Mining Aspects on the Social Network

Sapna Juneja, Abhinav Juneja, Rohit Anand, Paras Chawla

Abstract*: This paper proposes an effective concept of mining the feedback of product given by the user. In return various solutions are suggested according to the ratings of the aspect and its corresponding weightage. The satisfaction of user is determined by the help of user's rating and weight of the aspect determines the significance of each aspect in the user's review. These methodologies are thus, important and play a significant role for the manufacturers and producers to improvise their product and eventually leading to rise in the market value of that particular product. The methodology here extracts the aspects from the feedbacks of users with the help of conditional probability and bootstrap technique. Also an approach that is supervised and is called by the name, Naïve Bayes is used to classify aspect ratings and the sentiment words are considered as properties or features.

***Index Terms: Aspect, Aspect Extraction, Aspect Weight, Aspect Ratings, Core Term, Naïve Bayes Conditional Probability*.**

I. INTRODUCTION

Today the world is growing rapidly in terms of the technology which is the cause of social media coming into the picture. Almost every kind of emotions are being shared on the social apps. So, all these opinions can be put a very good use. But these opinions are so large in number that mining such sentiments might raise an issue. To find a solution for such an problem, a lot of work has been done [1], mining the required data from such opinions [2, 3, 4, 5], monitoring the emotions [6, 7, 8, 9], and so on. In here, we will work at building solutions for these tasks as mentioned.

- Mining the aspects related to product reviews
- Analysing the views of users on aspects
- Observing the weight emphasised on each aspect

An aspect is the classifying the user's opinion in terms of positivity and negativity. For instance, let us consider a customer drinking coffee. He can rate the coffee based on the body, taste, aroma and acidity. Based on these emotions the aspects can be mined. The problem here may be that some aspects might be mentioned explicitly but there are some aspects which user might not have mentioned like about the temperature of the coffee while served [10]. Here we will discuss both the aspects that is explicit and implicit ones.

Revised Manuscript Received on July 18, 2019

Sapna Juneja, Department of Computer Science, B.M.I E.T., Sonepat, India.

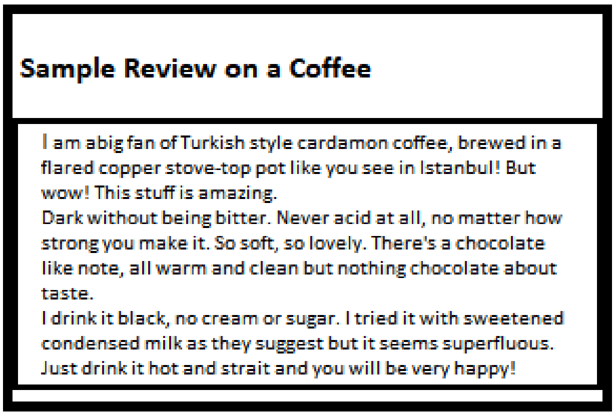
Abhinav Juneja, Department of Computer Science, B.M.I.E.T., Sonepat, India.

Rohit Anand, Department of Electronics and Communication Engineering, G.B.Pant Engg. College, New Delhi, India.

Paras Chawla, Department of Electronics and Communication Engineering, Chandigarh University,Mohali, India.

Also, another problem might arise which involves the irrelevant aspects. Therefore, this issue will also be discussed in the paper.

Initially the models used for such problems were unsupervised [11]. They used frequency-based approach but the problem was that less frequent and implicit aspects were ignored. And therefore results were not good as expected [12, 13]. These problems resulted in the usage of supervised models Hidden Markov Model (HMM) and Conditional Random Field (CRF). However, these techniques can be a little bit costlier than traditional ones.



Sample Review on a Coffee

I am abig fan of Turkish style cardamon coffee, brewed in a flared copper stove-top pot like you see in Istanbul! But wow! This stuff is amazing.
Dark without being bitter. Never acid at all, no matter how strong you make it. So soft, so lovely. There's a chocolate like note, all warm and clean but nothing chocolate about taste.
I drink it black, no cream or sugar. I tried it with sweetened condensed milk as they suggest but it seems superfluous. Just drink it hot and strait and you will be very happy!

