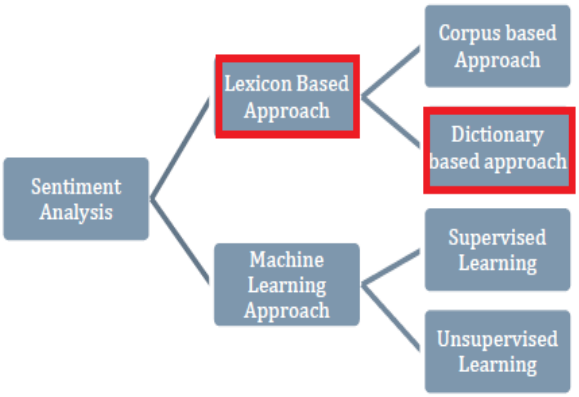

Fig. 2. Classification methods of Sentiment Analysis

Feature selection is also called as variable selection or attribute selection. In feature selection the data available which is redundant or irrelevant can be removed. It is automatic selection of attributes using efficient algorithms for enhanced results. The simplest algorithm is to minimize the error rate by checkingall the possible subset of features. The evaluation metrics are highly effected by these algorithms.

Vector space model or term vector model is an algebraic model for representing and filtering the text documents. It is an information retrieval method that captures the importance of a word in the document which is helpful and easier way of feature extraction.

II. LITERATURE REVIEW

Customers have a lot of impact on what other people think. A survey conducted on sentiment analysis methods and approaches [7] shows how opinion mining plays a crucial role in business decision making.Many of the enterprises take decisions and many users buy products based on reviews provided.

Taking this as an advantage manipulation of reviews [8] is done which can be detected by examining the writing style of the review.There are different approaches of conducting sentiment analysis. Machine learning algorithms are used to automatically detect the pattern existing in the data. Dictionary based method is the easiest method to perform SA using WordNet and SentiwordNet dictionaries.

Ontology based method is used to capture the structure of specific domain. Division of tasks into subtasks [9] are evaluated in four ways namely addressing the data, exploiting the dataset, selecting the features, approaches employed. Using machine learning algorithms classification of reviews has been done based on form, function and behaviour [10]

To give the brief outcome from a review which helps customers analyse faster, extraction of a specific feature from the entire review been done using different text summarization techniques [11].

A lexicon based approach on twitter data has been carried out [12] using bag of words and a sentiment dictionary with heuristic approach for pre-processing[13]. To achieve more accuracy while performing sentiment analysis using lexicon approach, a new way of polarity assignment has been done using recall, precision and f-measure [14].

Smart phonesplay a crucial role in day to day life. To analyse the brand and model of the mobile [15],customer reviews play a major role. They change and influence the buying patterns of customers. Websites like Amazon,Flipkart,and Snapdeal etc. provide a platformto consumers to provide theirexperiences. To extract useful data from large datasets,data is classifiedinto positive and negative sentiments.

Classification was performed using Naive Bayes model, Support Vector Machines and Decision tree models to measure the accuracy.An algorithm to mine the reviews into positive and negative has been developed. This algorithm was used to provide a comparison between two mobiles in terms of six features namely camera, battery, screen, sounds, and design [16] using MATLAB.

To provide continuous updates on customer requirements Kano Analysis [17] has been performed. Kano analysis evaluates customer requirements using subjective and analytical approach. Analytic approach deals with keyword occurrence, whereassubjective approach involvesexpert judgement.

Feature selection has diverse effect in deciding a product in terms of quality and functioning. Feature selection reduces the computation time and increases the prediction performance and better understanding of the data [18].

The three main classes of feature selection are wrappers, filters and embedded methods [19]. Wrapper method is a predictive method which uses every new subset to train the model.

These are computationally intensive and provide best feature set. Filter method uses a proxy method to find the feature subset. Filters are less computationally intensive compared to wrappers but useful for finding the relation between features. Embedded methods give intermediate performance in terms of computational complexity as it catches all group of techniques as part of model selection process.

