International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-6, April 2020

A Three Stage Summarization Framework for Product Recommendation from Opinion Tweets

Salina Adinarayana, E Ilavarasan

Abstract: ***This paper presents a three stage framework for feature -based rank of opinions around target product and features in a collection of product review tweets as twitter corpus. The main stages of the framework are discussed and presented the analysis of the sentiment with respect to features of the digital camera sales in ecommerce sites.***

Keywords: framework, target product.

I. INTRODUCTION

Demographics discover shoppers in online shopping based on their age, income, marital status and other physical features. Psychographics attempt to find out why a shopper buys a certain product. To support online shoppers, proposing a useful approach based on opinion ranking to generate product opinions summary to take wise decision before attempting product purchase. In the previous work, authors have implemented different algorithms for balancing and classifying opinions from imbalanced twitter corpus to generate document level product opinion summary. In this paper, it extends the previous work by taking the results of the algorithms on imbalanced twitter corpus consisting of tweets in 1011 digital camera review documents. Details of the review tweets are shown in Table 1.This paper addressing an important aspect of how the algorithms are selected by CIL_Classifier for generating document wise opinion summary. First of all, the twitter corpus is pre processed using two layer features vector generation approach [1] to eliminate the missing values, special characters and unnecessary blank spaces for improved performance of the classifiers.

Table 1 Product Review Tweets

Dataset	Product	Product Review Documents	Attributes	No. of Reviews in each Document
Twitter Product Reviews	Digital Camera	1011	672	@ 5-10 reviews

Even after pre processing, to eliminate imbalanced nature of dataset and classify the opinions in dataset, CIL_Classifier is applied on the dataset, so that it will generate document level product opinion summary. However this is not enough for a prospective shopper in online shopping to take a better purchase decision.

Revised Manuscript Received on April 30, 2020.
***** Correspondence Author
Salina Adinarayana*, Department of Computer Science and Engineering, Raghu Institute of Technology, Visakhapatnam, India
E Ilavarasan, Department of Computer Science Engineering, Pondicherry Engineering College, Pondicherry, India

This paper focuses on how feature based opinion summary is produced as recommendation of the document level opinion summary from CIL_Classifier. To do this, LDA topic-based representation method with Gibbs sampling is considered. First, Decision tree based CIL_Classifier algorithms balances and Classify the digital camera reviews at document level, in order to explore the most clear cut opinion summary from the document level, summary of opinions have further segmented and ranked the product features for concise opinion summary as product recommendation. For this discriminative representation, 1011 digital camera review documents from twitter as datasets. Latent Dirichlet Allocation (LDA) is topic model approach used to classify the product reviews based on selected set of features from the classified review documents taken from CIL_Classifier. digital camera reviews are collected as a set of 1011 documents from twitter corpus for evolution which are imbalanced state, these imbalanced product reviews are balanced using CIL_Classifier initially, as a result it produced only broad level recommendation from the review documents for the prospective shoppers ,This broad level opinion recommendation is not enough to make a buying decision on that product .So that to give a concise opinion recommendation ,These classified reviews are further processed the Document level opinion summary to extract concise opinion based summary on important aspects of the product. This paper proposes three stage model for product recommendation list. In stage1, generating document wise opinion summary using CIL_Classifier, stage2 addresses segmentation model to segment the document wise opinion summary on feature wise followed by opinion ranking in stage 3 to explore feature level opinion recommendation for the products.

II. LITERATURE REVIEW

The authors in [2], for recommender systems initial formulations are based on direct correlation statistics and predictive modeling, not appealing the wider range of practices in statistics and ML literature. In [3] the authors discussed a technique on the use of relative information in recommender systems and proposed a data warehouse approach to predict the ratings as per user's context. In the paper [4], authors have introduced an automated weighting scheme for ratings. In the paper, the proposed approach [5] has three techniques for identifying different features with respect to a movie for collaborative filtering. In [6] authors discussed on user reviews with non-explicit rating labels, and propose a sentiment analysis based nearest neighbor model to improve the existing recommendation methods performance. In paper [7] authors have analyzed the recommendation techniques by taking the recommendations as input and applied different recommendation algorithms.