Fig. 1:*Sample review on a coffee*

Procedure used to extract the aspects from the opinions is based on conditional probability combined with the bootstrap technique. The idea is to create an assumption that the universal set of all the aspects are already present with the aspect words, which ate generally known as *core terms* (terms used to describe aspects). The above prediction is generally true and acceptable as the aspects are small in number and can be easily mined. In some cases, there are very few core terms or even no core terms at all which can be a challenging situation. Therefore, the set of core terms is regularly updated and new terms are added to resolve the issue. This is done with the help of the concept of the conditional probability and technique of bootstrap.

The identification of aspects comes out to be a very helpful tool for the owners as they get to know the user's opinions and sentiments regarding the product, which in return leads to the improvement and upgradation of the quality of the product.

Generally, the users tend to give an overall review regarding their own experience of using the product. However, this overall review is the summation of the feedback of particular aspects of the products. Here comes, the *weights* into the picture. Weights are here to act as the parameters to measure the significance of each aspect [14, 15].

To be more clearer, the weight of the aspect is balanced with the frequent occurrence of the word pointing towards the particular aspect. Further, a approach which is supervised and called as the Naïve Bayes classification method comes into play and analyses the user's sentiments.

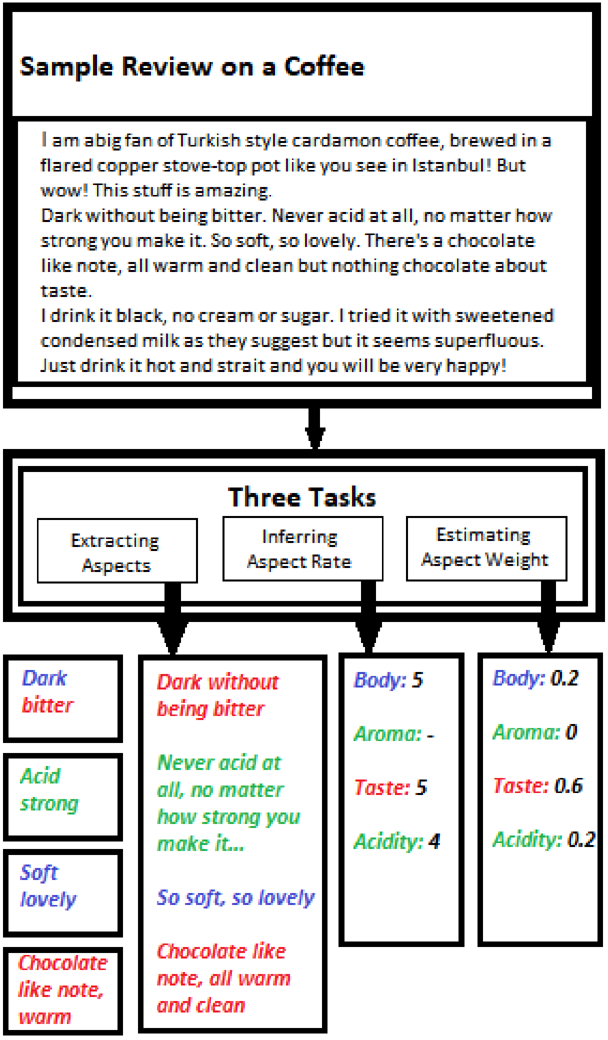

Sample Review on a Coffee
I am abig fan of Turkish style cardamon coffee, brewed in a flared copper stove-top pot like you see in Istanbul! But wow! This stuff is amazing.
Dark without being bitter. Never acid at all, no matter how strong you make it. So soft, so lovely. There's a chocolate like note, all warm and clean but nothing chocolate about taste.
I drink it black, no cream or sugar. I tried it with sweetened condensed milk as they suggest but it seems superfluous. Just drink it hot and strait and you will be very happy!
Three Tasks
Extracting Aspects
Inferring Aspect Rate
Estimating Aspect Weight
Dark bitter
Acid strong
Soft lovely
Chocolate like note, warm
Dark without being bitter
Never acid at all, no matter how strong you make it...
So soft, so lovely
Chocolate like note, all warm and clean
Body: 5
Aroma: -
Taste: 5
Acidity: 4
Body: 0.2
Aroma: 0
Taste: 0.6
Acidity: 0.2

Fig. 2:*Block Diagram of Step-wise processing of the review*

II. PROPOSED METHODOLOGY

A lot of study and research has been going on lately, in the field of big data, to extract the opinions of the users based on the experience of various products. The user's opinion and extracting the corresponding aspect from their reviews has been a task on which various methods are being proposed to help the product manufacturers to upgrade their product according to improvisations required, and significantly changing the user's experience towards betterment, eventually benefiting the market of the product, up to a much greater extent. According to a survey [16], conducted on opinion mining and sentiment analysis, it is clearly stated that the two very important tasks are

- Identification of aspect
- Rating Inference based on aspect

In the survey mentioned above, some very interesting methods are proposed including frequency-based, lexicon-based, machine learning and topic modeling.