Vector space model is a way of information retrieval which represents notations and definitions necessary to identify the concepts and relationships [20].

III. METHODOLOGY

A. Senti-Score Calculation

The Fig. 3 gives an idea about the steps involved in calculating the correlation between ratings and reviews. This can be explained in four categories: Data collection, Text cleaning, Tokenization, and Senti-score calculation.

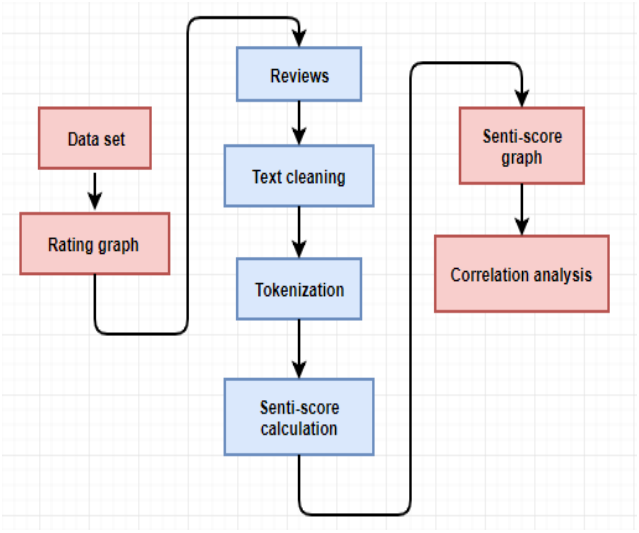

Fig. 3. Design flow

1) **Data Collection**

Sentiment analysis can be carried out on any product reviews. The product chosen here is mobiles which is a trending gadget which everyone holds. Data is collected through online websites such as Amazon, Flipkart etc. The main attributes of the data set include Name of the product, Rating, Review, and Date on which the rating and review was given.

2) Text Cleaning

Data collected from web is unstructured hence needs pre-processing which includes text cleaning. Text cleaning is removing unwanted data. It includes removal of numbers, punctuations, stop words, and white spaces.

3) Tokenization

Tokenization is the process of breaking the sentence into words (tokens) understandable by the machine. Tokenization also involves stemming, which removes the endings and returns the base or dictionary form of the word.

4) Senti-score calculation

Data which is clean and in the form of tokens is called as structured data which is used for sentiment analysis. The software is provided with two bag of words containing positive words and negative words. This approach of lexicon based sentiment analysis is referred as dictionary based lexicon approach. Each review is analyzed by counting the positive words (PW) and negatives words (NW). By using the below mentioned formula senti-score is calculated.

$$Senti - Score = Number\ of\ PW - Number\ of\ NW$$

TABLE I. Example of reviews classified based on senti-score obtained

Rating	Review	Positive word count	Negative word count	Senti-Score	Category
1	Don't buy, waste of money, poor camera, heating problem, backup is poor, too bad	1	7	-6	Extreme Negative (<-5)
1	Bad display break. Display worth Rs 8000	1	2	-1	Negative (-1 to -5)
5	Awesome but battery problem	1	1	0	Neutral (0)
1	Feature is good but mobile is heating	2	1	1	Positive (1 to 5)
5	Very good model phone, works very very good, camera is good both front and rear, I am very well satisfied with the mobile	9	1	8	Extreme positive (>5)

B. Vector Space Model

Feature extraction is done using Vector Space Model involving steps shown in Fig 4.The reviews obtained after preprocessing are used to calculate TFIDF (Term Frequency & Inverse Document Frequency). TFIDF is calculation of Term frequency and Document frequency of a word i.e., number of occurrences of a word in a review and occurrence of a word in all the reviews respectively.

