**International Journal of Scientific and Innovative Mathematical Research (IJSIMR)**
*Volume 7, Issue 6, 2019, PP 22-29*
*ISSN No. (Print) 2347-307X & ISSN No. (Online) 2347-3142*
*DOI: http://dx.doi.org/10.20431/2347-3142.0706004*
*www.arcjournals.org*

# Application of Tamil Syntax based on Authorship Attribution Using Neural Network and Classification Methods

**G. MANIMANNAN[1*], R. LAKSHMI PRIYA[2]**

[1]*Assistant Professor, Department of Mathematics, TMG College of Arts and science, Chennai*

[2]*Assistant Professor, Department of Statistics, Dr. Ambedkar Govt. Arts College, Chennai*

***Corresponding Author:** ***G. MANIMANNAN,*** *Assistant Professor, Department of Mathematics, TMG College of Arts and science, Chennai*

**Abstract:** *Application of Neural Networks with regard to author attribution as a problem of pattern recognition and proven results of their applications make them as promising techniques for the future in continuing authorship determination. Learning Vector Quantization (LVQ) is a neural network technique that develops a codebook of quantization vectors and makes use of these vectors to encode any input vector. In this paper an attempt is made to identify authorship attribute of disputed articles using LVQ and verify with the results obtained by traditional canonical discriminant analysis. This study demonstrates that statistical methods of attributing authorship can be clubbed successfully with neural networks to produce a powerful classification tool. Comparisons are made using means of sixteen articles of syntax identified from thirty - one articles written in Tamil language by three contemporary scholars namely, Mahakavi Bharathiar (MB), Subramaniya Iyer (SI) and T. V. Kalyanasundaranar (TVK) of identical repute to determine the authorship of twenty-three un-attributed articles pertaining to the same period.*

**Keywords:** *Authorship Attribution, Syntax, Neural Network, Learning Vector Quantization and Canonical Discriminant Analysis.*

## 1. INTRODUCTION

The present field of authorship attribution began with discipline known as stylometry. Stylometry is the study of quantifiable of human language, or the statistical analysis of literary style (Holmes, 1995). This involves attempting to formally capture the creative, unconscious elements of language particular to individual writers and speakers. Although researchers have studied writing for centuries, the discipline of stylometry is fairly recent, and while its origins date back to the late 19$^{th}$ century, the field as it is now began with work on *Federalist Papers* in 1968 (Mosteller and Wallace, 1968). Even though the study of stylometry began before the advent of modern computing; this discipline has easily been applied to computers. The principle of stylometry is formalization of writing style into quantifiable features and subsequent comparison of those features. As such, it includes vast calculations perfectly suited to a computer rather than a human.

Stylometry mainly concerns itself with authorship attribution studies, although chronological studies on dating of work within the corpus of an author have also investigated. Writing in a forensic background, Bailey (1969) proposed three rules to define the situation necessary for authorship attribution:

- The number of putative authors should constitute a well- defined set.
- The lengths of the writings should be sufficient to reflect linguistic behavior of the author of the disputed text and also those of the candidates;
- The texts used for comparison should be commensurate with the disputed writing.

A computational stylistic study of doubtful authorship should involve comparisons of the disputed text with works by each of the possible candidate authors using suitable statistical tests on quantifiable features of the texts – features which reflect the style of writing as defined above.

One modern addition tool available for computational stylometry is artificial neural network (ANN). Its computational methods are loosely based on the concept of biological neuron, idea being that simple, trained processing elements will result in much more difficult behavior when used in combination. Figure 1 shows an example of structures of a typical ANN. The majority of ANNs may be used to inductively learn a theory from input and output patterns.