The approaches based on the frequency are one of the earliest approaches [11]. In the above mentioned approach, the nouns as well as the phrases of nouns are used as the candidates of the aspect [17, 18]. An algorithm for mining the data is used by Hu and Liu [10] for the identification of nouns and the phrases of nouns [19]. The count of most occurring frequencies is considered and the ones with high occurrences are kept. As simple as this method might seem, but it is quite effective method. There are even a few companies which are using the method mentioned above in order to effectively increase the business growth [11]. But the major challenge, that exists here in the method above is that a lot of noise is created, i.e. irrelevant data which consists of non-aspects items is created.

In order to deal with such challenges, the methods of filtering have been [13]. These methods are merged with the frequency-based approach to eliminate the non-aspect elements. A solution which is similar to the above one, performs extraction on the nouns/aspects according to their frequency of occurrence and the information in it [12]. Initially, the search is performed for the seed words for each noun.

Table 1: An example of Seed words

	Aspects	Seed Words
HOTEL	Value	Value, Price Worth
	Room	Room, rooms
	Location	location
	Cleanliness	Dirty, Clean
	Check in	Staff
	Service	Food, Staff

Then they find the information to other words related to their aspect. For instance, it can find "$" or "Dollars" for the price or cost of the aspect. However, these approaches based on rules and frequency requires the manual assistance into parameters.

Now, to resolve the issues of approaches based on frequency, the concept of modeling has been proposed into the picture. The concept of modeling enables one to uncover the topics from large collection of texts. This concept requires the two basic models which are, PLSA (Probabilistic Latent Semantic Analysis) and LDA (Latent Dirichlet Allocation).

The concept of modeling is implemented by the authors in [4, 13], that can directly be connected to the aspects. For mining aspect, this concept is presented by. In the first step, they the aspects are verified using the concept of modeling. And then secondly, only the adjectives are taken into considerations to evaluate the sentiment lying behind the aspect. Topics like JST (Joint Sentiment Topic) and Reverse-JST are also presented in [4]. Latent Dirichlet allocation (LDA) is the basis of both the models mentioned above. The emotions as well as their polarity i.e. positive or negative can be mined using the models. The yield of at least 76.6% accuracy is obtained by both JST and RJST on Pang and Lee [7] dataset.

As the modeling concept tends to distinguish words for each aspect, in the words representing the aspects and the words representing the emotion behind that aspect is separated. In order to proceed and perform this study, two parameters are used to encode the above properties. We create a weighted graph which is bipartite, for each review. Aspect labels are learnt by either no supervision or weak or full supervision. In case of no supervision, only ratings of aspect are used. In case of weak supervision, manually labelled sentences are used in small numbers to un-label data. And in case of full supervision, only data which is labelled manually is used.

As mentioned in another method called AIR (Aspect Identification and Rating) model is used to mine the reviews and ratings. In the AIR model, the sampling of distribution of word of the aspect for each review is influenced by an aspect rating. The basis of this approach is LDA model. Overall ratings given by the reviewers affect the sampled latent aspect rating which further influences the word sampling for each aspect and their extraction, which is quite different from traditional approach. The imbalance of aspects present in the shorter reviews is handled by AIR which is further enhanced.

Although, the approach of the modelling concept is based on probabilistic inference, it has its own shortcomings which are responsible for it not being used in real life opinion analysis applications. Let us take for an instance the volume of data it requires, which is very huge. Not only this, but a proper synchronization is also required to achieve the desired results. Frequent and common topics can be easily searched but when it comes to search those which are locally frequent but less global terms, it becomes a challenging task to perform. Often, these local ones are of the most use because they are the most significant for particular entities in which the customer is interested. To summarize it, we can say that generally the concept of modeling is not so significant for a number of real time applications to analyze the sentiments [11].

