

Feature Extraction and Feature Selection process in Authorship Identification for Tamil Language

A. Pandian, R Ragavi, V.V.Ramalingam

Abstract – The concept of authorship identification and stylometry analysis have fascinating issues to be dealt with. Articles framed by various authors can be classified by measuring attributes linked to literary style along with attribute authorship related to texts that are newly explored. This forms a necessary function in different fields like psycholinguistics, cybercrime investigation, political socialization, etc... Numerous classical variations of statistical methods are presented and imbibed by the literary style. In the approach of Text processing outstanding information is fetched from the Tamil dataset combining quantifiable parameters from it. In the case of unidentified authors, it becomes tedious to classify the poems. The existing paper suggest that poem or text that is unidentified can be retrieved by categorizing potential author's earlier work and organizing the unfamiliar text or poem in Tamil language by constructing a classifier. In the proposed approach various stages include: feature extraction and selection utilizing decision tree. Whole process is split up into two - training and testing. All the known poets are organized within training dataset – contains 5 authors each 80 poems whereas the unknown poets are categorized under testing dataset. Through imbibing this methodology different poetic authors can be identified within the Tamil vernacular, resulting in valuable contribution to the society. The work considers a persons or authors distinctive style of writing, computes the features relativity. The method proposed is highly effective as demonstrated in the outcome of experiments performed on the actual dataset. The proposed techniques of decision tree effectively yields higher functionality in comparison with other existing approaches.

Keywords: Classification, Tamil Articles, Feature Extraction, Feature Selection, stylometry, Training dataset, Testing dataset, Authorship.

I. INTRODUCTION

The term Text mining alternatively known as text data mining, somewhat related to text analytics, is a method where information of superior quality is extracted from the text. The various levels or stages in Text mining includes: framing the input text by parsing, appending few derived linguistic features and discarding other, eventually inserting in the database, by utilizing structured data, deriving various strategy of Authorship identification incorporates text analysis so that the original author is being identified amidst various groups of candidate authors. The principle on which authorship attribution relies is as stated: considering the training data which includes set of texts of a known author, to find the author of the unchecked text (texts denotes testing data) the anonymous text is matched to a single author of the candidate set.

Revised Manuscript Received on April 30, 2019.

* Correspondence Author

Dr.A.Pandian¹, department of computer science, SRMIST, SRM University, Chennai, India. Email: pandiana@srmist.edu.in

R Ragavi², department of computer science, SRMIST, SRM University, Chennai, India. Email: ragavi_r17@srmuniv.edu.in

Dr.V.V.Ramalingam³, department of computer science, SRMIST, SRM University, Chennai, India. Email: ramalinv@srmist.edu.in

Consider a poem with unknown author the task is to trace the author to which the Tamil poem belongs within the available features list of every candidate authors. This research of Authorship identification in poems is altogether a new approach and not being much worked upon as in Tamil language. During the research formation, it was affirmed with intense knowledge that no research and published work concerning authorship attribution exists in Tamil poems. Tamil poems are being inspected for research purpose of classification task. In the approach of Text processing outstanding information is fetched from the dataset combining quantifiable parameters from it. The existing paper suggest that poem or text that is unidentified can be retrieved by categorizing potential author's earlier work and organizing the unfamiliar text or poem in Tamil language by constructing a classifier. In the proposed approach various stages include: feature extraction and selection utilizing decision tree and feature classification by imbibing Naives' Bayesian technique. The dataset are used for fetching various features in order to carry out categorization. The features being extracted are: lexical, syntactic and semantic. Lexical parameters include noun, pronoun, verb and adjective. Syntactic features being noun phrase, verb phrase and prepositional phrase. Semantic features incorporates a group of features that boost the words meaning and make it more intense. For deriving contributing and non-contributing features within the dataset the method of Feature selection is performed by utilizing the decision tree. Decision tree construction is based on the attribute that contains the highest information gain. The issues pertaining to computational cost and imprecise classifier accuracy due to unrelated data can be rectified using the approach of Feature selection. The Naives' Bayesian algorithm conducts classification which can be split up into two types viz training and testing. Authorship identification relies on the following sources:

- mount of reputed authors should form a well-defined set.
- o highlight the authors linguistic habits linked to disputed text there must exist adequate lengths of the writing.
- or comparing the texts utilized must be in proportion to the disputed writing.

In authorship identification the impact of word sequences is being analyzed. Both the topic and stylistic text features are observed by the researchers. The work presents authorship identification by incorporating set of word sequences which merge content words and functions. Experiments are conducted with Naïve Bayesian classifier on a poem based dataset. The journal classification is stated as: Section 2 includes working of earlier author.