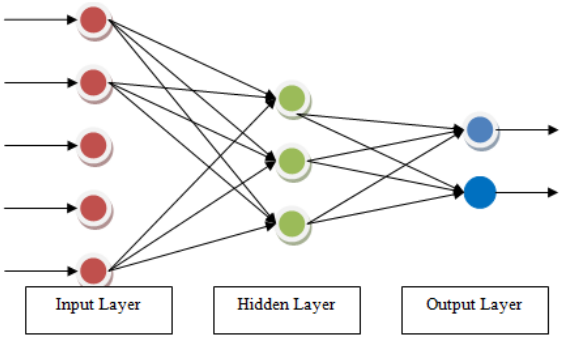



**Figure1.** *Structure of Artificial Neural Network*

### 1.1 Review of Literature

In recent years, many scholars have successfully demonstrated this technique of machine learning field which can be applied to authorship attribution. Merriam and Mathews (1994) have trained a multi layer perception network to distinguish the works of Shakespeare and Marlowe. Tweedie *et al.* (1996) have provided a useful review of applications of ANNs in the area of computational stylometry and have used this machine-learning package for reanalysis of *Federalist Papers*. Kjell (1994) have taken up authorship study using letter-pair frequency features with neural network classification. One of the most extensive in Statlog project (Michie, Spiegelhalter and Taylor, 1994) statistical methods, machine learning, and neural networks were compared using a large number of different data sets. This study has recommended LVQ as one of the best good neural network classification.

In this Paper, the researcher present the first version of learning vector quantization, a supervised learning neural network model to classify the un-attributed articles and has shown how potent it is in the discriminative power and the capability to learn and represent implicit knowledge of the computational stylistic data set. Recently, scores of researchers have made systematic studies to investigate links between neural network based techniques and traditional statistical classification techniques. In majority of cases, these two techniques have been seen as alternatives, or in fact, neural networks, in general, have been seen as a subset of statistics. This approach has led, on one hand, to a fruitful analysis of existing neural networks, and on the other, it has enriched the current statistical methods with newly traced viewpoints, and sometimes has pioneered to a useful synthesis of the two fields.

## 2. DATABASE

The present study deals with the literary works of three contemporary Tamil scholars, namely, Mahakavi Bharathi (MB), T. V.Kalyanasundaranar (TVK), and Subramaniya Iyer (SI). In the Pre–Independence period, these three scholars have written number of articles on India's Freedom Movement in the magazine called *India*. Initially, all the three scholars have written articles by attributing their names. The oppressive attitude of the then British Regime made all the three writers to write articles on the same topic anonymously in the same magazine. All the attributed and un-attributed articles written on India’s Freedom Movement in that magazine were compiled and brought out as a book entitled *Bharathi Dharisanam* in the year 1975. For this quantitative attribution study, all attributed articles of these three scholars written on India’s Freedom Movement in the year 1906 are considered. Our study is based on nineteen articles of MB, five of TVK and seven of SI and twenty-three un-attributed articles. Sixteen articles syntax were considered for this study. In stylometry, important decisions to be made about the features to be selected and the methods to be used (Mealand D. 1997). As any other Indian language, Tamil shows the finite verb at the end of a sentence. That is to say, Tamil is a verb final language. Structurally speaking there are four major

types of simple sentences in Tamil, and they are 1. Noun Phrase (NP) + Verb Phrase (VP), 2. Noun Phrase (NP) + Noun Phrase (NP), 3. Noun Phrase (NP) + Adjectives (ADJ), 4. Noun phrase (NP) + Genitive (GEN). These sentences are classified on the basis of the items that occur in the predicate slot. The exact lists of variables of this study with their meanings are given in Table 1.

**Table1.** *List of Sixteen combinations of syntax*

| Variable Name | Abbreviation of Variable |
|---|---|
| NP+VP | Noun Phrase followed Verb Phrase |
| NP+NP | Noun Phrase followed Adjectives |
| NP+ADJ. | Noun phrase followed Genitive |
| NP+GEN. | Verb Phrase followed Noun Phrase |
| VP+NP | Verb Phrase followed Verb Phrase |
| VP+VP | Verb Phrase followed Adjectives |
| VP+ADJ. | Verb Phrase followed Genitive |
| VP+GEN. | Adjectives followed Noun Phrase |
| ADJ.+NP | Adjectives followed Verb Phrase |
| ADJ.+VP | Adjectives followed Adjectives |
| ADJ.+ADJ. | Adjectives followed Genitive |
| ADJ.+GEN. | Genitive followed Noun Phrase |
| GEN+NP | Genitive followed Verb Phrase |
| GEN+VP | Genitive followed Adjectives |
| GEN+ADJ. | Genitive followed Genitive |
| GEN+GEN. | |