In [8], authors suggested that LDA may not necessarily perform well when working with documents that are short in length. In [9], authors discussed the importance of automated systems by analysing views of millions of people. In [10], authors discussed that a review document has probability distribution over topics and set of words are distributed over these topics. In [11], authors had discussed different applications of PLSA. In the paper [12] discussed Correlated Topic Model (CTM) to discover topic and authors had concluded that this technique has improved classification accuracy. In [13], authors had studied a method for detecting the topic evolution in Scientific literature. In [14] authors have reviewed the importance of Twitter to examine earthquakes detection. As per the discussion of authors in [15], bulkiest topics which were modeled are proved as they are almost separable. In [16] authors had concluded that meaningful features will characterize certain image streams. The authors in [17] had explored different topic models were individually developed for disability review data. Remaining parts of the paper are organized as follows: in section3 proposed CIL_classifier is discussed, section 4 discuses opinion segmentation model. In[18] authors had presented a model to handle hate speech offensive language with all sample inputs based on sentiment analysis using Twitter API. In [19] authors had presented Bi-directional recurring neural network to improve the performance of the opinion mining by taking Indian movies as dataset.

1. Proposed CIL_Classifier for Opinion summary

The idea of CIL_Classifier is shown in Fig 1, it selects one of the classifier: from CSIDL, OSIDL and USIDL algorithms after dividing the multi class reviews in twitter corpus into binary class instances [20] to balance the imbalanced twitter corpus and generate the opinion summary document level.

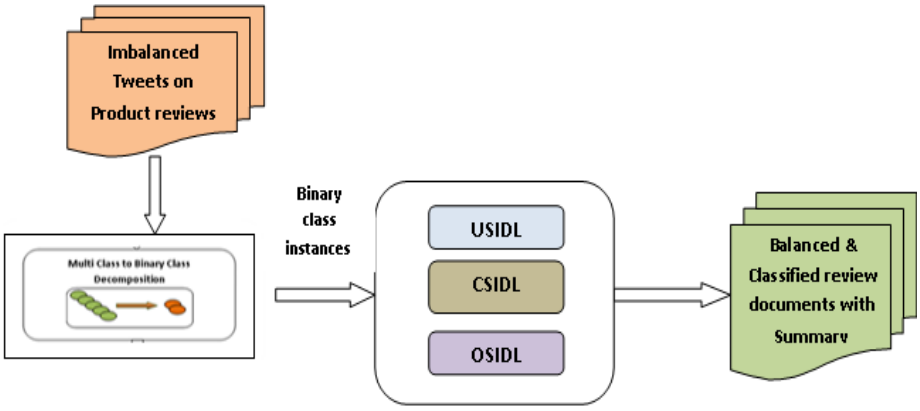

Figure 1 CIL_Classifier

If the product reviews are with more number of majority opinions, CIL_Classifier selects USIDL classifier [21] to balance the twitter corpus and generate the document level opinion summary.

If the product reviews are with more number of minority opinions, OSIDL algorithm [22] is selected by CIL_Classifier to balance the twitter corpus and produce the high level opinion summary document wise. If the product reviews are with more number of noisy, missing and borderline instances, CIL_Classifier selects CSIDL algorithm to balance the twitter corpus and produce the high level opinion summary document wise. CSIDL is similar to the C4.5 algorithm and uses the hybrid approach [23].Because this research use the hybrid approach, only a single tree is needed, and the attributes to split by are only attributes that describe users.

These attributes can be computed based on the users past ratings and the content of the items. All these three approaches in CIL_Classifier are based on Decision tree based models.

After balancing and classifying the twitter review documents on product transactions, a broad level opinion summary is generated, to narrow down the opinion summary for better and concise summary w.r.t selected features, opinions segmentation model is implemented on the opinion summary.

2. Opinion segmentation model

The idea of Opinion segmentation model is to apply LDA topic model with the Gibbs sampling for segmenting the opinions documents feature wise so that more elaborated opinion summary based on product aspects is generated. This model returns an object that contains a lot of information. Of particular interest to us are the documents to topic assignments, the top features in each document and the probabilities associated with each of those terms. In this opinion segmentation, each document in the 1011 documents is considered to be a mixture of all features the assignments in a file, lists the top feature which is with the highest probability. Each feature contains all words in the review corpus; it list only the top 6 terms in another file. Finally, the last file lists the probabilities with which each feature is assigned to a document. The highest probability in each row corresponds to the feature assigned to that document. The “goodness” of the main assignment can be assessed by taking the ratio of the highest to second-highest probability and the second-highest to the third-highest probability and so on. In general, if a document has multiple features with comparable probabilities, it simply means that the document articulate to all those features in proportions indicated by the probabilities. Summary of product review document will convince the shopper the expressed opinions did indeed range over all those features.

