IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-4, April 2019 E-ISSN: 2347-2693

Feature Selection and Summarization of Customer Reviews Using Fitness Based BPSO

B.Suganya[1*], S.C.Lavanya[2], T.Gowrisankari[3]

[1,2,3]Dept. of Computer Science and Engineering, Dr.Mahalingam College of Engineering and Technology, Pollachi, India

[*]*Corresponding Author: suganyagovindharaaj@gmail.com, Tel.: +91-9791745051*

DOI: https://doi.org/10.26438/ijcse/v7i4.462467 | Available online at: www.ijcseonline.org

Accepted: 10/Apr/2019, Published: 30/Apr/2019

***Abstract*—** Significant growth of e-commerce has led to huge number of reviews for a product or service. It provides different aspects of service or a product for the users. Sentiment analysis techniques are used to extract feature and opinion in a concise summary form from the customer reviews. Feature based summarization system uses term frequency and feature opinion learner to generate the summary. Fitness value based binary particle swarm optimization for feature selection is proposed to select the best feature subset. The feature selection in BPSO uses fitness value based on the term frequency and opinion score. In BPSO efficient summary is generated using the multi-objective function based on feature weight score and similarity between term frequency and position. The Recall-Oriented Understanding for Gisting Evaluation (ROUGE) toolkit is used to measure the performance of the Multi objective fitness based BPSO. An experimental result proves that multi-objective FBPSO algorithm improves the feature selection and summary generation accuracy.

Keywords*—Feature selection, Multi-objective, Fitness, Binary Particle Swarm Optimization, Summarization*

I. INTRODUCTION

In a general web page, the reviews are written in natural language scheme and are free of texts with unstructured paradigm. In comparison, numerical and categorical data are well structured, which make them relatively easy to handle. On the contrary, customer reviews are unstructured data. To be handled, these data demand knowledge from different areas, e.g., database, information retrieval, information extraction, machine learning, and natural language processing. With the great and rapid growth of web contents, customer reviews become available where a customer is able to express opinions on products and services. This trend has seen increasingly attention in sentiment analysis or opinion mining. In the opinion mining community, there are many challenging research topics such as subjectivity classification, sentiment classification, and opinion summarization [1].

Subjectivity classification [2] is the task of classifying the sentences or the documents which contain opinions from factual. It is useful for many natural language processing applications such as question answering, information extraction, and so on. The task of sentiment classification is to judge whether a review expresses a positive or negative opinion. The systems assign a positive or negative sentiment for the whole review document. Sentiment classification is useful; it does not imply the underlining information, about the reviewer likes and dislikes. Opinion summarization is the task of producing a sentiment summary, which consists of sentences from reviews that capture the author's opinion. The summarization task is interested in features or objects on which customers have opinions. This is different from traditional text summarization that involves reducing a larger corpus of multiple documents into a short of paragraph that conveys the meaning of text.

Feature based summarization system analysis the customer reviews and considers only the subjective sentences. Subjectivity and objectivity determination [3] works in two phases – training and classification. For training phase, subjective clues are used as a trained data, and it is used later to identify the subjective unigrams for new dataset. In the second phase, the classification is performed based on the probability of unigrams from the test dataset using the training data. Then pre-processing of subjective sentences is performed, in which, only subjective sentences are submitted to a pipeline for Parts-Of-Speech (POS) tags. Then tags are applied to the subjective sentences in the dataset document to obtain the best feature set.

POS tagging [3] is used for sentence splitting and to assign lexical categories to the words in text. Maxent tagger from Stanford NLP is used for tagging the sentence. There are 36 tags in Maxent tagger. The system used 20 tags among 36 tags of Maxent tagger to get the features which express the

sentiment and also the opinion words which related to those words. Linguistic filtering pattern is used to extract the features and opinions. Relates the features and opinions and review summarization that aim to summarize customer reviews by selecting the informative review sentences are performed. Maximum entropy model is used to predict which feature word should be related with the opinion word with maximum probability. In order to use the maximum entropy to classify product feature-opinion candidates, use syntactic information to classify product feature-opinion pair.

The term frequency [4] of each feature is calculated and the top ranked features are selected based on the average value of the term frequency. Then for each of the top ranked features relative importance of each feature according to related opinion score is evaluated. ROUGE (Recall Oriented Understudy for Gisting Evaluation) is used as a performance metric to evaluate the quality of the summary generated.