For a comparative analysis frequency counts of stylistic features must be normalized to the text length in an article. In this study since each sentence is considered as a sample, to normalize the syntax, the raw frequency counts of each syntax word is divided by the number of words in each sentence and then multiplied by hundred to express it in percentage. Sixteen articles of syntax are identified from each sentence. If we have n sentences and if we identify p syntax from each sentence, then we have a data matrix of size n x p. Thus, each article is converted as a data matrix and these data matrices form the basis for this quantitative study.

Chi-square analysis of nineteen articles of MB establishes that these articles do not differ from one another in terms of frequency distribution of occurrence of these stylistic features. Similar results were obtained in case of other two scholars (Manimannan and Bagavandas, 2001). Hence all nineteen articles of MB were considered as one article for this study and also, five articles of TVK and seven articles of SI. Hence nineteen articles of MB consist of three hundred and fifty three sentences, five articles of TVK consist of three hundred and eighty two sentences and seven articles of SI consist of three hundred and fifteen sentences. As there are three authors, there are three data matrices with size (353x16), (382x16) and (315x16) respectively. The main objective of this study is to explore authorship attribution of the un-attributed articles with those of MB, SI and TVK.

## 3. Neural Network Approach

One feature that Stylometry shares with other applications is that it is basically a case of pattern recognition. However, in majority of cases of doubtful authorship we do not know what the pattern is and we probably would not be able to recognize it even if we did. Neural networks have ability to recognize the underlying organization of data, which is of vital importance for any pattern reorganization problem. Tweedie e*t. al.* (1996) proposed their application in Stylometry will be useful for a number of reasons

- Neural networks can learn from the data themselves. Implementing a rule-based system in linguistic computing may become complex as the number of distinguishing variables increases and even the most complex rules may still not be good enough to completely characterize the training data. In essence, neural networks are more adaptive.
- Neural networks can generalize. This ability is particularly required in the literary field, as only limited data may be available.
- Neural networks can capture non-linear interactions between input variables.
- Neural networks are capable of fault tolerance. Hence a particular work, which is not in line with the usual writing style of an author, will not affect the network to a considerable extent. Thus

neural networks appear to promise much for the field of stylometry. Their application would appear to be worthy of investigation.

### 3.1. Learning Vector Quantization (LVQ)

Learning vector quantization was developed by Kohonen (Kohonen, 1986). In the year 1990, he has summarized three versions of this algorithm (Kohonen, 1990). It is a supervised learning technique that can classify input vectors based on vector quantization. The first version of learning vector quantization (LVQ1) is used for this attribution study. LVQ1 training process proceeds with an input vector being randomly selected (along with the correct class for that vector, thus the supervised learning) from the "labeled" training set. Given an input vector $u_i$ to the network, the "output neuron" (i.e., the class or category) in LVQ1 is deemed to be a "winner" according to

$$\min_{\forall j} d(u_i, w_j) = \min_{\forall j} \| u_i - u_j \|^2 \quad (1)$$

### 3.2. Algorithm

Let $\{u_i\}$, for i = 1, 2, …, N be the set of input vectors, and the network synaptic weight vectors (Voronoi vector) are denoted as $\{w_j\}$, for j = 1, 2,…,m. We also let $Cw_j$ be the class (or category) that is associated with the weight vector $w_j$ and $Cu_i$ is the class label of the input vector $u_i$ to the network. The weight vector wj is adjusted in the following manner:

(i). If the class associated with the weight vector and the class label of the input are the same, that is, $Cw_j = Cu_i$, then

$$w_j(k+1) = w_j(k) + \mu(k)\left[u_i - w_j(k)\right] \quad (2)$$

where $0 < \mu(k) < 1$ (the learning rate parameter)

(ii). But if $Cw_j \neq Cu_j$, then $w_j(k+1) = w_j(k) - \mu(k)\left[u_i - w_j(k)\right]$ (3)

and the other weight vectors are not adapted.