Also, the other approach based on lexicon methods is used. It also comes under the category of unsupervised approach. The emotions are predicted from the reviews using the dictionary. In return, this also help one obtain the polarity and the core strength. The dictionary contains the words and phrases and with them, the sentiments as well as orientation of the opinions are also attached. The score for any particular sentiment can be computed [8]. Xiaowen Ding, Minqing Hu make use of sentiment classification of the aspect and the sentence [10]. A method called EXPRS (An Extended Page Rank algorithm enhanced by a Synonym lexicon) is presented by Yan et al. . The above method then helps to mine the features and properties of the product. For this, the nouns and the phrases of the nouns are mined initially and then followed by the mining of the dependency relations between nouns and phrases of the noun and their corresponding words of sentiments. The relations like subject-predicate, adjective-modifications, relative-clause-modifying and verb-object are included in relations of dependency. The synonyms for the properties of the product are also introduced in it. The elimination of nouns with no feature also takes place which is based on the concept of proper nouns, name of brands, verbal as well as personal nouns. Movie review was taken as an instance and a methodology was proposed by Peñalver-Martinez et al. which performed the analysis on the sentiments on the reviews of the movie, based on the aspects. A domain ontology named as Movie Ontology to mine the features from the reviews. And then the score of the sentiment is calculated by the utilization of SentiWordNet. However, it had a drawback, that was creating such a lexicon of the sentiments considering the facts that it has high cost, and also it takes a long time period to build such dictionaries.

For higher accuracy the approaches based on machine learning [20] can also be used to perform classification of the sentiments. Here also, the method can be either supervised or unsupervised. In case of methods that are supervised, two different sets of annotated data can be considered- the first one for training and the second one for testing. For supervised learning, Decision Tree (DT), Support Vector Machine (SVM), Neural Network (NN), Naïve Bayes, and Maximum Entropy (ME) are a few classifiers which are applied commonly. A method, for selection of feature, with SVM i.e. Support Vector Machine classifier used to classify sentiments for large movie review data set and which is based on Gini Index is presented by Asha et al. Due to the introduction of Gini Index, the accuracy of the selection of feature while analyzing the sentiments, is improved.

Duc-Hong Pham and Anh-Cuong Le have designed an architecture of multiple layers to represent the knowledge and different levels of sentiment in an input text. Then the neural networks are introduced and combined with the above representation to construct a model for predicting the overall feedback of the products. However, such techniques and methods demand a set of manual power for labelled data for the model to be trained [21]. And this eventually leads to further challenges like higher costs and more efforts.

III. RESULT ANALYSIS

The ambition of the task is to allot a part of labels of aspect taken from the set which consists of all the labels of aspect of a given product, to every sentence in a feedback given by some user or customer. The determination of labels of an aspect has the aspect words and terms as its basis. Some initial core terms are provided for each and every label present in the universal set. However, the challenging situation may arise in some cases where very few core terms or even no core terms are present at all, for the all the feedbacks. This may eventually result in the assignment of labels which are completely incorrect for the sentences these are being assigned for. Thus, it becomes an important task to take a wider set of core terms for the aspect words based on the data which is provided by the user in the form of opinions shared by user as the feedback. Therefore, in the methods which are discussed above use of Bayes or Hidden Markov Model is preferred. And as discussed above, the method relies on use of conditional probabilistic model which is merged with the Bootstrap technique to construct the set of aspect words.

Here, figure 3 describes the 4 aspects of a product that is coffee that is represented by the respective aspect words. The symbol that denotes core terms is represented by O, the symbol that denotes the words appearing is represented by X. The four aspects of the coffee which are *body*, *taste*, *aroma*, and *acidity* are already mentioned. The sets of the core terms which are corresponding to the aspects are as follows:

- {body}
- {taste, aftertaste, finishing, mouthfeel}
- {aroma, smell, flavour}
- {acid, acidity}

These terms are then enriched by putting the words which have a high probability of their appearance in the similar sentences that they occur in. The four circles shown in the figure represent the group of aspect words. Overlapping of these circles indicate there might be some aspect word which are present in the multiple aspects.