Section 3 portrays the suggested decision tree and outlook of various levels. Section 4, demonstrates experimental outcome. At last, Section 5 concludes presents future research work thereby concluding the paper.

II. RELATED WORKS

Al-Falahi Ahmed et.al, performs authorship attribution task by proposing the Arabic poetry. The method of Markov Chains is being imbibed which considers multiple features like Sentence length, Rhyme, Characters, Word length and Initial word of the sentence as input fields. The experimental data set is split up into two categories: training dataset containing known authors and test dataset having unknown authors [1] [2].

Ramdani Mohamed et.al, possess authorship attribution linked to Arabic poetry utilizing method of machine learning. The text mining classification algorithms namely Naïve Bayes NB and Support Vector Machine SVM utilizes input fields in terms of poetry features like Poetry Sentence length, Characters, Rhyme, Word length, Meter and Initial word of the sentence. To address and resolve the issue of determining who the poet of the available unknown is text the strategy of style markers is imbibed to detect the author [3].

Mohammad A. Alnagdawi et.al have detected the poem meter name (called Bahar in Arabic) by Arud science, which offers a strategy to categorize Arabic poems into 16 meters, aiding the user in tracing out meter name concerning any Arabic poem by making use of context free grammar (CFG). Right from the initial phase to end results the remedies linked to the problems are examined applying regular expression and CFG [4] [5].

Steven H. H. Ding proposed in his work, The authorship analysis (AA) which includes study of revealing the invisible attributes of authors from available textual data. Relating to writing styles reflected in the text the author's identification and sociolinguistic features are fetched. The performance of the mentioned approach is examined on issues concerning authorship characterization, authorship identification and authorship verification in respect to blog, Twitter, review, novel, and essay datasets [6][7][8].

A. Pandian has fetched Tvarious features from Mukkoodar Pallu's poems and thereby proposed the training classifier. Classification of authors can be done on different unknown poems. By implementing C4.5 algorithm classification accuracy approach Fis conducted for dataset classification. Varied characteristics are depicted namely number of characters, number of sentences and the classification accuracy when C4.5 algorithm is imbibed [9].

Michael R. Schmid et.al Usage of e-mail communication is usually criticized for engaging in socio-technology attacks likely phishing, spamming, spreading malware and theft of personal identity. The model that is build up aids in detecting the most credible author of the text. Regrettably, many current researches merely pay attention to improvise predictive accuracy not bothering about intrinsic data gathered from the evidence. [10] [11].

Pandian. A et.al, points out the necessity of e-mail for quick and prompt communication. Altered or forged information are mailed to multiple users that looks like genuine. Massive dimensional signature is transformed into a 2D Pattern with help of Fisher's Linear Discriminant Function (FLD) thereby making further processing convenient. The Radial Basis Function (RBF) network

considers these signatures in 2D pattern for training data to later comprehend linear data [12].

R. Lakshmi Priya et.al, proposes authorship attribution for articles linked to ambiguous authorship forming as one key application domain of Stylometry. Here unknown authorship articles or writings are considered for classification in contrast with the articles framed by the modern Tamil scholars in the same time period which are Mahakavi Bharathiar (MB), Subramniya Iyer (SI), and T. V. Kalyanasundaranar (TVK). Principal Component Analysis (PCA) and Multivariate Discriminant Analysis (MDA) applications are taken into account for discussion [13] [14].

Ahmed Fawzi Otoom et.al, emphasizes upon constructing a brilliant methodology that has the ability to categorize a new available article amidst any one of the seven classes which are designated to seven types of authors. To implement this, a novel dataset is proposed containing 12 features and 456 instances associated to the 7 authors. Moreover, the feature set suggested is merged with robust classification algorithms that aids in differentiating amidst various authors [15].

Ahmed Alwajeih et.al, assumes the issue of authorship authentication which tends to a traditional issue of concern in linguistics which has turned out to be tedious enough with widespread usage of Internet, expansion and rise in the quantity of unchecked texts and difficulty in examining claims that are published online. The Arabic language is targeted in which this issue is rarely examined instead of it being important. [16].

Navinder Kaur et.al, stated the authorship attribution as the issue of author identification of an unknown text. Based on the approach of machine learning, Authorship Attribution can be portrayed as a multiclass, single-label text-categorization method. Numerous attributes acts as the input values like word ngram, character ngram for the linear SVM (Support Vector Machine) classifier thus computing the validity of the suggested system relying upon Precision, Recall, F-score and Accuracy [17].