Therefore, the update rule for modifying a weight vector in 2 is the standard one if the class is correct. In other words, according to the learning rule in 2, the weight vector $w_j$, is moved in the direction of the input $u_i$ if the class labels of the input vector and the weight vector agree. However, if the class is not correct, the weight vector is moved in the opposite direction away from the input vector according to 3. The learning rate parameter $\mu(k)$ is monotonically decreased in accordance with the discrete-time index k (e.g., linearly decreased in time, starting at 0.01 or 0.02; However, many times 0.1 is used as the initial value).

The convergence properties of LVQ have been studied by Baras and LaVigna (Baras and LaVigna, 1990), and their approach is based on stochastic approximation theory. The weights can be initialized by using several methods. In this Paper the first m (total number of classes) vectors from the set of training vectors are used to initialize the weight vectors, that is, $w_j(0)$ for j = 1, 2,.., m. Another approach is to randomly initialize the weight vectors (within the dynamic range of the input vectors).

The stopping condition can be based on the total number of desired training epochs, or on monitoring the convergence of the weight vectors. Another stopping condition can be based on monitoring the learning rate parameter directly, and when it is sufficiently small, training can be terminated. In this Paper, we establish the stopping condition to be the total number (predefined) of iterations. The basic LVQ algorithm can be summarized as follows:

***Step 1***. Initialize all weight vectors $w_j(0)$), initialize the learning rate parameter $\mu(0)$, and set $k = 0$.

***Step 2***. Check the stopping condition. If false, continue; if true, quit.

***Step 3***. For each training vectors $u_i$, perform step 4 and 5:

***Step 4***. Determine the weight vector index $(j = q)$ such that $\min \left\| u_i - w_j(\mathrm{k}) \right\|^2$ and the weight vector $w_q(k)$ that minimizes the square of the norm.

***Step 5***. Update the appropriate weight vector $w_q(k)$ as follows:

If $Cw_q = Cu_i$, then $w_q(k+1) = w_q(k) + \mu(k)\left[u_i - w_q(k)\right]$

If $Cw_q \neq Cu_i$, then $w_q(k+1) = w_q(k) - \mu(k)\left[u_i - w_q(k)\right]$

***Step 6***. Set $k \leftarrow k+1$, reduce the learning rate parameter, then go to Step 2. The learning rate parameter μ can be reduced in accordance with k (discrete–time index) sing

$\mu(k) = \dfrac{\mu(k+1)}{(k+1)}$ for $k > 0$.

The neural network architecture for LVQ is shows in *Figure 2*

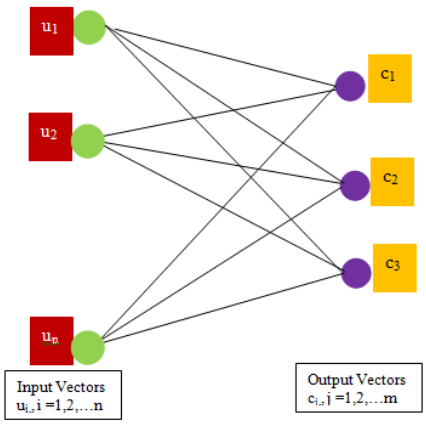



**Figure2.** *Neural Network architecture for Learning Vector Quantization*

## 4. DATA ANALYSIS

Here, the researcher associates the nineteen articles of MB to group 1, seven of SI to group 2 and five of TVK to group 3, and thereby have totally thirty-one articles (vectors) associated with three classes. One method to initialize the codebook vectors $w_1$, $w_2$, and $w_3$ is to use first three articles, one from each group of articles of each author. The associated groups for $w_1$, $w_2$, and $w_3$ are then 1, 2, and 3 respectively. The remaining twenty-eight articles of three scholars can be used for training.