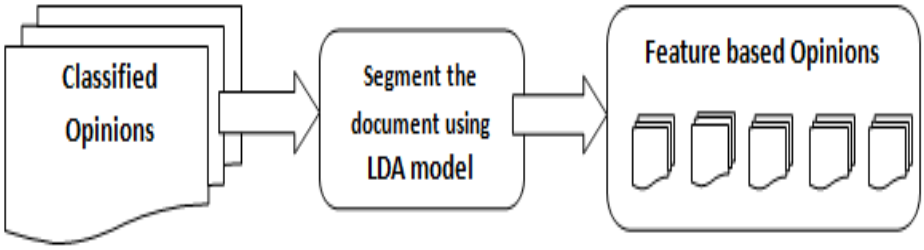

Figure 2:Segmenting the opinion features

III. RANKING FEATURES

In this process each document wise summary from CIL_Classifier is segmented further into number of feature based opinions. For example in a product review of a digital camera, segment the product review based on the features: quality, price, brand, Camera lens, burst mode, power source, Image sensor, Focal-plane shutter etc. In the review, document wise ranking is done with pairing function. A pairing function is used here to uniquely encode opinion-feature and score of that feature in to a single score of that document. Finally the scores of all the documents are generated to form final recommended opinion summary.
Algorithmic steps are as follows:

2.1 Algorithm: Rank_Summary

Let the classified opinions are in different documents of the review dataset D

Input: Classified opinions of the review dataset D

Output: Recommended feature wise opinion summary with its rank in D^i and D^s arrays

Procedure: Rank_Summary Algorithm

Step 1: for i=1 to count (D)

Step1.1: for each Document D_i

Step 1.1.1: For each opinion Oi in Document D_i

Step 1.1.1.1: for each feature f_i in opinion O_i

Step 1.1.1.1: Segment each opinion instance O_i in D_i into number of opinions based on feature f_i using topic model LDA with Gibbs Sampling and rename it as feature based opinion O_i^i

Step 1.1.1.2: next feature f_i

Step 1.1.2: next opinion O_i

Step 1.1.3: Assign score to each O_i^i by counting its occurrence in the current document D_i

Step 1.1.4: Generate opinion-score pair $<O_i^i, Score_i^i>$ for each feature based opinion in D_i

Step 1.1.5: Rank the feature wise opinions in document D_i according to $<O_i^i, Score_i^i>$ pair and extract the opinions O_i^i with highest scores to form the summary of that Document D_i .Store this ranked O_i^i in D_i^i

Step1.1.6: Generate the Document-opinion score pair $<D_i, O_i^i>$ and generate the Document Score D_i^s using a pairing function.

Step 1.1.7: Display document wise opinion summary D_i^i and its opinion rank D_i^s

Step 1.2: next D_i

Step 2: next i

3. Analysing the opinion Segmentation for feature based opinion summary

There are many reasons to segment each document opinion summary in the twitter corpus aspect wise.

Every new buyer as a prospective customer will not satisfy with only one aspect or feature of the product, he need to observe different aspects of the product to finalize his buying decision. In that case topic modeling on these product transactions based opinions is essentially useful not only for the new buyer but it is also useful for the seller of that product. It is also useful for the seller of the product to analyze the buyer expectations on the products.

Summary from CIL_Classifier is a set of opinion documents, will not reveal feature wise opinion summary so each of these opinion documents are further segmented into number of feature based opinions. To segment each document summary, LDA topic model is used, which typically results each summarized document being represented as a vector with set of features based opinions as shown in Figure 3, where each entry of this vector describes the particular opinion feature in summarized opinion document. Use this vector as a list of feature based opinions describing the document, in addition to other feature based opinions that might be collected resulting feature vector can then be used along with clustering algorithm to perform clustering of documents.

3.1 Generate Opinion document wise opinion score

By generating document wise opinion score, it produces the rank of the document. The detail flow for generating opinion score document wise for summary generation is shown with a process flow chart shown in Figure3.

