# Feature Selection using Normalized Weight Method for Tamil Text Classification

**N. Rajkumar, T.S. Subashini, K. Rajan, V. Ramalingam**

***Abstract*: The Feature Selection process simplify the Tamil text classification work at present we are in the information age, in this period all the applications has great growth in the domain of World Wide Web, so regional language like Tamil materials such as web pages, e-mails, e-books, and digital data has grown enormously so the retrieval of this Tamil digital document is more wanted by Tamil Document searcher. For quick retrieval of needed Tamil digitized documents among the millions of Tamil web documents, these documents should be classified by content according to their classes. The Tamil Text classification is a background work for many Tamil NLP applications such as query response, information extraction, information summarization, etc. the implementation of text categorization is very important in the information retrieval field. The text categorization assigns a document an appropriate category from a predefined group of categories. Tamil Text Classification classifies the documents based on Tamil text in a Document. Tamil language words are very rich in morphology and hence Tamil language consists of very large set of word forms. So it is important to reduce the features of Tamil text. This paper discusses about Feature selection Using Normalized weight from the huge set of key words from the preprocessed corpus. The Feature selection done by Term Weighting (TF*IDF) normalized method is reducing the size of the key word list which is very useful for training and testing Tamil text classification algorithms.***

***Keywords: Stop word, Stemming, Feature Selection, Text mining, Text Classification, NLP***

## I. INTRODUCTION

**N**owadays, as online as well as offline text data is increasing drastically, the need to extract only the required information from this huge volume of data is increasing rapidly this has given rise to a new field namely Text Mining which analyses natural language text to extract needed useful information for a specific purpose. In comparison to numerical data, text is an unstructured in format, ambiguous, and more difficult to manipulate and process. Text mining includes text summarization, text classification and text clustering. The primary goal of Text classification is documents are classified into categories or classes which already defined. The sample of text could be a document, email, search query, news article, customer feedback, tweet, user product review etc. Applications of text classification include categorizing news article contents into topics, and organizing web pages content into hierarchical categories, email, and sentiment analysis, filtering spam, support tickets, predicting user intent from search queries, and analyzing customer feedback.

## II. LITERATURE REVIEW

Literature related to Preprocessing stop-word removal study in[1]. Authors [2][3] have made a list of Gujarati language stop words by creating a frequency list from a Gujarati corpus. The authors in [4] proposed algorithm for stop word for Sanskrit language an in [5] for Hindi language[6] for Arabic language. In [7] proposed an aggregated method for automatically building key word stop-word lists in information retrieval systems. The authors in [8][9][10] specify about suffix stripping, porter stemmer, affix stripping algorithm for stemming. The author in [11] clearly explains about calculate weight for the word in a key word by which is used select the feature. In[12][13] clarify the views for the feature selection Tamil text classification. The author in [16][17] give a good idea about machine learning algorithm for the Tamil text classification

## III. PRE-PROCESSING

Text can come in a variety of forms from a list of individual words, to sentences to multiple paragraphs with special characters (like tweets for example). Preprocessing step transforms document into key word List.

**Tab 1.Tamil Corpus Taken**

| Category NAME | Total No of words (Tokenization) | Unique No of Word list from Corpus |
|---|---|---|
| All Combine | 2,56,700 | 98,560 |
| Agriculture | 22,462 | 7,735 |
| Science | 26,045 | 11,207 |
| Sports | 18,636 | 5,129 |
| Astrology | 21,375 | 5,573 |
| Literature | 35,612 | 16,949 |
| Spiritual | 26,786 | 8,450 |
| Cinema | 17,546 | 7,252 |
| Politics | 27,456 | 9,923 |
| Medical | 42,896 | 14,299 |
| Business | 26,673 | 9,112 |

**Revised Manuscript Received on April 15, 2020.**
***** Correspondence Author
**N. Rajkumar,** Research Scholar,Department of Computer and Information Science, Annamalai University,India, raju.prg@gmail.com
**T.S. Subashini,** Department of Computer Science and Engineering, Annamalai University, Tamilnadu, India, rtramsuba@gmail.com
**K. Rajan,** Department of Computer Engineering, Muthiah Polytechnic College, Annamalainagar. tamizhkavi@gmail.com
**V. Ramalingam,** Department of Computer Science and Engineering, Annamalai University, Tamilnadu, India, aucsevr@gmail.com