We initialize the learning rate parameter to $\mu(k=1) = \mu(1) = 0.1$ and decrease it by $k$ every training epoch, for example $\mu(2) = \dfrac{\mu(1)}{2}$, $\mu(3) = \dfrac{\mu(2)}{3}$, etc. According to the algorithm given above, we set $k = 2$, and $\mu(2) = \dfrac{\mu(1)}{2}$ = 0.05, then go to step 2 and check the stopping condition. The next training epoch we start with the first training article among the remaining twenty-nine articles of the three scholars. The MATLAB coding is used with the number of training epochs set to 20,000 to find the *codebook vectors*. After 20,000 training epochs, the final weights give the following *codebook* vectors in *Table 2*.

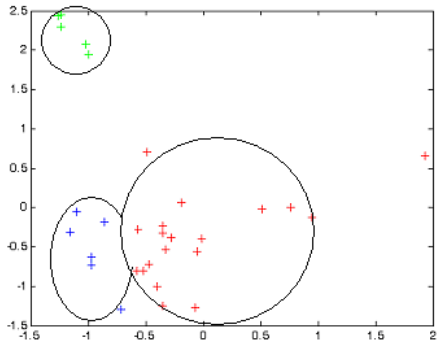
**Table2.** *Codebook vectors after 20,000 training epochs*

| W1 | W2 | W3 | W1 | W2 | W3 |
|---|---|---|---|---|---|
| 65.5005 | 122.955 | 70.5998 | 5.8381 | 2.1660 | 1.4240 |
| 24.6500 | 38.4275 | 11.9448 | 5.6772 | 2.1660 | 1.5040 |
| 7.7487 | 5.2769 | 5.2561 | 6.0263 | 2.1660 | 1.4791 |
| 10.9729 | 6.9630 | 4.0745 | 5.6772 | 2.1660 | 1.3929 |
| 5.6751 | 52.1495 | 1.8903 | 5.6772 | 2.1660 | 1.3929 |

| | | | | | |
|---|---|---|---|---|---|
| 13.4438<br>6.6548<br>5.6300 | 14.1218<br>3.5633<br>2.8200 | 6.4602<br>3.1572<br>2.1261 | 5.7217<br>5.6772<br>5.6772 | 2.1660<br>2.1660<br>2.1660 | 1.3918<br>1.3929<br>1.0403 |

Finally, if we calculate the minimum distance between each of the un-attributed articles and the computed weight vectors $w_1$, $w_2$, and $w_3$ of the codebook shown above, it will give the class to which each un-attributed articles. The experiment reveals that the Tamil Scholar, Mahakavi Bharathiar has written all the twenty-three un-attributed articles.

The convergence profiles for the elements of the training epoch are not shown due to lack of space. Finally, if we calculate the minimum distance between each of the un-attributed articles and the computed weight vectors w1, w2, and w3 of the codebook shown above, it will give the class to which each un-attributed articles belongs, and it is computed using the MATLAB code. At first, the undisputed articles were presented to the LVQ1 neural network technique to verify its efficiency. This neural network should then be able to identify an article of MB and so on. The end result of this intra-analysis is given in Figure 3. The convergence result is given in Table 3. Thus LVQ1 with MATLAB code has identified correctly all attributed articles. All twenty-three disputed articles are presented to LVQ1 neural network. This experiment attributes all twenty-three un-attributed articles to the Tamil Scholar Mahakavi Bharathiar and final result is given in figure 4. The convergence result is given in Table 4.