G. Manimannan et.al, proposed of an effort towards attribute authorship relying on the stylistic attributes of specific articles formulated on Indian freedom movements and being published in the India magazine. This affirms that through Stylometry evidence, unknown articles are placed attributed to known articles. The control article being secluded from known and unknown articles. Based on Bharathiar writings two unattributed articles can be linked to it [18]. Sanjanasri J.P presents and states that with the rampant rise in World Wide Web, the regional language contents in form of web pages, e-books, e-mails, and digital repositories is effortlessly obtainable, approachable and has tremendously gain its roots. Resultant, the information can be retrieved amidst massive web documents using the prominent method of automatic document classification. NLP applications such as information extraction, query response, information summarization. Numerous NLP applications namely information extraction, query response, information summarization etc. works on the thumb rule of text classification [19] [20].

Anderson Rocha et.al precisely predicts the author's identification in sites like Twitter where the message may not go beyond 140 characters. Mostly every author exhibits a particular habit of writing and creating their own content. Using machine learning algorithms such features are being evaluated [21].

Sadia Afroz proposes the method of stylometry which analyzes the style of writing, for identification of unknown authors pertaining to unknown text. This technique haveyield better outcome in earlier tests yet their functionality needs to be tested on a demanding dataset pertaining to the area of interest of the security researchers. [22] [23].

III. PROPOSED WORK

3.1 Project Overview

Stylometry is a method which analyzes the style of writing, for identification of unknown authors pertaining to unknown text. Furthermore this approach can be spread over to nearly all regional languages globally. Classification of poems having anonymous author tends to be tedious as found by many literary reviewers. Through imbibing this methodology different poetic authors can be identified within the Tamil vernacular, resulting in valuable contribution to the society. In the proposed approach: feature extraction and selection utilizing decision tree. To differentiate amidst multiple authors Naïve Bayes classifier algorithms is merged alongside proposed feature set. The test outcome reveals that the dataset successfully yields in higher classification performance accuracy.

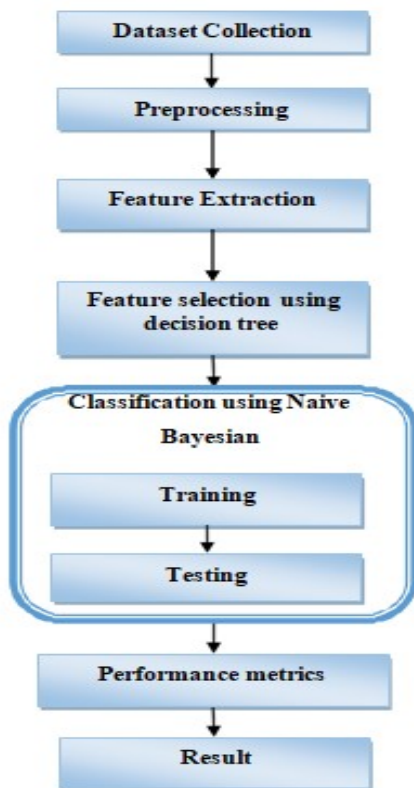


Fig.1 Proposed Overall Architecture

3.2 Dataset Collection

It withholds different poem sets that is gathered from various standardized articles. Write ups penned down by potential authors are gathered from plentiful resources and then computerized. Poems are being assembled pertaining

to numerous Tamil articles. The database consists of nearly every type of poem. Required data from the dataset is being retrieved and further passed on for processing to the data preprocessing module. Around 100 poems of each mentioned authors are collected: Ilango Adigal, Ottakoothar, Bharathidasan, Chinnaswami Subramaniya Bharati, Kumara Gurubarar. The prepared collection of poetry texts is utilized for training data, and the left over are allotted to testing data. That is each author having 100 poems thus 80 poems from dataset is considered for training and remaining 20 poems from dataset is considered for testing.

3.3 Preprocessing

It withholds compilation of diverse well known poet's texts for examination, derived from poetry websites and encyclopedias. The selection of poets belonging to different eras was done randomly. Thereafter the processing included, filling up the missing text, smooth noisy text, detect or eliminate the outliers, and rectify inconsistencies. Source system withholds unclean data hence Data cleaning becomes a valid requirement. Presence of alphanumeric and punctuation is not common in these sort of poetries hence nearly all such poetic texts undergoes initialization process that is: stripping of all punctuation, strip and alphanumeric.