In goal behind the word preprocessing is to convert every document in to feature vector, that is, to break the document into individual key words. The preprocessing **phase** have various sub-phases namely Tokenization, Stop Word Removal, Stemming as shown in Fig. 1

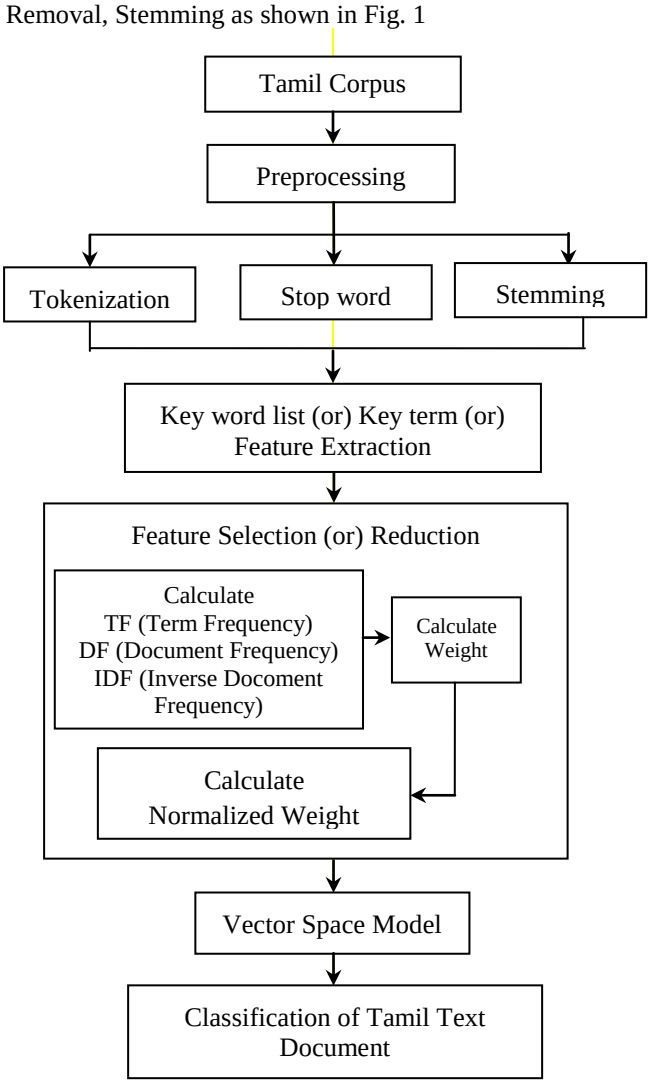


**Fig 1.Shows the steps in Tamil Text Classification**

**Tokenization** divides the documents into tokens where a token is a sequence of characters, which excludes white spaces and punctuation. **Stop words** have low discrimination value and the words like articles, prepositions, and conjunctions these words are also known as negative dictionary or noise words that present frequently in documents and these words do not have importance to do learning tasks. Therefore, these unwanted types of words lead to the degradation in result so stop word process is important. Stop words are eliminated in the document by dictionary based method and frequency based method. The next step of preprocessing is **stemming** which eliminate some character from the word for to find out the root format fix stripping method.

**Tab 2.Preprocessing result for above corpus**

| Preprocessing | | | | | Stemming |
|---|---|---|---|---|---|
| Stop Word | | | | | |
| Special symbol | Dictionary Based | Length Based | High Frequency | Low Frequency | Suffix Stripping |
| 11940 | 51,068 | 4567 | 11450 | 65679 | 25450 |
| 1400 | 4,967 | 396 | 1345 | 6156 | 2450 |
| 1050 | 5,345 | 437 | 1023 | 6569 | 2245 |
| 1200 | 5,046 | 489 | 1078 | 6498 | 2356 |

## IV. FEATURE SELECTION

In the preprocessing the stop words are removed and the stemming process reduces the keyword list by grouping different forms of a particular root word. Even after performing these two steps, there are many insignificant words which have appeared in more than one document categories and have negative effect on classification. Many classification algorithms cannot handle high-dimensional data sets in text classification problem. Hence, we need to reduce the dimensional space and improve the performance of the classifiers. FS uses machine learning algorithms to choose appropriate words that can represent the original text, which can reduce the dimensions of the feature space. The significance of a particular word for a selected category can be identified by estimating the term weight.