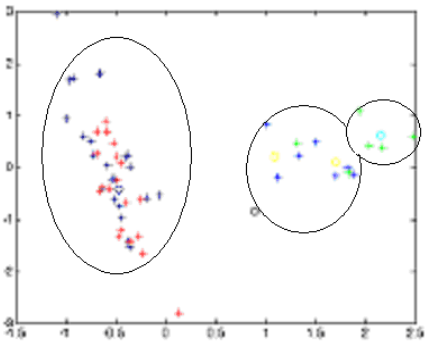


*Note: + red = MB, + Blue = SI and + Green = VK*

**Figure3.** *Classification map after training LVQ technique for MB, SI and VK*

**Table3.** *Convergence results after training LVQ technique for MB, SI and VK*

| Articles | Group1 | Group2 | Group 3 |
|---|---|---|---|
| MB | **19** | 0 | 0 |
| SI | 0 | **7** | 0 |
| VK | 0 | 0 | **5** |



*Note: + red = MB, + Light Blue = SI, + Green = VK and + Dark Blue = un attributed*

**Figure4.** *Classification map after training LVQ technique for MB, SI, VK and Disputed Articles*

**Table4.** *Convergence results after training LVQ technique for MB, SI, VK and Un attributed Articles*

| Articles | Group1 | Group2 | Group 3 |
|---|---|---|---|
| MB | **19** | 0 | 0 |
| SI | 0 | **7** | 0 |
| VK | 0 | 0 | **5** |
| Unattributed | **23** | 0 | 0 |

### 4.1. Cross – Validation

This study proposes to re-analyse the same data-set using canonical discriminant analysis to verify the results obtained through LVQ1 neural network computation. Canonical discriminant analysis is a multivariate method developed for testing the significance between two or more pre-defined groups of objects. The main objectives of this analysis are (1) to determine a set of linear discriminant functions with ordered discrimination power between groups identified a priori (2) to test whether the means of these groups are significantly different, and (3) to assign individual objects of unknown origin to the given known groups. The main assumptions of this analysis are that there are multiple groups that can be unambiguously well defined in advance and all individuals of unknown origin can be assigned to one and only one such group (Klecka 1980).

An intra-analysis on three authors (MB, TVK and SI) is made using canonical discriminant analysis to determine whether these three author's styles of using syntax are different or not. Classification map (Figure 5) and classification table (Table 5) of this analysis indicates that all the three author's styles of using syntax are distinct. The same analysis is repeated again to attribute authorship for disputed articles. The cross- validation results also indicate that, all twenty-three un-attributed articles go with the Tamil Scholar Mahakavi Bharathiar. It can be seen that un-attributed articles (Table 6, Figure 6) shows that twenty three articles are classified as articles of Bhartahiar. On the basis of these two analyses we can definitely attribute twenty three articles to Mahakavi Bharathiar.

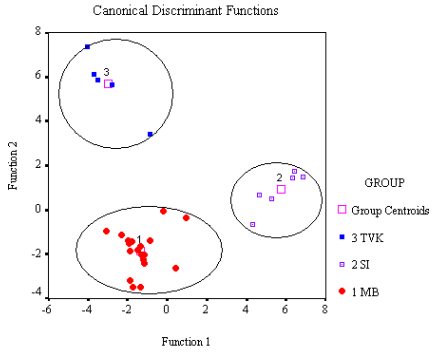



**Figure5.** *Classification plots for MB, SI and VK*

**Table5.** *Classification results for MB, SI and VK*

| Articles | Group1 | Group2 | Group 3 |
|---|---|---|---|
| MB | **19** | 0 | 0 |
| SI | 0 | **7** | 0 |
| VK | 0 | 0 | **5** |

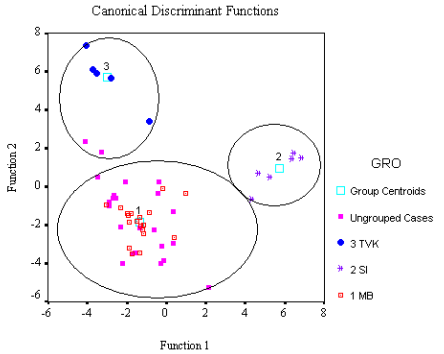



**Figure6.** *Classification plots for MB, SI, VK and Disputed Articles*

**Table6.** *Classification results for MB, SI, VK and Disputed Articles*

| Articles | Group1 | Group2 | Group 3 |
|---|---|---|---|
| MB | **19** | 0 | 0 |
| SI | 0 | **7** | 0 |
| VK | 0 | 0 | **5** |
| Unattributed | **23** | 0 | 0 |