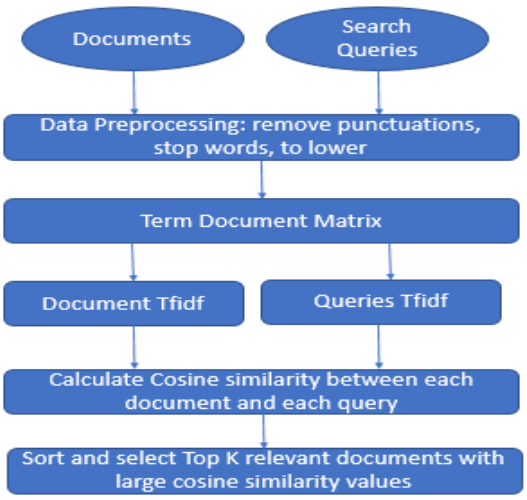

Fig 4. Vector Space Model

TFIDF is calculated as follows.

$$TFIDF(\omega) = tf.\log\left(\frac{N}{df(\omega)}\right)$$

TFIDF helps to find out the higher frequency termsby obtaining higher and positive tfidf values and terms which doesn't carry any value with negative tfidf values such as "the", "an", "is" etc. This is used to classify the reviews in terms of its features as shown in Table II.

Table- II: Vector attributes Selection

Target Attribute	Vector Attributes
Audio	Calling, Volume, Breaking, Voice
Battery	Usage, Backup, Life, Minutes, Charge
Deign	Front, Rear, Resolution, Clarity
Camera	Display, Body, Sleek
Fingerprint	Gesture, Scanner
Memory	GB, Ram, SD Card
OS	Update, Android, Software, Snapdragon
Performance	Lag, Speed, Hang
Others	Budget, Warranty

C. Mapping Features

The following Cosine Similarity formula is used to map a review to its respective featurewhere D1 is a Document comprising single review and D2 is a document comprising attribute data of any specific feature.

$$sim(D1, D2) = \frac{\sum_i x_{1i} x_{2i}}{\sqrt{\sum_j x_j^2}\sqrt{\sum_k x_k^2}}$$

By comparing the review for each feature separately we obtain similarity value. Similarity value is zero for those words which are not related.

IV. RESULTS

By plotting the ratings graph on yearly basis we can get the histogram with no data which indicates the end of product ratings which indirectly implies the end of product life time in the market. Using the senti-score obtained a new graph for the obtained ratingsis plotted. A comparison is made between the actual plot and senti-score plot to find out the correlation and actual view of the review.Fig. 5. Represents the graph for year 2017 and the corresponding senti-score graph is shown in the Fig. 6. With the obtained new ratings, we compare the average rating of a particular product with trusted sites which is shown in Table III.