The Feature selection is the process of identifying a subset of features from the exhaustive list of features, which represent the text. The commonly used feature selection algorithms in the text processing tasks are as follows:

**InformationGain (IG):**

$$InfGain(w) = -\sum_{i=1}^{|C|} P(c_i)\log P(c_i)$$

**ExpectedCrossEntropy (ECE):**

$$= P(w)\sum_{i=1}^{|C|} P(w|C_i)\log\frac{P(w|C_i)}{P(C_i)},$$

**Mutual Information (MI):**

$$MutualInfo(w) = \log\frac{P(w|C_i)}{P(w)} = \log\frac{P(w,C_i)}{P(w)P(C_i)}$$

**Odds Ratio (OR):**

$$OddsRatio(w,c_i) = \log\frac{P(w|c_i)(1-P(w|\bar{c}_i)}{(1-P(w|c_i))P(w|c_i)},$$

**$\chi$2 (CHI):**

$$\chi^2(w, c_i) = \frac{|C|(A1 \times A4 - A2 \times A_3)^2}{(A_1 + A_3)(A_2 + A_4)(A_1 + A_2)(A_3 + A_4)},$$

**Gini Index (GI):**

$$Gini_A(Q) = 1 - \sum_{i=1}^{1} \frac{S_i}{S} Gini(Q).$$

**TheImprovedGiniIndex (IGI) Algorithm:**

$$IGI(w) = \sum_{i=1}^{|C|} P(w \mid c_i)^2 \, P(c_i \mid w)^2,$$

The term weight for each term is estimated as follows:

The term frequencies (TF) of a particular term in each category are calculated.

These term frequencies are divided by the total number of documents(DF) in which this term appears.

The Weight for the term of the document is calculated by

$Wt(t_j,d_i) = TF(t_j,d_i) * 1 / DF(t_j)$

Where,

TF (tj, di) - Term Frequency

IDF (tj) - Inverse Document Frequency

DF - Document Frequency

## V. IMPLEMENTATION AND RESULTS

The Term Frequency is calculated by the word count of a document. The Term Frequency is calculated for each Term in all the key word list of the entire category. Some words are selected from the key word list and calculated Term Frequency is shown Tab 3.

**Table 3. Term Frequency in Each category**
**c1-Agriculture,c2-Science,c3-Sports,c4-Astrology,c5-Literature ,c6-Spritual,c7-Cinema,c8-Politics,c9-Medical,c10-Business**

| Words | W c1 | W c2 | W c3 | W c4 | W c5 | W c6 | W c7 | W c8 | W c9 | Wc 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| சாகுபடி | 93 | 0 | 0 | 0 | 0 | 0 | 0 | 18 | 4 | 3 |
| விஞ்ஞானிகள் | 8 | 38 | 0 | 0 | 0 | 2 | 3 | 0 | 5 | 0 |
| மருந்து | 25 | 34 | 0 | 0 | 3 | 3 | 0 | 0 | 94 | 8 |
| தமிழ் | 21 | 10 | 0 | 7 | 67 | 3 | 36 | 119 | 11 | 7 |
| குரு | 0 | 0 | 0 | 121 | 33 | 39 | 3 | 0 | 0 | 0 |
| வருவாய் | 4 | 0 | 0 | 3 | 4 | 0 | 0 | 4 | 0 | 17 |
| பயிற்சி | 3 | 5 | 82 | 0 | 9 | 1 | 3 | 5 | 30 | 9 |
| திருமணம் | 0 | 1 | 0 | 6 | 9 | 29 | 36 | 0 | 1 | 1 |
| முதலீடு | 0 | 2 | 0 | 11 | 2 | 0 | 1 | 3 | 0 | 40 |
| நடிகர் | 0 | 0 | 1 | 0 | 5 | 0 | 59 | 5 | 1 | 4 |

After calculating Term Frequency (TF) for all categories, calculate weight by dividing it by the value of Document Frequency (DF). The Table 4 shows the weight of terms in agriculture document.

**Table 4.Term weight for Agriculture category**

The table 5 shows the term weight for the selected words

| Words | TF | Wt=Tf*IDF Agriculture | DF |
|---|---|---|---|
| சாகுபடி | 93 | 23.4 | 4 |
| விஞ்ஞானிகள் | 08 | 1.6 | 5 |
| மருந்து | 25 | 4.2 | 6 |
| தமிழ் | 21 | 2.3 | 9 |
| குரு | 00 | 0.0 | 4 |
| வருவாய் | 04 | 0.8 | 5 |
| பயிற்சி | 03 | 0.3 | 9 |
| திருமணம் | 00 | 0.0 | 7 |
| முதலீடு | 00 | 0.0 | 6 |
| நடிகர் | 00 | 0.0 | 6 |

from the key word list .The weight is calculated with help of Term Frequency and Inverse document Frequency.

