Research Journal of Applied Sciences, Engineering and Technology 8(2): 203-211, 2014
DOI:10.19026/rjaset.8.960
ISSN: 2040-7459; e-ISSN: 2040-7467
© 2014 Maxwell Scientific Publication Corp.
Submitted: March 14, 2014 Accepted: April 22, 2014 Published: July 10, 2014

Research Article

A Study of Authorship Attribution in English and Tamil Emails

[1]A. Pandian and [2]Mohamed Abdul Karim
[1]Department of MCA, SRM University, Chennai-603 203, India
[2]College of Applied Sciences, Sohar, Ministry of Higher Education, Oman

Abstract: The aim of our study is to identify author of unknown emails of Tamil and English. The recent approaches in Authorship Attribution show that apart from lexical measures some other features of written language are considerably effective as discriminators of author style. However, there have been no attempts to compare the attribution potential of these features. The aim of the present study, then, has to examine the effectiveness of several styles-markers in authorship attribution between the following two languages, English and Tamil equally important, however, we have to compare the usefulness of the chosen style-markers across a two languages the results proved high attribution effectiveness can be achieved in both the language.

Keywords: Echo state neural network, english emails, fishers linear discriminant method, lexical features, radial basis function, syntactic features, Tamil emails

INTRODUCTION

The study of identifying the owner (author) of Text/Email/Message/blog is called Authorship Attribution (AA). Currently, there were very few works on AA for Tamil emails (Bagavandas *et al*., 2009) have been done when compare to English. Previous authorship studies contain lexical, syntax (Grieve, 2007; Luyckx and Daelemans, 2008) structural and content-specific features, word based features including word length distribution, words per sentence and vocabulary richness were successful in earlier authorship studies. Syntactic features, called style markers, consist of all-purpose functional words.

The importance of text classification techniques rooted in machine learning marked as a pivotal turning point in authorship attribution studies. The use of such methods is straightforward: Training texts are used as labeled numerical vectors. They use learning methods to find boundaries between classes (authors) that minimize some classification function. The nature of the land boundaries depends on the learning method used. These methods facilitate the use of classes of boundaries that extend well beyond those implicit in methods that minimize distance. The earliest methods applied various types of neural networks using small sets of functional words as features. Graham *et al*. (2005) used neural networks on a wide variety of features. Other studies used k-nearest neighbor (Zhao and Zobel, 2005), rule learners (Koppel and Schler, 2003; Abbasi and Chen, 2005) and Bayesian regression (Genkin *et al*., 2007; Madigan *et al*., 2005; Argamon *et al*., 2003). Support Vector Machine (SVM) learning is suitable for text categorization as any other learning method and find the same for authorship attribution (De Vel *et al*., 2001; Diederich *et al*., 2003), Winnow (Koppel *et al*., 2002; Argamon *et al*., 2009).

The studies since that of Mosteller and Wallace have shown the use of function words for authorship attribution in different scenarios (Holmes *et al*., 2001a, b; Baayen *et al*., 2002; Binongo, 2003). Typical modern studies using function words in English use lists of a few hundred words, including pronouns, prepositions, auxiliary and modal verbs, conjunctions and determiners. Results of different studies using somewhat different lists of function words have been similar, indicating that the precise choice of function words is not crucial. For documents such as email formatting, structural features can be used for authorship attribution (Corney *et al*., 2002) (Fig. 1).

MATERIALS AND METHODS FOR ENGLISH EMAILS

Materials: The Table 1 describes the sequence of operations of the proposed system in this study for email authorship categorization. The proposed system is the combination of FLD and RBF algorithms.