Rest of the paper is organized as follows, Section I contains the introduction of Feature selection and summarization of reviews , Section II contain the related work of fitness based BPSO used in feature selection and multi objective based summarization process and , Section III explains the methodology used and architecture diagram of Fitness based Feature Selection and Summarization System, Section IV describes the results and discussion of the Feature selection and summarization using fitness based BPSO compared with the traditional probabilistic method, Section V concludes research work with future directions.

II. RELATED WORK

Naïve Bayes classifier [5] classifies the sentence as subjective if it contains any of the learned subjective patterns, and as objective if it contains any objective pattern. The initial training data used by the Naïve Bayes classifier was generated by the rule-based classifiers, which look for the presence or absence of a set of general subjectivity clues. It uses a greater variety of features than the rule-based classifiers and it exploits a probabilistic model to make classification decisions based on the combinations of these features. Naïve Bayes classifier is able to reliably label a different, and perhaps more diverse, set of sentences in the unlabeled corpus. Initially, self-training builds a single Naïve Bayes classifier using the labeled training data and all the features. Then it labels the unlabeled training data and converts the most confidently predicted document of each class into a labeled training data. The chosen sentences form a brand new training set that is used to retrain the Extraction Pattern (EP) Learner and then the naïve Bayes classifier.

The authors of [6] focused on aspect based opinion mining of customer reviews in an efficient way. Stanford-POS tagger to parse each sentence of review and yield the POS tags of each word (word is noun, adjective, verb, adverb, etc.). After POS tagging is done, extract features that are nouns or noun phrases using the pattern knowledge. For opinion words extraction, extract features that are used to find the nearest opinion words with adjective/adverb. To decide the opinion orientation of each sentence, three subtasks are performed. First, a set of opinion words (adjectives) is identified. If an adjective appears near a product feature in a sentence, then it is regarded as an opinion. Opinion words are extracted from the review using the extracted features. For identifying feature, both explicit and implicit features as a future work because both these features are useful for providing more accurate results in determining the polarity of feature before summarizing them.

In Lexicon based approach [6], entities and aspects were collected manually in a base list. Afterwards, this list was extended using the community-generated synonym lexicon. The extraction of the aspects is carried out as a simple search. Due to the fact that some aspects span over more words, the longest possible aspect phrase is taken. The linking of the opinion phrases to the aspects is done using a distance-based approach applied on the sentence-level. All strong positive or negative opinion phrases are linked to the next aspect found in a sentence according to the word position. The result is an opinion tuple giving the opinion phrase, the tonality (sn = strong negative, sp = strong positive) and the aspect itself. If more aspects than opinion phrases are found, the opinion phrase is linked to both aspects. If more opinion phrases than aspects are found, e.g., one aspect and two opinion phrases, only the nearest phrase is linked to the aspect.

In Binary Particle Swarm Optimization based Feature Subset Ranking for Feature Selection [7], each dataset is first divided into two sets: training set and a test set. K-Nearest Neighbor with n-fold cross-validation is employed to evaluate the classification accuracy in both of the training set and the test set, which are divided into n folds, respectively. If a dataset includes D features, D feature subsets will be evolved and ranked. The feature subsets search process starts from finding the best subset including 1 feature and ends with the feature subset with D features. The d^{th} feature subset includes d features, where d is a positive integer from 1 to D. There are many combinations for a feature subset with a particular number of features, use the dth feature subset to represent the best combination with d features in this method. In each dataset, the aim is to determine the number of top-ranked feature subsets that can achieve classification accuracy close to or even better than the classifier with all features. Feature subset ranking provides an effective way for feature selection. Using the same number of features, BPSO based feature subset ranking can achieve higher classification accuracy. This suggests that BPSO could be

used to find a subset of complementary features to improve the feature selection.

Fitness proportionate based feature subset selection method [8] is used to make the selected feature subset as powerful as possible, and aim to improve PSO's performance in discrete space. In order to achieve this, two major problems of PSO when the search space is discrete is considered and try to prove that the traditional way of calculating velocities is the main cause of the two problems. A new way to calculate velocities based on fitness values is then proposed and based on that a new binary version of PSO called FPSBPSO is proposed. To prove the efficiency of method by utilizing the method to perform the feature selection process in classification problems. If the selected feature subset with FPBPSO can return better classification results than that with traditional PSO, proposed scheme can be regarded as an efficient one.