**Table 5. Term Weight for all categories**

| Words | W c1 | W c2 | W c3 | W c4 | W c5 | W c6 | W c7 | Wc 8 | W c9 | W c10 |
|---|---|---|---|---|---|---|---|---|---|---|
| சாகுபடி | 23.3 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 4.5 | 1.0 | 0.8 |
| விஞ்ஞானிகள் | 1.6 | 7.6 | 0.0 | 0.0 | 0.0 | 0.4 | 0.6 | 0.0 | 1.0 | 0.0 |
| மருந்து | 4.2 | 5.7 | 0.0 | 0.0 | 0.5 | 0.5 | 0.0 | 0.0 | 15.7 | 1.3 |
| தமிழ் | 2.3 | 1.1 | 0.0 | 0.8 | 7.4 | 0.3 | 4.0 | 13.2 | 1.2 | 0.8 |
| குரு | 0.0 | 0.0 | 0.0 | 30.3 | 8.3 | 9.8 | 0.8 | 0.0 | 0.0 | 0.0 |
| வருவாய் | 0.8 | 0.0 | 0.0 | 0.6 | 0.8 | 0.0 | 0.0 | 0.8 | 0.0 | 3.4 |
| பயிற்சி | 0.3 | 0.6 | 9.1 | 0.0 | 1.0 | 0.1 | 0.3 | 0.6 | 3.3 | 1.0 |
| திருமணம் | 0.0 | 0.1 | 0.0 | 0.9 | 1.3 | 4.1 | 5.1 | 0.0 | 0.1 | 0.1 |
| முதலீடு | 0.0 | 0.3 | 0.0 | 1.8 | 0.3 | 0.0 | 0.2 | 0.5 | 0.0 | 6.7 |
| நடிகர் | 0.0 | 0.0 | 0.2 | 0.0 | 0.8 | 0.0 | 9.8 | 0.8 | 0.2 | 0.7 |

This term weights are normalized by dividing each value by sum of the term weights for each term. These normalized term weights are shown in the Table 6.

**Table 6.Normalised term weights**

| Words | Wc1 | Wc2 | Wc3 | W c4 | W c5 | Wc6 | Wc7 | W c8 | Wc9 | W c10 |
|---|---|---|---|---|---|---|---|---|---|---|
| சாகுபடி | 0.79 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.15 | 0.03 | 0.03 |
| விஞ்ஞானிகள் | 0.14 | 0.68 | 0.00 | 0.00 | 0.00 | 0.04 | 0.05 | 0.00 | 0.09 | 0.00 |
| மருந்து | 0.15 | 0.20 | 0.00 | 0.00 | 0.02 | 0.02 | 0.00 | 0.00 | 0.56 | 0.05 |
| தமிழ் | 0.07 | 0.04 | 0.00 | 0.02 | 0.24 | 0.01 | 0.13 | 0.42 | 0.04 | 0.02 |
| குரு | 0.00 | 0.00 | 0.00 | 0.62 | 0.17 | 0.20 | 0.02 | 0.00 | 0.00 | 0.00 |
| வருவாய் | 0.13 | 0.00 | 0.00 | 0.09 | 0.13 | 0.00 | 0.00 | 0.13 | 0.00 | 0.53 |
| பயிற்சி | 0.02 | 0.03 | 0.56 | 0.00 | 0.06 | 0.01 | 0.02 | 0.03 | 0.20 | 0.06 |
| திருமணம் | 0.00 | 0.01 | 0.00 | 0.07 | 0.11 | 0.35 | 0.43 | 0.00 | 0.01 | 0.01 |
| முதலீடு | 0.00 | 0.03 | 0.00 | 0.19 | 0.03 | 0.00 | 0.02 | 0.05 | 0.00 | 0.68 |
| நடிகர் | 0.00 | 0.00 | 0.01 | 0.00 | 0.07 | 0.00 | 0.79 | 0.07 | 0.01 | 0.05 |

Finally the features for each category are selected based on the normalized term weights. The below table 7 lists some of the selected words for four categories