3.4 Feature Extraction

The keyword Feature extraction is incorporated in computers and machine learning. Concurrently with text processing, feature extraction initializes with set of measured data and building a list of derived values that proves as essential and informative. The classifier can be constructed relying upon the features retrieved from the dataset. The features present in the feature set includes lexical, syntactic and semantic. Lexical features covers frequency of N grams, punctuation and special characters. Syntactic features encompass frequency of language-specific parts-of-speech and function of words. Semantic features incorporate a group of features that boost the words meaning and make it more intense.

Feature list:

Uyir(12) - அ, ஆ, இ, ஈ, உ, ஊ, எ, ஏ, ஐ, ஒ, ஓ, ஔ,

Mei(18) -க், ங், ச், ஞ், ட், ண், த், ந், ப், ம், ய், ர், ல், வ், ழ், ள், ற், ன்

Uyirmei(216) -க, ங, ச, ஞ, ட, ண, த, ந, ப, ம, ய, ர, ல, வ,

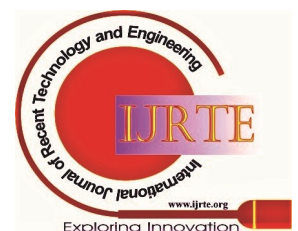
ழ, ள, ற, ன, கா, ஙா, டா, ணா, தா, நா, பா, மா, யா, ரா,

லா, வா, ழா, ளா, றா, னா, கி, சி, ஞி, டி, ணி, தி, நி, பி, மி,

யி, ரி, லி, வி, ழி, ளி, றி, னி, கீ, நீ, சீ, ஞீ, டீ, ணீ, ி, தீ, நீ, பீ,

மீ, யீ, ரீ, லீ, வ, ி, ழீ, ள, ி, றீ, ன, ி, கு, ஙு, சு, னு, டு,

னு, து, று, பு, மு, யு, று, லு, வு, ழு, ளு, று, னு, கூ, னு, ஞு, ஞா, ஞு, னு,



2. **False Negative (FN):** If instance being classification result positive it's classified as negative.
3. **True Negative (TN):** I If instance being classification result negative it's classified as negative.
4. **False Positive (FP):** I If instance being classification result negative it's classified as positive.

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (Y_i - R_i)^2} \quad (1)$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (2)$$

TP, TN, FN, FP are considered as four attribute values fetched from classification and for examining outcome.

IV. RESULT AND DISCUSSION

The experiments concerns authorship identification related to Tamil poetry for which a set of texts framed by Tamil poets are presented. Various testing features includes characters, sentences length, words length, rhyme, first-word in sentences and meter. The proposed techniques of decision tree is effective and yielding higher functionality in comparison with other existing approaches.

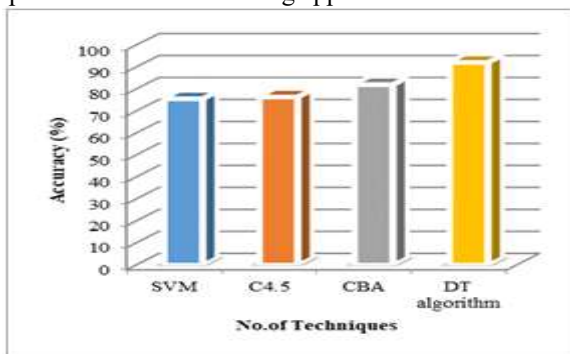


Fig 5: Performance of Techniques based Feature selection

Fig.4 mentioned above depicts Feature selection used in Decision tree algorithm comparing with several techniques like Support Vector Machine (SVM) algorithm, C4.5 algorithm, CBA (Classification Based Associations) algorithm. Feature selection results in high and effective accuracy performance compared to rest of the techniques.

Table 1: Performance of classification techniques

S.No	No. of Techniques	Accuracy (%)	Processing Time (ms)
1	Support Vector Machines (SVM)	79.3	0.88
2	Functional Tree (FT)	82	0.72
3	SEE 5	93.5	0.65
4	Naïve Bayesian algorithm	96.48	0.50