PSO is used to develop a multi-objective feature selection algorithm, CMDPSOFS [11], which is based on the ideas of crowding, mutation, and dominance. CMDPSO has never been applied to feature selection problems. In order to address the main issue of determining a good leader (g_{best}), CMDPSOFS employs a leader set to store the non-dominated solutions as the potential leaders for each particle. A g_{best} is selected from the leader set according to their crowding distances and a binary tournament selection. Specifically, a crowding factor is employed to decide which non-dominated solutions should be added into the leader set and kept during the evolutionary process. The binary tournament selection is used to select two solutions from the leader set, and the less crowded solution is chosen as the g_{best}. The maximum size of the leader set is usually set as the number of particles in the swarm. Mutation operators are adopted to keep the diversity of the swarm and to improve the search ability of the algorithm. A dominance factor is adopted to determine the size of the archive, which is the number of non-dominated solutions that the algorithm reports. The solutions (feature subsets) in the final archive are used for classification on the test set in each data set.

Ahmed M. Al-Zahrani et.al [9] proposed particle swarm optimization (PSO) to evaluate the effectiveness of different state-of-the-art features used to summarize Arabic text. The PSO is trained on the Essex Arabic summaries corpus data to determine the best particle that represents the most appropriate simple/combination of eight informative/structure features used regularly by Arab summarizers. In all evolutionary computations, the choice of the fitness function is crucial, since the PSO evaluates the quality of each particle to move the solution space towards the optimized area. Thus, the function responsible for calculating and evolving the value of the quality for each particle is the fitness function. Therefore, the most important step in executing the PSO algorithm is to define a fitness function that can lead the swarm to the optimized solution based on the application and data by maximizing or minimizing the fitness function value.

Automatic text summarization aims to produce summaries for one or more texts using machine techniques. A novel statistical document summarization system [10] for Arabic texts is proposed. It uses a clustering algorithm and an adapted discriminant analysis method: MRMR (minimum redundancy and maximum relevance) to score terms. MRMR scores feature on the basis of how much discriminant information they hold. In summarization, there is a need of highly discriminant terms, which allow us to select a specific sentence and not the other sentences. For a term's relevance, it is the discriminating power of a specific term within classes, i.e., the more a term's frequency varies significantly through classes, the more it is discriminant. It is actually better to think about the opposite case; if a term has the same frequency mean (or is close to the term's frequency mean in all classes), it is not quite as interesting, and consequently, will receive a low relevance score.

Mohammed Salem Binwahlan et al. [11] investigate the effect of the feature structure on the features selection using particle swarm optimization. The particle swarm optimization is trained to learn the weight of each feature. The features used are different in terms of the structure, where some features were formed as combination of more than one feature while others as simple or individual feature. Therefore the effectiveness of each type of features could lead to mechanism to differentiate between the features having high importance and those having low importance. The combined features have higher priority of getting selection more than the simple features. In each of the iteration, the particle swarm optimization selects some features, then corresponding weights of those features are used to score the sentences and the top ranking sentences are selected as summary. The selected features of each best summary are used in calculation of the final features weights.

Summarization depends on the maximum relevance and minimum redundancy functions. Minimum redundancy and maximum relevance method [12] is proposed. To perform extractive summarization, sort sentences within a document and keep those that maximize relevance and, at the same time, cover up, at best, information contained in the source document. Measuring the relevance of a specific sentence is the main novelty in our proposition. For a term's relevance, it is the discriminating power of a specific term within classes. In fact, if two terms share very close relevance scores, it should be able to sort them on the basis of their redundancy scores. Term's redundancy is the mean of its mutual information with all other terms, i.e., shares an important amount of information with the rest of the terms.

If two terms share high mutual information, this shows how much one term attracts the other. Therefore, if a term has a high redundancy score, which means that it attracts many terms, this will reduce its discriminating power. Finally, the result is finding a set of terms, which describe the best sentences and, at the same time, does not attract many terms as the summary.