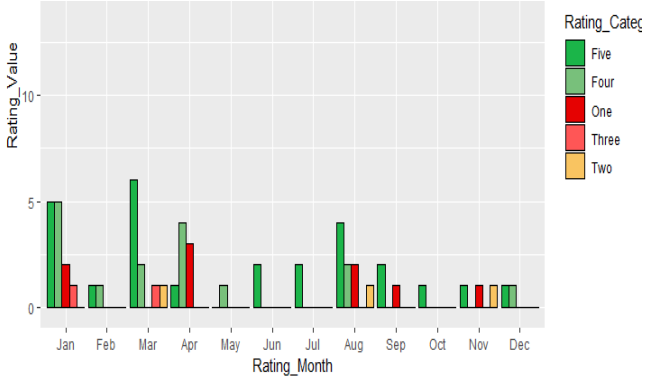

Fig. 5. Graph with ratings obtained in 2017

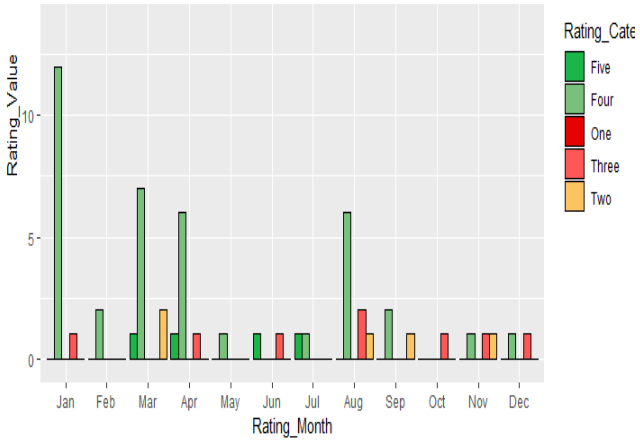

Fig. 6. Graph with senti-score obtained in 2017

Table- III: Average rating comparison

Mobile	Mouthshut Rating	Gadgets 360 Rating	Our Rating
Vivo v3	3.9	4.2	3.7
Vivo v5+	3.4	4.8	3.6
Vivo v7	2.7	4.2	3.5

With the help of obtained new ratings and reference vector for each attribute obtained in Table II aggregation of average ratings for each attribute is obtained. When plotted a graph with these averages for all the three versions of mobiles (VivoV3, VivoV5+, and VivoV7) we can generalize the product based on its features. Fig 7 shows a clear idea of success factors and failures of every mobile. We can say that VivoV3 is successful in all its features compared to other versions whereas VivoV7 being a new version was unsuccessful in many features namely audio, battery, camera etc.

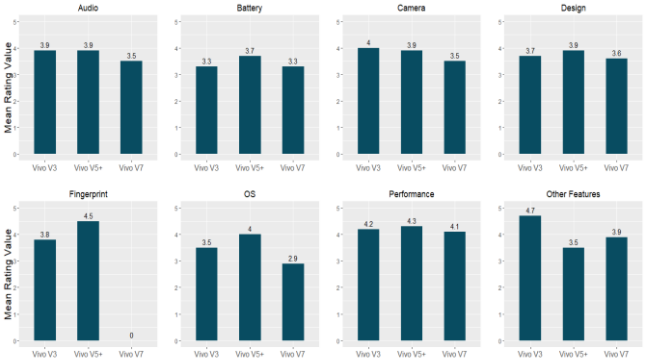

Fig. 7. Average rating for mobiles based on features

V. CONCLUSION AND FUTURE SCOPE

Customers depend a lot on reviews and ratings of a product before actually purchasing it. Sentiment analysis plays a key role in determining the facts about products. This facts can involve customer satisfaction, success/failure, future needs etc. One among the facts is analyzing proper correlation between ratings and reviews. In this paper the correlation has been addressed using lexicon based sentiment analysis which is dictionary based approach. Choosing a product depends on the interest of the developer.

As mobile phones are trending in the market the analysis has been carried out on them.The need for analyzing the correlation between ratings and reviews has been successfully carried out by generating sentiment score from the cleaned text i.e., reviews. This method clearly shows the actual rating for a particular review which helps the customers end with a good product purchase. Also taking this data as base, feature analysis has been carried out using vector space model. From this approach every product was deeply examined based on its features (Audio, Battery, Camera, Design, Fingerprint, OS, and Performance) which helps customer choose a product with best features. It also helps the producer create a new version which fulfills all the drawbacks of the older version.

Though the lexicon based approach suits best saving lot of time and money there needs an analysis of sentences rather than words which gives a more clear vision for analysis. A comparative study between lexicon and machine learning can be made to analyze the results more efficiently. Vector Space model being easy and based on linear algebra it mostly depends on search attributes provided. Also, a comparative study can be made by implementing other feature selection algorithms.