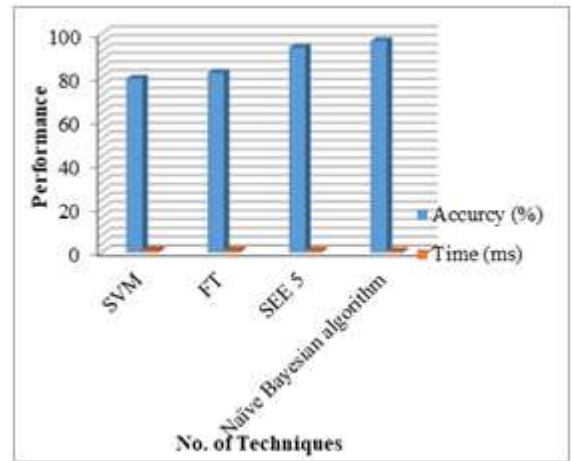


Fig 6: Performing of Techniques based Classification

Fig.5 mentioned above, Author identification is implemented in naïve Bayesian algorithm comparing with different techniques like Support Vector Machine (SVM) algorithm, Functional Tree (FT) algorithm, SEE 5 algorithm. The Author Identification results in high and effective accuracy performance compared to rest of the techniques.

V. CONCLUSION

The issue concerning authorship attribution is focused in the research paper, which is assumed as a practical concept in various aspects of information science research. Impact of textual classification has been analyzed related to discrimination of documents penned down by unique authors. An approach is designed for a methodology to gather additional information pertaining to author's peculiar styles. The outcome portrayed in the research work is backed by the rule optimization procedure conducted on the interested dataset focusing on to resolve the issue concerning a particular authorship attribution. Consequently, by creating a subsequent large database, with standardized sections, which is hierarchically stored, retrieving general attributes in Tamil language, gives a complete authorship identification system.

REFERENCES

1. Al-Falahi Ahmed, Ramdani Mohamed, Bellafkih Mostafa, Al-Sarem Mohammed "Authorship Attribution in Arabic Poetry", 2015, IEEE.
2. Bhargava Urala K, A G Ramakrishnan, Sahil Mohamed "Recognition of open vocabulary, online handwritten pages in Tamil script", 2014 IEEE.
3. Al-Falahi Ahmed, Ramdani Mohamed, and Bellafkih Mostafa "Machine Learning for Authorship Attribution in Arabic Poetry", International Journal of Future Computer and Communication, Vol. 6, No. 2, 2017.
4. Mohammad A. Alnagdawi, Hasan Rashideh, Ala' Fahed Aburumman "Finding Arabic Poem Meter using Context Free Grammar", Journal of Commun. & Comput. Eng, Modern Science Publishers, Vol. 3, Issue 1, 2013, P.P. 52 – 59.
5. Mahmoud Khonji, Youssef Iraqi, Andrew Jones "An Evaluation of Authorship Attribution Using Random Forests", International Conference on Information and Communication Technology Research (ICTRC 2015), ©2015 IEEE, p.p. 68 – 71.

Feature Extraction and Feature Selection process in Authorship Identification for Tamil Language