**Table7 .Sample Features selected for four categories**

| Categories | | | |
|---|---|---|---|
| **Agriculture** | **Science** | **Sports** | **Business** |
| சாகுபடி (0.79) | விஞ்ஞானிகள் (0.68) | பயிற்சி (0.56) | முதலீடு (0.68) |
| மருந்து (0.15) | மருந்து (0.20) | ---------- | வருவாய் (0.53) |
| விஞ்ஞானிகள் (0.14) | தமிழ் (0.04) | ---------- | பயிற்சி (0.06) |
| வருவாய் (0.13) | ---------- | ---------- | மருந்து(0.05) |

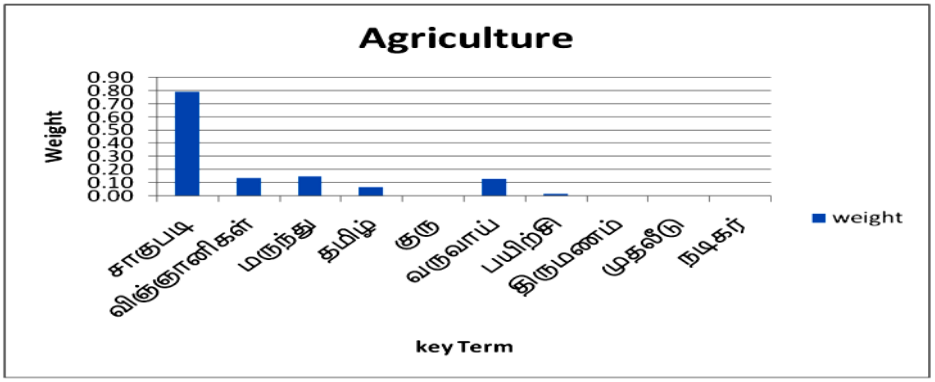


**Fig. 2 Category Agriculture**

The above graph specifies the assignment of the weight for the word in the table 6. of Agriculture category

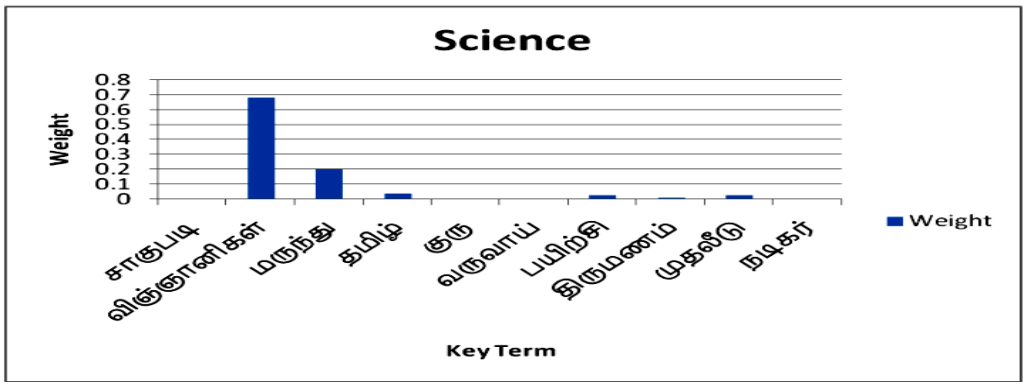


**Fig. 3 Category Science**

The above graph specifies the assignment of the weight for the words in the table 6. of science category.

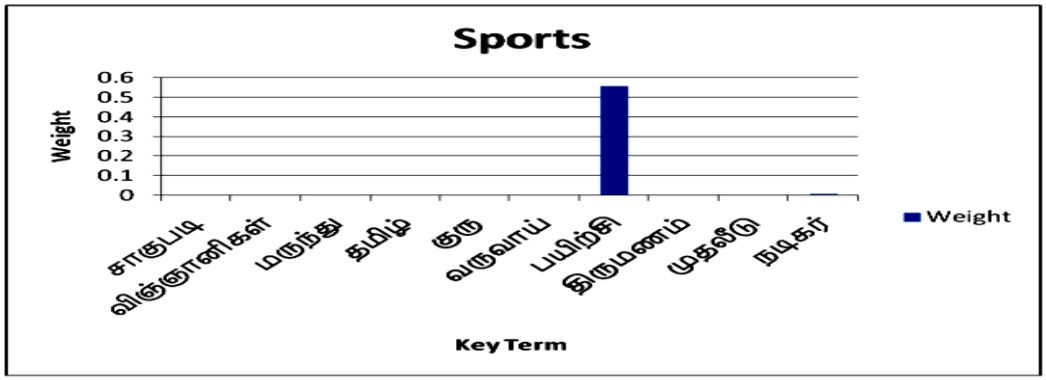


**Fig. 4 Category sports**

The above graph specifies the assignment of the weight for the words in the table 6. of sports category.