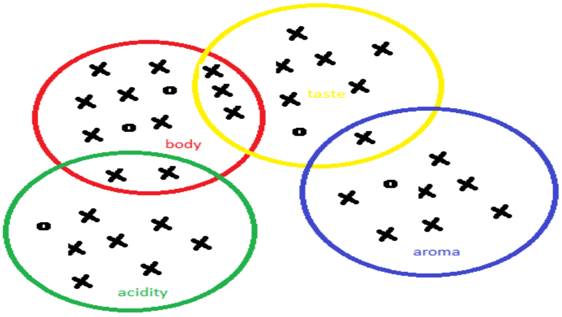


Fig. 3:*Core terms with aspects*

Let us consider a supposition as A={a_1,a_2,…,a_K}which is a set or group of K number of aspects. A group or set of the words which happens to appear in the sentences and is labelled as a_j for every $a_j a_j$, is present in such a way that their frequency of occurrences exceeds a threshold which is given is obtained. The overlapping of the two or more sets signifies that some core terms may be present in more than one aspect.

a. Initially, the sentences which hold minimum one word in the original core terms of that particular inspected are tracked.
b. Secondly, all the words that constitutes the nouns, phrases of nouns, adjectives, and adverbs in the specific sentences are located.
c. Now, the threshold θ is considered and depending on the number, the words which occur more than threshold θ are put into the group of aspect words.
d. Then the core terms are excluded and the new set of aspect words is put under consideration.
e. Further the sentences which are required, are searched and found.
f. The steps mentioned above are repeated until the number of words found is zero.

IV. CONCLUSION

This conclusion of this paper, cam be summarized with the help of the three basic tasks which are, (1) Mining the aspects related to product reviews, (2) Analyzing the views of users on aspects, and (3) Observing the weight emphasized on each aspect.

With the help of the above tasks, the aspect of the user's review can be determined. The determination of the aspect with the help of various methodologies and techniques which are discussed in the paper, enables the manufacturer to obtain the feedback from users in way which can be analyzed to arrive on some results. Then, these results are helpful to decide further improvements to be made in the product. It drastically improves the product quality, leading to a better user experience. Due to this the market of the product can also arise up to a much greater extent.

The methodologies are simple one and do not need overall ratings. Also, it works very well on real world applications. These techniques can be a bit costly and may require some extra efforts.

The issue of mining of aspect from the data which is not labelled can be considered in future. Not only this, but the above concept can be implemented to other areas as well which include, movies, business and other domains.

REFERENCES

[1] S Park, K Lee , JSong: Contrasting opposing views of news articles on contentious issues. Proceedings of the 49th annual meeting of the association for computational linguistics (ACL-2011),2011.

[2] MVan Den Camp, A Bosch:The socialist network. Decis Support Syst, 2012,pp.761-69.

[3] S K Li, Z Guan, LY Tang :Exploiting consumer reviews for product feature ranking. J Comput Sci Technology, 2012, pp. 635-49.

[4] C Lin, Y He, R Everson, S Ruger: Weakly supervised joint sentiment-topic detection from text,IEEE 2012, pp.1134-45.

[5] J Zhan, HT Loh, Y Liu: Gather customer concerns from online product reviews—a text summarization approach. Expert Sysem 2009, pp.2107-15.

[6] Y Dang, Y Zhang, H Chen, A lexicon-enhanced method for sentiment classification: an experiment on online product reviews. IEEE, 2010, pp.46-53.

[7] B Pang,L Lee: A sentiment education: sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the,42nd annual meeting on association for Computational Linguistics. 2004.

[8] M Taboada, J Brooke, M Tofiloski, KVoll, M Stede: Lexicon-based methods for sentiment analysis.,2011, pp.267-307.

[9] P D Turney:Semantic orientation applied to unsupervised classification of reviews. Proceedings of the 40th annual meeting on association for computational linguistics, 2002, pp.417–24.

[10] M Hu,B Liu:Mining and summarizing customer reviews, Proceedings of the Tenth ACM SIGKDD international conference on knowledge discovery and data mining, New York, 2004, pp. 168–77.

[11] B Liu: Sentiment analysis and opinion mining. Synth Lect Human Lang Technology1–67.CrossRef, 2012.

[12] C Long, J Zhang, X Zhut:A review selection approach for accurate feature rating estimation. Proceedings of Coling ,2010.

[13] S Moghaddam ,M Ester: Opinion digger: an unsupervised opinion miner from unstructured product review, Proceeding of the ACM conference on Information and knowledge management , 2010.

[14] Li Chen, Feng Wang:Preference-based clustering reviews for augmenting e-commerce

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

recommendation. Knowledge Based System, 2013, pp.44-59.