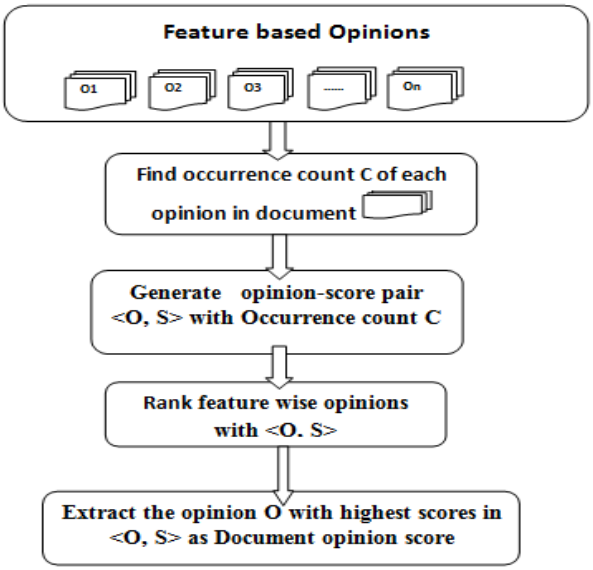

Figure 3: Generating Document wise ranked opinion summary

IV. RESULT ANALYSIS

For modelling the features in the opinion document, topic model LDA with the Gibbs sampling is used. Opinions summary in a set of review documents in bag of words representation are collected from CIL_Classifier, to infer the underlying feature based opinions in the opinion summary of each document. Gibbs Sampling is an algorithm in a family of algorithms in the Markov Chain Monte Carlo (MCMC). The MCMC algorithms aim to construct a Markov chain that has the target posterior distribution as its stationary distribution. Gibbs Sampling is based on sampling from conditional distributions of the variables of the posterior. For example, to sample x from the joint distribution $p(x) = p(x_1, ..., x_m)$, where there is no closed form solution for p(x), but a representation for the conditional distributions is available, using Gibbs Sampling. This procedure is repeated 'a' number of times until the samples begin to converge to what would be sampled from the true distribution. While convergence is theoretically guaranteed with Gibbs Sampling, there is no way of knowing how much iterations are required to reach the stationary distribution. Therefore, diagnosing convergence is a real problem with the Gibbs Sampling approximate inference method. However, in practice it is quite powerful and has fairly good performance. Typically, an acceptable estimation of convergence can be obtained by calculating the log-likelihood or even, in some situations, by inspection of the posteriors.

To calculate the probability distribution of each feature assignment, Gibbs Sampling samples the discrete distribution from and the chosen feature is set in the z array and the appropriate counts are then incremented. Typically, an acceptable estimation of convergence can be obtained by calculating the log-likelihood or even, in some situations, by inspection of the posteriors. For implementing this feature segmentation with LDA model we have selected all the classified opinions in the twitter dataset with 1011 document instances. Five important features namely: Resolution,Memory,Flashtype,Burst mode and Optical zoom as feature1,feature2,feature3,feature4 and feature 5 respectively are considered on which the classified opinions should be segmented, however one may chose more number of features if required for segmentation. For these five features in feature based opinions O_i rank is generated them within the entire review dataset by averaging the feature wise probability distributions.

4. Analyzing feature distribution for document ranking

Let us analyze the feature-wise word distribution in the first 10 opinion documents on five features of the shopper/customer interest in the graph shown in figure 7.5. In document D1, it has highly correlated words on Resolution with probability of 0.492147 out of 1.

Document D2 has highly correlated words on Resolution with a probability of 0.259912 out of 1.

In document D3, it has the words with most likely belongs to Memory with a probability of 0.412346 out of 1.

In document D4, it has the words mostly correlated to Resolution with a probability of 0.36612 out of 1. Document D5 has the words inclined mostly towards Memory with a probability of 0.260536 out of 1.

In document D6, it has highly correlated words on Resolution with a probability of 0.341772 out of 1. Document D7 is with highly correlated words on Flash type with a probability distribution of 0.374795 out of 1. In document D8, it has highly correlated words on Flash type with a probability of 0.358434 out of 1. Document D9 has the words mostly correlated to Burst mode with a probability of 0.415094 out of 1. In document D10, it has the set of words mostly related to Resolution with a probability of 0.356502 out of 1. Percentage of feature distribution in the opinion documents of different digital camera of the dataset are shown in Figure 2.