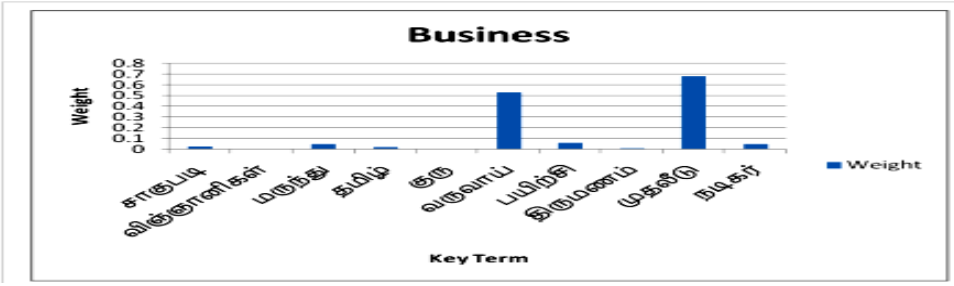


**Fig. 5 Category Business**

The above graph specifies the assignment of the weight for the word in the table 6. of business category

## VI. CONCLUSION

There has been considerable amount of work done in Feature selection by weighting method using Term Frequency and Inverse Document Frequency. In this paper words from 10 different categories are used for estimating term weights and are assigned normalized term weights. Based on this weight, features are selected. The terms with lower term weights are eliminated from the keyword list. These weights represent the importance of each term for different categories. Features are selected based on normalized term weights for each category and are added to the dictionary of each category which are very useful for Tamil document classification

## REFERENCES

1. A. El-Khair, "Effects of stop words elimination for arabic information retrieval: a comparative study", International Journal of Computing & Information Sciences, vol. 4, no. 3, pp. 119-133, 2006.
2. Ashish T, Kothari M and Pinkesh P, “Pre-Processing Phase of Text Summarization Based on Gujarati Language”, International Journal of Innovative Research in Computer Science & Technology (IJIRCST) Vol-2, Iss-4, July 2014.
3. Rakholia R.M., Saini J.R. “A Rule-Based Approach to Identify Stop Words for Gujarati Language”, Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications. Advances in Intelligent Systems and Computing, vol 515. Springer, Singapore, 2017
4. Jaideepsinh K. Raulji , Jatinderkumar R. Saini, “Stop-Word Removal Algorithm and its Implementation for Sanskrit Language”, International Journal of Computer Applications (0975 – 8887) Volume 150 – No.2, 2016
5. V. Jha, N. Manjunath, P. D. Shenoy and K. R. Venugopal, "HSRA: Hindi stopword removal algorithm," 2016 International Conference on Microelectronics, Computing and Communications (MicroCom), Durgapur, 2016, pp. 1-5.
6. R. Al-Shalabi, G. Kanaan, J. M. Jaam, A. Hasnah, E. Hilat, "Stop-word removal algorithm for arabic language", Proceedings of 1st International Conference on Information & Communication Technologies: from Theory to Applications CTTA ‘04, pp. 545-550, 2004.