REFERENCES

1. Tan, A.H., 1999, April. Text mining: The state of the art and the challenges. In Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases (Vol. 8, pp. 65-70). sn.
2. Pang, B. and Lee, L., 2008. Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval, 2(1–2), pp.1-135.
3. Behdenna, S., Barigou, F. and Belalem, G., 2016, August. Sentiment analysis at document level. In International Conference on Smart Trends for Information Technology and Computer Communications (pp. 159-168). Springer, Singapore.
4. Wilson, T., Wiebe, J. and Hoffmann, P., 2005, October. Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of the conference on human language technology and empirical methods in natural language processing (pp. 347-354). Association for Computational Linguistics.
5. Schouten, K. and Frasincar, F., 2016. Survey on aspect-level sentiment analysis. IEEE Transactions on Knowledge & Data Engineering, (1), pp.1-1.
6. Prabowo, R. and Thewall, M., 2009. Sentiment Analysis: A combined approach. Journal of Informetrics, 3{2), pp.143-157.
7. Abirami, A.M. and Gayathri, V., 2017, January. A survey on sentiment analysis methods and approach. In Advanced Computing (ICoAC), 2016 Eighth International Conference on (pp. 72-76). IEEE.
8. Hu, N., Bose, I., Koh, N.S. and Liu, L., 2012. Manipulation of online reviews: An analysis of ratings, readability, and sentiments. Decision Support Systems, 52(3), pp.674-684.
9. Ravi, K. and Ravi, V., 2015. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. Knowledge-Based Systems, 89, pp.14-46.
10. Singh, A. and Tucker, C.S., 2017. A machine learning approach to product review disambiguation based on function, form and behavior classification. Decision Support Systems, 97, pp.81-91.
11. Hu, M. and Liu, B., 2004, July. Mining opinion features in customer reviews. In AAAI (Vol. 4, No. 4, pp. 755-760).
12. Ray, P. and Chakrabarti, A., 2017, February. Twitter sentiment analysis for product review using lexicon method. In Data Management, Analytics and Innovation (ICDMAI), 2017 International Conference on (pp. 211-216). IEEE.
13. Vu, Linh & Le, Thanh., 2017. A lexicon-based method for Sentiment Analysis using social network data. In International Conference on Information and Knowledge Engineering (IKE), 2017
14. Sonawane, S.L. and Kulkarni, P.V., 2017, October. Extracting sentiments from reviews: A lexicon-based approach. In Intelligent Systems and Information Management (ICISIM), 2017 1st International Conference on (pp. 38-43). IEEE.
15. Singla, Z., Randhawa, S. and Jain, S., 2017, June. Sentiment analysis of customer product reviews using machine learning. In Intelligent Computing and Control (I2C2), 2017 International Conference on (pp. 1-5). IEEE.
16. Singh, W., 2017, March. Sentiment analysis of online mobile reviews. In 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT) (pp. 20-25). IEEE.
17. Min, H., Yun, J. and Geum, Y., 2018. Analyzing Dynamic Change in Customer Requirements: An Approach Using Review-Based Kano Analysis. Sustainability, 10(3), p.746.
18. Chandrashekar, G. and Sahin, F., 2014. A survey on feature selection methods. Computers & Electrical Engineering, 40(1), pp.16-28.
19. Guyon, I. and Elisseeff, A., 2003. An introduction to variable and feature selection. Journal of machine learning research, 3(Mar), pp.1157-1182.
20. Raghavan, V.V. and Wong, S.M., 1986. A critical analysis of vector space model for information retrieval. Journal of the American Society for information Science, 37(5), pp.279-287.

AUTHORS PROFILE

Susmitha Gunti, doing Master of Technology in Department of Computer Science and Engineering at Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Karnataka – 576104
Email: Susmitha.gunti@gmail.com
Susmitha has done project research in Datamining which is trending in current scenario. Also interested in Machine Learning and Artificial Intelligence. She has consistently scored top 5 in academics.

Krishnamoorthi Makkithaya, Professor in Department of Computer Science and Engineering at Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Karnataka – 576104
Email: k.moorthi@manipal.edu Krishnamoorthi is interested in Computer Engineering Research. He is an NBA coordinator for Post graduates at Manipal Institute of Technology. He has published various papers on Data mining and Machine Learning.

Deepthi S, Assistant Professor Sr. Scale in Department of Computer Science and Engineering at Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Karnataka – 576104
Deepthi is interested in Cloud Computing and Information security. She has published a paper on "Secure audio steganography using lifting wavelet transform".
Email: deepthi.s@manipal.edu