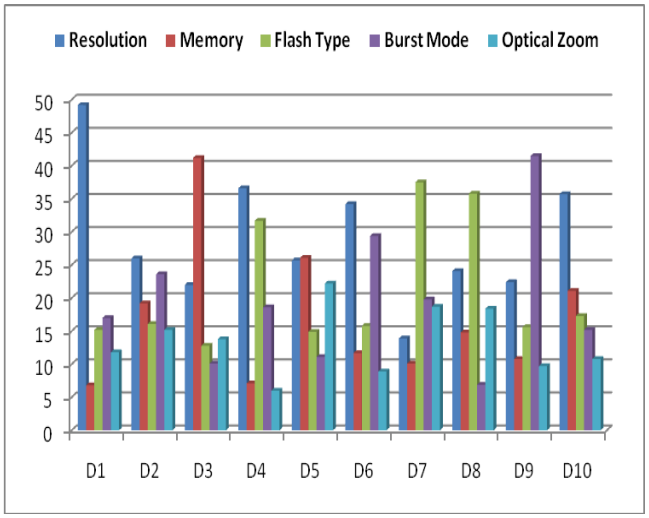

Figure 2 Frequency of word distribution feature wise in opinion document

In general, if an opinion document has multiple features with comparable probabilities, it means that the opinion document speaks to all those features in proportions indicated by the probabilities. The graph shown in Figure2 describes the document rank feature wise.

To calculate the feature wise opinion rank, the probabilities are averaged to find the ranked features as recommendations that will best describe the product opinion on which the purchase decisions can be made.

Table 2 shows the feature wise opinion ranking with average probability distribution of all 1011 reviews instances of the digital cameras.

Table 2 Probability based feature ranking

FNo.	Feature	Average Probability	Rank
Feature1	Resolution	23.06	1
Feature2	Memory,	19.26	3
Feature3	Flash type,	18.47	5
Feature4	Burst mode	18.81	4
Feature5	Optical zoom	20.40	2

As per the averaged probabilities of the feature wise opinions of the product, Resolution is with highest rank that best describes the product in terms of its importance.

Graph corresponds to opinion rank for the feature wise opinions are shown in Figure3.

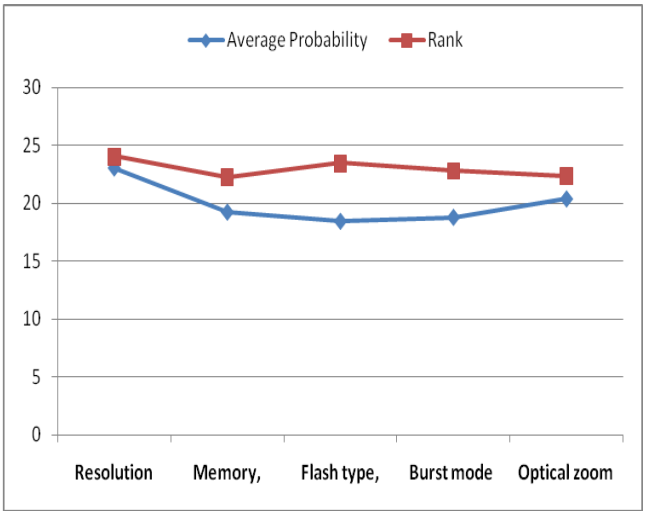

Figure 3 Opinion rank of the features corresonds to opinion documents

V. CONCLUSION

In this paper feature wise opinion summary is presented as product recommendation for different digital cameras as products based on selected features of shopper's interest. This work defines CIL_Classifier for generating document wise opinion summary from imbalance twitter multi class opinions. There after opinion segmentation is done followed by document ranking which directly maps each review opinion sentence into shopper selected features. This work is mostly appropriate and useful for those shoppers who are interested to purchase particular product using product reviews on the basis of selected features and also useful for vertical online shopping sites that sells only particular products.