7. Tsz-Wai Lo, Rachel & He, Ben &Ounis, Iadh. (2005). Automatically Building a Stop word List for an Information Retrieval System.. JDIM. 3. 3-8.
8. Ramachandran, VivekAnandan and Krishnamurthi, Ilango. Ilango. "An Iterative Suffix Stripping Tamil Stemmer". Proceedings of the International Conference on Information Systems Design and Intelligent Applications (2012): Volume 132, 583-590.
9. M.Thangarasu and R.Manavalan, "Design and Development of Stemmer for Tamil Language",International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 7, July 2013.
10. Rajalingam, Damodharan. "A Rule Based Iterative Affix Stripping Stemming Algorithm For Tamil." 12th International Tamil Internet Conference. 2013. pp. 28-34.
11. Rajan, V.Ramalingam, M.Ganesan, S. Palanivel, B. Paiappan " Automatic classification of Tamil documents using vector space model and artificial neural network."Expert Systems with Applications 36 (2009) 10914–10918.
12. Sanjanasri J, Anand Kumar M," A Computational Framework for Tamil Document Classification using Random Kitchen Sink" International Conference on Advances inComputing, Communications and Informatics (ICACCI) ,2015.
13. NadanaRavishankar, ShriramRaghunathan," corpus based sentiment classification of tamil movie tweets using syntactic patterns", , | IIOABJ | Vol. 8 ||2017|.
14. Songtao Shang, Minyong Shi, Wenqian Shang, and Zhiguo Hong, "Improved Feature Weight Algorithm and Its Application to text Classification", Hindawi Publishing Corporation, Volume 2016, Article ID 7819626, 12 pages
15. LI Yong-fei, "A Feature Weight Algorithm for Text Classification Based on Class Information", Published by Atlantis Press, Paris, France, Proceedings of the 2012 2nd International Conference on Computer and Information Application
16. Hanumanthappa, ,NarayanaSwamy M .,"indian language text documents categorization and keyword extraction", I J C T A, 9(3), 2016, pp. 1473-1481.
17. M NarayanaSwamy, , M. Hanumanthappa, " Indian Language Text Representation and Categorization Using Supervised Learning Algorithm", International Journal of Data Mining Techniques and Applications, Vol:02, December 2013, Pages: 251-257.


## AUTHORS PROFILE

**N. RAJKUMAR,** M.C.A., M.Phil., working as Assistant Professor in Dept. of Computer Science, Govt. Arts and Science College, Kumbakonam, Tamil Nadu, India. He has more than 13 years of teaching experience and currently pursuing his Ph.D in the Department of Computer and Information Science, Annamalai University, Annamalai Nagar, Tamil Nadu, Indi. He published 3 papers (1 International Journal, 1 International Conferences, 1 National Conferences). His main research work focuses on neural networks, Natural languages processing, Tamil Computing, Computational Linguistics and machine learning.

**Dr. T.S. SUBASHINI** is working as Professor of Computer Science and Engineering in the Department of Computer Science and Engineering, Faculty of Engineering and Technology, Annamalai University, Tamil Nadu, India. She has been on its faculty for nearly two and half decades. She earlier served as Hardware Engineer in Sterling Computers Ltd., Chennai. She received her Ph.D from the Annamalai University and her area of research is Medical Image Analysis. In her doctoral work, she has developed a computer aided diagnosis system for breast cancer using Mammogram images. She has completed a major Research Project, titled "Content Based Image Retrieval coupled CAD for breast cancer using Digital Mammograms" which was funded by UGC, Govt. of India. She has published research articles in Elsevier, Taylor and Francis and ACM International journals. She is also a reviewer of Springer and Elsevier journals. She has to her credit 60 International journal papers. She has published more than 65 papers in National and International Conferences. She has organized several National Workshop and Conferences. She organized an International Conference on Pattern Recognition and Multimedia Signal Processing on January 2015. She is regularly invited to deliver lectures in various conferences and programmes for imparting skills in research methodology to students. Her research interests include Image and Video processing, Computer Vision and Pattern Classification and Image Forensics.

**Dr. K. Rajan,** is working as Head of the Department in Computer Engineering at Muthiah Polytechnic College. He received his Bachelor of Engineering (Electronics and Communication) from Madras University, M.S.(Engg. By Research) in Software Engineering and Ph.D (CSE) from Annamalai University. He has more than 30 years of teaching experience. He is a CMI Level 5 certified Manager and Leader. He is member of ISTE, CSI, INFITT, IAENG and CSTA. He has authored E-Text Books and E-Lectures for DOTE-Tamilnadu. He has published several research papers on national and international conferences in the areas of Tamil Computing, NLP, Computational Linguistics and Machine Learning.

**Professor Dr. V. Ramalingam** received B.Sc (Maths) and M.Sc (Statistics) degree from Annamalai University in 1978 and 1980 respectively. He was University Rank holder in both UG and PG studies in Annamalai University. He received M.Tech degree in Computer Applications from Indian Institute of Technology Delhi in the year 1995. He completed his Ph.D degree in Computer Science and Engineering at Annamalai University in 2006. He has served 29 years in Annamalai University. He published 145 papers (59 International Journals, 21 International Conferences, 26 National Journals, 39 National Conferences). His research interest includes neural networks, image and video processing, and natural language processing.