III. METHODOLOGY

In Particle Swarm Optimization based feature selection and summarization of customer reviews, the dataset document reviews are represented as a feature vector [during data preprocessing. After all the texts are transformed into corresponding feature vectors, Fitness based feature selection [13] of the reviews is performed. It is often seen in genetic algorithms. It selects the solution based on its fitness score on a particular task. The higher fitness score a solution has the more chances the solution will have of getting selected because the solution is considered to be more suitable for the task. The extracted features are assigned a feature weight score based on multi objective function of maximum relevance and minimum redundancy to generate a summary.

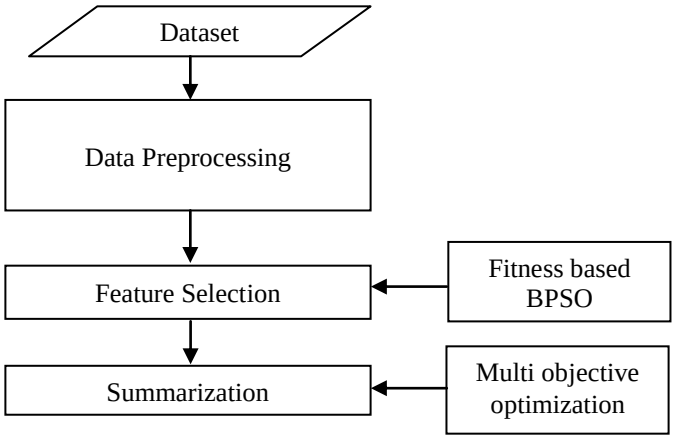

Figure 1 Block Diagram of Fitness based Feature Selection and Summarization System

A. *Data Preprocessing*

The review dataset document is collected. Data Preprocessing [14] of dataset document includes stop word removal, stemming and tokenization. Document is then transformed under a specific model and represented as a feature vector during data preprocessing. One commonly used model for document representation is unigram bag-of-words model (BoW). Under BoW model, number of dimensions of each feature vector is the number of different words in the whole text dataset. The vector assigns "1" to dth dimension if the text contains corresponding word, and assigns "0" if it does not.

B. *Feature Selection using fitness based BPSO*

Initialize N number of particles of the swarm randomly, a position of particle is denoted by xi velocity is denoted as v_i. Particle position is initialized as randomly in the search space. Set the particle's best known position and velocity to its initial value. The velocity and position is calculated as follows:

$$\text{Velocity, } v_i = \text{Max (term frequency of the particle) + Opinion score of the particle} \quad (1)$$

$$\text{Position, } x_i = \text{Min}(\text{Similarity}(\text{Max}_{\text{termfrequency}}, x_i) \quad (2)$$

Until a termination criterion is met, repeat the following: For every particle i=1, 2… N, Compute the fitness value, fitness function is $\max(v_i)$ and $\min(x_i)$. Based on the fitness score, the g_{best} and p_{best} value is updated. Return the best feature subset found by the swarm.

C. *Summarization Using Multiobjective Optimization*

Maximum Relevance and Minimum Redundancy are the two multi objective functions [15] used to generate a summary. To perform extractive summarization, sort sentences within a document and keep those that maximize relevance. If a term has the same frequency means, it is not quite as interesting, and consequently, will receive a low relevance score. In fact, if two terms share very close relevance scores, it should be able to sort them on the basis of their redundancy scores. Term's redundancy is the mean of its mutual information with all other terms. Therefore, if a term has a high redundancy score, which means that it attracts many terms, this will reduce its discriminating power. The fitness function for summarization depends upon the maximum relevance and minimum redundancy function and is given by,

$$\text{Fitness function} = \text{Max(feature weight score) + Min(Particle's position)} \quad (3)$$

Based on the fitness value, the top ranked features are generated as a summary.

IV. RESULTS AND DISCUSSION

A. *Data Collection*

The dataset comprises of customer reviews for hotels which are collected from Trip Advisor site. In this work, the hotel reviews of cities such as Beijing, London are used. Extracted fields include date and the full review.

B. Evaluation Metric

Evaluation metrics are important to identify the efficiency of the system. The performance of the system is calculated using the ROUGE-N metric [16].