## 5. CONCLUSION

Computational literary analysis provides an unique opportunity for researchers to experiment with high dimensional data. This comparative study made use of neural network technique of learning vector quantization and multivariate statistical techniques is to distinguish the writing styles of three Tamil scholars, namely, Mahakavi Bharathiar, T. V.Kalyanasundaram and Subrmaniya Iyer and also to attribute authorship for disputed articles. The materials for this study were nineteen articles of Mahakavi Bharathiar, five articles of T. V. Kalyanasundaranar, seven articles of Subramaniya Iyer and twenty-three disputed articles. All these articles were written on India's freedom movement in same period. The stylistic features of this study are sixteen article of syntax. It is clearly seen that all unattributed articles we classified as articles of Bhartahiar. On the basis of these two analyses all twenty three unattributed articles merge with the writing style of Mahakavi Bharathiar.

## REFERENCES


[1] Bailey, R. W. (1969). *Statistics and style: A historical survey*. In L. Dolezel & R. W. Bailey (Eds.), *Statistics and style*. New York, NY: American Elsevier Publishing Company Inc,pp.217- 236, 1969.

[2] Baras, J. S., and LaVigna, A., *Convergence of Kohonen's Learning Vector Quantization*,

[3] International Joint Conference on Neural Networks, San Diego, CA, vol. 3, pp. 17-20, 1990.

[4] David Mealand., *Measuring Genre Differences in Mark with Correspondence Analysis*, Literary and Linguistic Computing, 12, No.4. 1997.

[5] Holmes D.I. and Forsyth R.S. *The Federalist Revisited: New Directions in Authorship Attribution*, Literary and Linguistic Computing, 10, 111-127,1995.

[6] Kjell, B. (1994). *Authorship Determination Using Letter-pair Frequency Features with Neural Network Classifiers*. Literary and Linguistic Computing, 9: 119-24,1994.

[7] Klecka W.R. *Discriminant Analysis*, Sage Publications, California, 1980.

[8] Kohonen, T., *Learning Vector Quantization for Pattern Recognition*, Technical Report TKK-F-A601, Helsinki University of Technology, Finland,1986.

[9] Kohonen, T., *Improved Version of Learning Vector Quantization*, Proceedings of the International Joint Conference on Neural Networks, San Diego, CA, vol. 1, pp. 545-50,1990.

[10] Kohonen, T., *The Self-Organizing Map*, Proceedings of the Institute of Electrical and Electronics Engineers, vol. 78, pp. 1464-1480,1990.

[11] Manimannan G. and Bagavandas M., *The Authorship Attribution: the case of Bharathiyar*, Paper Presented at National Conference on Mathematical and Applied Statistics,

[12] Nagpur University, Nagpur, 2001

[13] Merriam, T. and Mathewa, R. *Neural Computation in Stylometry II: An Application to the Works of Shakespeare and Mralowe*. Literary and Linguistic Computing, 9:1-6,1994.

[14] Mosteller F. and Wallace D.L., Applied Bayesian and Classical Inference, The Case of the Federalist Papers, Addition-Wesley, Reading, 1964.

[15] Tweedie, F. J, Singh, S., and Holmes, D. I., *Neural Network Applications in Stylometry. The Federalist Papers*. Computers and the Humanities, 39(1), 1-10, 1996.


**Citation:** *G. MANIMANNAN & R. LAKSHMI PRIYA. (2019). Application of Tamil Syntax based on Authorship Attribution Using Neural Network and Classification Methods. International Journal of Scientific and Innovative Mathematical Research (IJSIMR), 7(6), pp.22-29. http://dx.doi.org/10.20431/2347-3142.0706004*