6. Steven H. H. Ding , Benjamin C. M. Fung , Senior Member, IEEE, Farkhund Iqbal, and William K. Cheung “Learning Stylometric Representations for Authorship Analysis”, IEEE TRANSACTIONS ON CYBERNETICS, 2017.
7. F. Howedi and M. Mohd, “Text classification for authorship attribution using naive bayes classifier with limited training data,” Comput. Eng. Intell. Syst., vol. 5, no. 4, 2014, p.p. 48 – 57.
8. Jafar Albadameh, Bashar Talafha, Mahmoud Al-Ayyoub, Belal Zaqabeh, Mohammad Al-Smadi, Yaser Jararweh and Elhadj Benkhelifa “Using Big Data Analytics For Authorship Authentication of Arabic Tweets”, IEEE/ACM 8th International Conference on Utility and Cloud Computing, © 2015, IEEE, p.p. 448 – 452.
- a. Pandian, V. V. Ramalingam and R. P. Vishnu Preet “Authorship Identification for Tamil Classical Poem using C4.5 Algorithm”, Indian Journal of Science and Technology, Vol. 9, No. 47, 2016.
9. Michael R. Schmid , Farkhund Iqbal, Benjamin C.M. Fung “E-mail authorship attribution using customized associative classification”, Published by Elsevier Ltd, Digital Investigation 14, 2015, p.p. 116 – 126.
10. Siddharth Swain, Gaurav Mishra and C. Sindhu “Recent Approaches on Authorship Attribution Techniques - An Overview”, International Conference on Electronics, Communication and Aerospace Technology (ICECA 2017), ©2017, IEEE, p.p. 557 – 556.
11. Pandian, A. and Md. Abdul Karim Sadiq “Authorship Categorization In Email Investigations Using Fisher’s Linear Discriminant Method With Radial Basis Function”, Journal of Computer Science, Vol. 10, No. 6, p.p. 1003-1014, 2014.
12. R. Lakshmi Priya, G. Manimannan “A Study of Ambiguous Authorship in Tamil Articles using Multivariate Statistical Analysis”, International Journal of Computer Applications ,Volume 86 – No 1, January 2014.
13. Pandian A , Mohamed Abdul karim Sadiq “Innovative Methods in Identifying Authors of Documents”, International Journal of Engineering and Technology (IJET), Vol. 6, No. 6, 2015, p.p. 2512 – 2520.
14. Ahmed Fawzi Otoom, Emad E. Abdullah, Shifaa Jafer, Aseel Hamdallah, Dana Amer “Towards Author Identification of Arabic Text Articles”, 5th International Conference on Information and Communication Systems (ICICS), ©IEEE, 2014.
15. Ahmed Alwajeeh, Mahmoud Al-Ayyoub, and Ismail Hmeidi “On Authorship Authentication of Arabic Articles”, International Journal of Computer Applications, 2014.
16. Navinder Kaur, Amandeep Verma “Authorship Attribution of Punjabi Poetry using SVM Classifier”, Computer Science and Software Engineering, Volume 5, Issue 5, 2015,p.p. 1055 – 1061.
17. G. Manimannan and R. Lakshmi Priya “Identification of Disputed Writings in Tamil Articles Using Multivariate Statistical Techniques”, IOSR Journal of Mathematics (IOSR-JM), Volume 10, Issue 2, 2014, PP 01-07.
18. Sanjanasri J.P and Anand Kumar M “A Computational Framework for Tamil Document Classification using Random Kitchen Sink”, 2015 IEEE, p.p. 1571 – 1577.
19. Jeaneth Machicao , Edison A. Corrêa, Jr. , Gisele H. B. Miranda, Diego R. Amancio, Odemir M. Bruno “Authorship attribution based on Life-Like Network Automata”, PLOS ONE, 2018, p.p. 1 – 21.
20. Anderson Rocha, Walter J. Scheirer, Christopher W. Forstall, Thiago Cavalcante, Antonio Theophilo, Bingyu Shen, Ariadne R. B. Carvalho, and Efstathios Stamatatos
21. “Authorship Attribution for Social Media Forensics”, IEEE, VOL. 12, NO. 1, 2017, p.p. 5 - 33.
22. Sadia Afroz, Aylin Caliskan-Islam, Ariel Stoleran, Rachel Greenstadt and Damon McCoy “Doppelgänger Finder: Taking Stylometry To The Underground”, 2014, IEEE Symposium on Security and Privacy, p.p. 212 – 226.
23. Haifa Alharthi, Diana Inkpen, Stan Szpakowicz “Authorship Identification for Literary Book Recommendations”, 27th International Conference on Computational Linguistics, 2018, p.p. 390 - 400.

AUTHORS PROFILE



Dr. A. Pandian received his MCA degree from Bharathidasan University, Tiruchi. He received his M.Tech degree from Punjabi University, Patiala, Punjab and M.Phil. degree from Periyar University, Salem. He has completed Ph.D.(Computer Science & Engineering) in SRM Institute of Science and Technology , Chennai. He has over twenty three years

of experience in teaching. He is working as Associate Professor in the Department of Computer Science & Engineering), SRM IST, Chennai. His

areas of interest are text processing, information retrieval and machine learning. He is a member of ISTE,IAENG, IACSIT and ISC. He have published more than thirty papers in many international conferences and refereed journals of repute. Also, he filed four patents in the Intellectual Property Rights of India.



R. Ragavi, completed her B.E.(CSE) in Valliammai Engineering college , Chennai and M. Tech (CSE) in SRM Institute of science and technology, Chennai. She has the sound knowledge in OOPS, Data Structures, Machine learning, Python, C#. Also, she has done value added course in Data Structures.



Dr. V. V. Ramalingam received his MCA degree from Bharathidasan University, Tiruchi and M.Phil degree from Periyar University, Salem .He received his M.Tech(CSE) and Ph.D.(CSE) in SRM Institute of Science and Technology , Chennai. He has over nineteen years of experience in teaching. He is working as Associate Professor in the Department of Computer Science & Engineering), SRM IST, Chennai. His areas of interest are information retrieval and machine learning. He is a member of ISTE and ISC. He have published more than thirty papers in many international conferences and refereed journals of repute. Also, he filed four patents in the Intellectual Property Rights of India.