Rogue-N Metric:

ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. It is essentially of a set of metrics for evaluating automatic summarization of texts as well as machine translation [18]. It works by comparing an automatically produced summary or translation against a set of reference summaries (typically human-produced). The ROUGE-N metric to evaluate the quality of a summary is given by the following formula,

$$\text{ROUGE-N} = \frac{S\in\{\Sigma_{ReferenceSummaries}\}\Sigma gram_{n\in S} Count_{match}(gram_n)}{S\in\{\Sigma_{ReferenceSummaries}\}\Sigma_{gram_{n\in S}} Count(gram_n)}$$

Where n stands for the length of the n-gram, i.e., is the maximum number of n-grams co-occurring in the candidate summary and the reference summaries. Recall, Precision and F-Measure in the context of ROUGE is used to get a good quantitative value. The formula for calculating these values is given by,

RECALL

Recall in the context of ROUGE simply means how much of the reference summary, the system summary is recovering or capturing.

$$\text{Recall} = \frac{\text{Number of overlapping words}}{\text{Total words in reference summary}}$$

PRECISION

Precision in the context of ROUGE is, how much of the system summary was in fact relevant or needed.

$$\text{Precision} = \frac{\text{Number of overlapping words}}{\text{Total words in system summary}}$$

F-MEASURE

F-Measure is the harmonic mean of precision and recall.

$$\text{F-Measure} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

C. PERFORMANCE EVALUATION

The experimental results for the Feature based summarization system and Fitness based BPSO system is measured using ROUGE metric.

Table 1. Feature based summarization system

FEATURE	ROUGE 1 METRIC		
	PRECISION	RECALL	F-MEASURE
hotel	0.2770	0.8849	0.4219
room	0.4832	0.5982	0.6793
staff	0.3328	0.5250	0.6563
location	0.2589	0.3956	0.3130
price	0.4661	0.7534	0.5759

Table 2. Fitness based BPSO for summarization

FEATURE	ROUGE 1 METRIC		
	PRECISION	RECALL	F-MEASURE
hotel	0.5797	0.7534	0.7820
room	0.5148	0.8144	0.7590
staff	0.5693	0.7932	0.7432
location	0.4920	0.8918	0.6595
price	0.4380	0.7260	0.5463

The analysis has been made on those results and the following inference is obtained. The performance value of the ROUGE metric shows that the fitness value based BPSO performs better than the traditional probabilistic based feature selection system. The accuracy of the summary generated is enhanced in the fitness value based BPSO system.

V. CONCLUSION AND FUTURE SCOPE

In this work, the summarized output is generated by adapting the Binary Particle Swarm Optimization algorithm. Based on the fitness function, it generates the best feature subset found by the swarm. Multi objective optimization based on maximum relevance and minimum redundancy for summarization can select most optimal set of sentences that represents the important information of the whole documents. This method also used to reduce the computational time in the case of multiple documents with more sentences. It has been observed that the ROUGE-N metric has been used to evaluate both the existing and the proposed system. Multi objective optimized summary using Binary Particle Swarm Optimization improves the performance of summary when compared to probabilistic ranking approach. The experimental tests conducted prove that the Fitness based BPSO approach leads to the better result in terms of improved accuracy of summarized content.

The fast convergence of particle swarm algorithms can become a downside in multi-objective optimization problems when there are many local optimal fronts. In such a situation a multi-objective particle swarm algorithm may get stuck to a

local Pareto optimal front. Future research will be on choosing different leader selection approach which presents the better results in terms of convergence and diversity in multi-objective scenarios. And selection of different initialization schemes for generating the swarm population to improve the performance of PSO.

REFERENCES

[1] ArtiBuche, Dr.M.BChandak, Akshay Zadgaonkar, "Opinion Mining and Analysis: A Survey", Proceedings of the International Journal on Natural Language Computing, Volume 2, No. 3, pp. 39-48, 2013.

[2] Gamgarn Somprasertsri, Pattarachai Lalitrojwong, "Mining Feature-Opinion in Online Customer Reviews for Opinion Summarization", Journal of Universal Computer Science, Vol.16, No.6, pp. 938-955, 2010.

[3] Dim En Nyaung, Thin Lai Lai Thein, "Feature-Based Summarizing and Ranking from Customer Reviews", International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol: 9, No: 3, 2015.

[4] Li-Ping Jing, Hou-Kuan Huang, Hong-Bo, "Improved Feature Selection Approach TFIDF in Text Mining", Proceedings of the First International Conference on Machine Learning and Cybernetics, Vol. 2, 2002.

[5] J. Wiebe, E. Riloff, "Creating Subjective and Objective Sentence Classifiers from Unannotated Texts", In Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing-10), Vol: 3406, pp. 486-497, 2010.