REFERENCES

1. Salina Adinarayana, E Ilavarasan," A Two layer Feature Vector generation for OSIDL Classifier", in International Journal of Management, Technology And Engineering, Volume 8, Issue 11, Nov,2018
2. Daniel Billsus and Michael J. Pazzani, "Learning collaborative information filters", In Proceedings of the Fifteenth International Conference on Machine Learning (ICML-98), pp.46–54, Madison, WI, 1998.
3. G. Adomavicius, R. Sankaranarayanan, S. Sen, and A. Tuzhilin, "Incorporating contextual information in recommender systems using a multidimensional approach," ACM Transactions on Information Systems, Vol. 23, Issue 1, pp. 103– 145, 2005.
4. R. Jin, J. Chai, and L. Si, "An automatic weighting scheme for collaborative filtering," In Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval,pp. 337–344.,2004.
5. N. Jakob, et.al, "Beyond the stars: exploiting free-text user reviews to improve the accuracy of movie recommendations", In Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion, pp.57–64, ACM, 2009.
6. N. Pappas et.al.," Sentiment analysis of user comments for one-class collaborative filtering over ted talks" In SIGIR, pp. 773–776. ACM, 2013.
7. Burke, R,"Hybrid Recommender Systems: Survey and Experiments". In User Modelling and User-Adapted Interaction 2007.
8. Vivek Kumar Rangarajan Sridhar," Unsupervised topic modeling for short texts using distributed representations of words", In Proceedings of NAACL-HLT, PP.192–200, 2015.
9. Terdiman D,"Report: Twitter hits half a billion tweets a day",http://news.cnet.com/8301-1023_3-57541566-93/report-twitter-hits-half-abillion-tweets-a-day.
10. Blei D, Ng A, Jordan MI,"Latent Dirichlet allocation", Journal of Machine Learning Research, 2013.
11. Liu, S., Xia, C., and Jiang, X., "Efficient Probabilistic Latent Semantic Analysis with Sparsity Control", In Proceedings of IEEE International Conference on Data Mining, pp. 905-910,2010.
12. Hofmann, T., "Ünsupervised learning by probabilistic latent semantic analysisl, Machine Learning", Vol. 42, Issue 1, pp. 177- 196,2001.
13. Yookyung Jo, John E. Hopcroft, and Carl Lagoze," The Web of Topics: Discovering the Topology of Topic Evolution in a Corpus", In proceedings of The 20th International World Wide Web Conference, 2011
14. Sakaki, T., Okazaki, M., and Matsuo, Y, "Earth quake shakes twitter users: real-time event detection by social sensors", In Proceedings of the 19th international conference on World wide web, pp. 851–860. ACM,2010.
15. Kejun Huang, Xiao Fu, and Nikolaos D. Sidiropoulos," Anchor-free correlated topic modeling: Identifiability and algorithm". In Proceedings of NIPS, 2016.
16. Chong Wang, David M. Blei, and Fei-Fei Li.," Simultaneous image classification and annotation", In Proceedings of CVPR, 2009.
17. E. Erosheva," A thesis on Grade of membership and latent structure models with application to disability survey data", Carnegie Mellon University, Department of Statistics,2002.
18. Guduri Sulakshana, R Siva jyothi and Aluri Lakshmi," Detection of Hate Speech and offensive Language on Sentiment Analysis using Machine Learning Techniques", International Journal of Innovative Technology and Exploring Engineering (IJITEE), Vol.9 Issue5, March 2020.
19. Kumar R G and Shriram R," Sentiment Analysis using Bi-directional Recurrent Neural Network for Telugu Movies", International Journal of Innovative Technology and Exploring Engineering (IJITEE), Vol.9 Issue2, Dec 2019.
20. Salina Adinarayana, E Ilavarasan," Two stage Decision Tree Learning from Multi-class Imbalanced Tweets for Knowledge Discovery", International Journal on Recent and Innovation Trends in Computing and Communication", Volume: 5 Issue: 6, June 2017.
21. Salina Adinarayana, E Ilavarasan," An Efficient approach for Opinion Mining from skewed Twitter corpus using Under sampling approach", In Proceedings of "IEEE 2017 International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET), Bharath Institute of Higher Education and Research Chennai, India, pp. 1-4.,Feb 2017
22. Salina Adinarayana, E Ilavarasan," An Efficient approach for Opinion Mining from Skewed Twitter corpus using Over Sampled Imbalance Data Learning", 2017 International Conference on Intelligent Communication and Computational Techniques (ICCT), Manipal University, Jaipur, pp. 42 – 47,Dec 2017
23. Salina Adinarayana, E Ilavarasan," An Efficient Decision Tree for Imbalance data learning using Confiscate and Substitute Technique", Materials today: Proceedings, Vol 5,Issue.1,pp. 680-687,Jan 2018.