[15] H Wang, Y Lu, C Zhai: Latent aspect rating analysis on review text data: a rating regression approach, Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining, 2010, pp.783-92.

[16] K Ravi:A survey on opinion mining and sentiment analysis: tasks, approaches and applications. Knowledge Based System, 2015, pp.14-46.

[17] A M Popescu, O Etzioni :Extracting product features and opinions from reviews, Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing.2005, pp. 339–46.

[18] J Zhu, H Wang, B K Tsau, M Zhau:Multi-aspect opinion polling from textual reviews, Proceedings of ACM international conference on information and knowledge management,2009.

[19] B Santorini: Part-of-speech tagging guidelines for the Penn Treebank Project, University of Pennsylvania, School of Engineering and Applied Science, Dept. of Computer and Information Science, 1990.

[20] Shivam Agarwal: Data Mining: Data mining concepts and techniques,International conference on machine learning and research advancement, IEEE, 2013.

[21] Xiuzhen Zhang, Shuliang Wang, Gao Cong,Alfredo Cuzzocrea:Social Big Data: Mining aspects, applications and beyond , Wiley, 2019.

AUTHORS' PROFILE

Sapna Juneja is Professor in Computer Science Deptt. in B.M.Institute of Engineering and Technology, Sonepat. Dr. Sapna has a teaching experience of more than 14 years including UG and PG courses. She received her Ph.D. Degree in Computer Science from MDU, Rohtak, India in 2018, and M.Tech (Computer Science) in 2010.Her topic of research is Software Reliability of Embedded System. Her areas of Interest are Software Engineering, Computer Networks, Operating System, Database Management Systems, Artificial Intelligence etc. She has guided several research thesis of UG and PG students in Computer Science and Engineering.

Abhinav Juneja is Professor and Head of Department of Computer Science Engg. in B.M.Institute of Engineering and Technology, Sonepat Dr. Abhinav has a teaching experience of 18 years including UG and PG courses.. He received his Ph.D. Degree in Computer Science from MDU, Rohtak, India in 2018 and M.Tech from GGSIPU, Delhi, India in 2007. His topic of research is Intelligent Selection of Software Reliability Growth Models. His areas of interest include Software Reliability, Software Engineering, Microprocessors, Database Management Systems, Internet of Things, Computer Architecture etc. He has mentored several engineering students in their research thesis and innovation based competitive events.

Rohit Anand is Assistant Professor in Electronics and Communication Engineering Department in G.B.Pant Engineering College, New Delhi, India. He has done B.E.(ECE) from M.D.University, Rohtak, India in 2001 and M.Tech. (ECE) from P.T.U.,Jalandhar, India in 2008. He has a teaching experience of more than 17 years including UG and PG Courses. He is currently pursuing Ph.D. from P.T.U., Jalandhar, India. He is a Life Member of Indian Society for Technical Education. He has published more than 30 papers in National and International Conferences and Journals. His research areas include Electromagnetic Field Theory, Antenna Theory, Image Processing, Optical Fiber Communication etc.

Paras Chawla received his B.Tech. (Honors) and M.Tech. degree in Electronics and Communication Engineering from Kurukshetra University; NIT, Kurukshetra and Ph. D. from Thapar University, Patiala. He has more than 14 years of teaching experience and currently working as Professor & HOD in ECE Department at Chandigarh University, Mohali (India). Total sixteen M.Tech. Dissertations of various fields of ECE has been guided by him successfully. He is also guiding 07 students of Ph.D. He received the "Coventor Scholarship award" from MANCEF, New Mexico-USA for his proposal titled "Performance & Analysis of RF Front Section of Mobile Terminal Using RF MEMS DC Contact Switches", under the name of conference "Commercialization of Micro-Nano Systems Conference (COMS 2010)". His team received consecutive two years "Tenderfoot Award" collaborated given by American Astronautical Society, American Institute of Aeronautics, NASA/Goddard Space Flight Centre, NASA/Jet Propulsion Laboratory & Naval Research Laboratory for "CanSat Competition", June 2015 & 2016, Burkett, Texas, USA. His main research interest includes microstrip antennas, RF MEMS, RF front end mobile terminal, wireless & mobile communication, Optimization Algorithm for RF circuits, LTE, and 5G. He has published more than 50 papers in various reputed National and International Journals/Conferences, one SCI book chapter and ten patents filled.