[6] Florian Wogenstein, J. Drescher, D. Reinel, S. Rill, J. Scheidt, "Evaluation of an Algorithm for Aspect-Based Opinion Mining Using a Lexicon-Based Approach", WISDOM '13, Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining, Article No. 5, 2013.

[7] Bing Xue, Mengjie Zhang and Will N. Browne, 'Single Feature Ranking and Binary Particle Swarm Optimisation Based Feature Subset Ranking for Feature Selection', Proceedings of the Thirty-Fifth Australasian Computer Science Conference, Melbourne, Australia, Vol.122, ACS, pp. 27-36, 2012.

[8] Zhou Z, Liu X, Li P, Shang L, "Feature selection method with Proportionate Fitness based Binary Particle Swarm Optimization", In: Simulated evolution and learning, pp. 582–592. Springer, New York, 2014.

[9] Ahmed M. Al-Zahrani, Hassan Mathkour, Hassan Abdalla, "PSO-Based Feature Selection for Arabic Text Summarization", Journal of Universal Computer Science, Vol. 21, No.11, pp. 1454-1469, 2015.

[10] Rasim M. Alguliev, Ramiz M. Aliguliyev, Nijat R. Isazade, "MR&MR-SUM: Maximum Relevance and Minimum Redundancy Document Summarization Model", International Journal of Information Technology and Decision Making, Vol.12, No.3, pp. 361-393, 2013.

[11] Mohammed Salem Binwahlan , Naomie Salim2 , Ladda Suanmali, "Swarm Based Features Selection for Text Summarization", International Journal of Computer Science and Network Security, Vol.9, No.1, 2009.

[12] Houda Oufaida, Omar Nouali, Philippe Blache, "Minimum Redundancy and Maximum Relevance for Single and Multi-document Arabic Text Summarization", Journal of King Saud University – Computer and Information Sciences, Volume 26, Issue 4, pp. 450–461, 2014.

[13] B.Suganya, V.Priya, "Particle Swarm Optimization Based Feature Selection and Summarization of Customer Reviews", International Conference on Emerging trends in Engineering, Science and Sustainable Technology, pp. 131-135, 2017.

[14] Lin Shang, Zhe Zhou, Xing Liu, "Particle Swarm Optimization-based Feature Selection in Sentiment Classification", Journal of Soft Computing – A Fusion of Foundations, Methodologies and Applications, Vol.20, Issue.10, pp. 3821-3834, 2016.

[15] Bing Xue, Mengjie Zhang and Will N.Browne, "Particle Swarm Optimization for Feature Selection in Classification: A Multi-Objective Approach", IEEE Transactions on Cybernetics, 43(6), pp. 1656-71, 2012.

[16] Josef Steinberger, Karel Jezek, "Evaluation Measures For Text Summarization", Computing and Informatics, Vol. 28, pp. 1001–1026, 2009.

Authors Profile

Ms. B.Suganya completed her Bachelor of Engineering in 2013 and Master of Engineering in 2017. She is specialized in Computer Science and Engineering from Anna University Chennai. She is currently working as Assistant Professor in the department of Computer Science and Engineering at Dr. Mahalingam College of Engineering and Technology, Pollachi, Coimbatore, TamilNadu, India. She is a Life time member of ISTE, since 2017. Her research interests are Data Mining and Human Computer Interaction.

Ms. S.C.Lavanya completed her Bachelor of Engineering in 2009 and Master of Engineering in 2011. She is specialized in Computer Science and Engineering from Anna University Chennai. She is currently working as Assistant Professor in the department of Computer Science and Engineering at Dr. Mahalingam College of Engineering and Technology, Pollachi, Coimbatore, TamilNadu, India. She is a Life time member of ISTE, since 2011. Her research interests are Data Mining and Human Computer Interaction.

Ms T.Gowrisankari completed her Bachelor of Engineering in 2013 and Master of Engineering in 2015. She is specialized in Computer Science and Engineering from Anna University Chennai. She is currently working as Assistant Professor in the department of Computer Science and Engineering at Dr. Mahalingam College of Engineering and Technology, Pollachi, Coimbatore, TamilNadu, India. She is a Life time member of ISTE, since 2015. Her research interests are Data Mining, IOT, Big data and Analytics and Human Computer Interaction.