Step 1: Emails have been used for Enron database.
Step 2: Tokenize the information of the enron emails. Create a dictionary of information. The template contains functional words like preposition, conjunctions, interjections, pronouns, verbs, adverbs, adjectives. This template has been used for filtering out

Corresponding Author: A. Pandian, Department of MCA, SRM University, Chennai-603 203, India

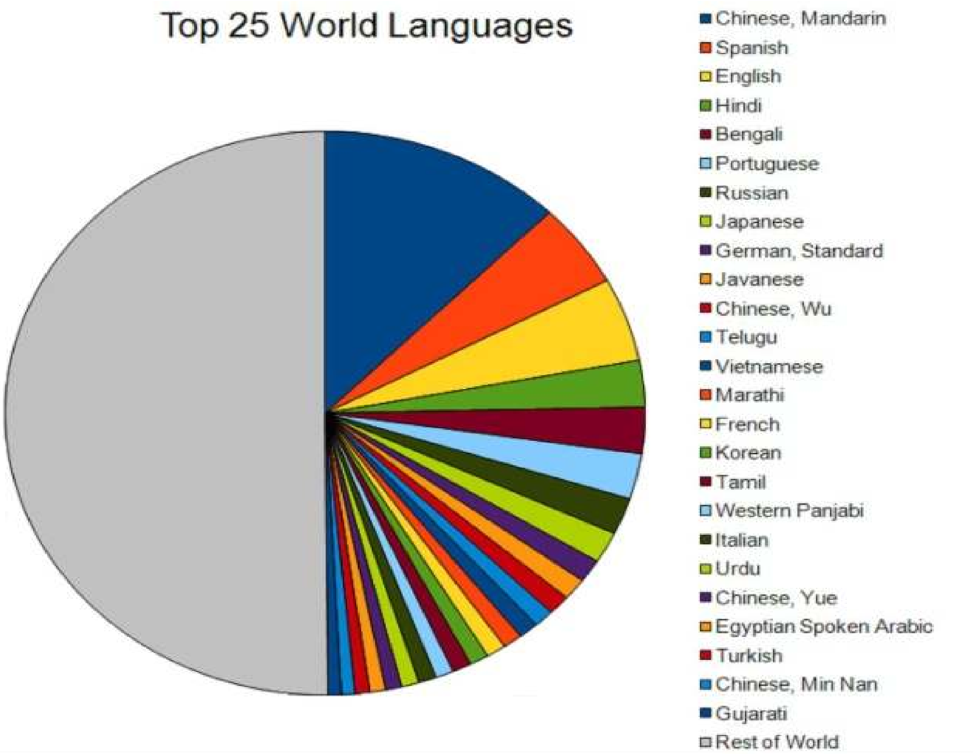

Fig. 1: Results of different studies of using function words in English

Table 1: Steps of the proposed system

Training the proposed system		
Step 1	Collecting emails	Enron dataset is used
Step 2	Preprocessing	Identifying words, filtering out the words in the email based on the dictionary of information available.
Step 3	Feature extraction	Character based, Word based and Syntactic based.
Step 4	Fisher's Linear discriminant method	Obtain projection vectors φ_1 and φ_2. Transform signature vector of higher dimension into 2-dimensional pattern for each email.
Step 5	RBF training	2-dimensional signature patterns are input to the RBF and final weights are obtained.
Testing the proposed system	Receive emails of an author not used for training the proposed system. Adopt step 2, step 3, step 4 and process with final weights obtained in step 5. Compare the output with a template to categorize the author.	

irrelevant information that will not be used for authorship analysis.

Step 3: Signature for each email is created by extracting features based on lexical characters, lexical words and syntactic properties. The total number of features for each email signature is 322. The details of the features (Farkhund *et al*., 2008, 2010) are as follows:

- Lexical analysis based on characters
- Total characters per line (NC)
- Ratio of digits to total characters (RD_T_C)
- Ratio of letters to total characters (RL_T_C)
- Ratio of uppercase letters to total characters (RUCL_T_C)
- Ratio of spaces to total characters (RS_T_C)
- Occurrences of alphabets to total characters (OA_T_C)
- Occurrences of special characters: < > j { } (OSC_T)
- Lexical word based analysis
- Number of Words (NW)
- Sentence length in terms of characters per line (SL)
- Average token length (ATL)
- Ratio of short words (1 to 3 characters) to T (RSWT)

- Ratio of word length frequency distribution of T (20 features) (RWLF)
- Average sentence length in terms of characters (ASLC)
- Ratio of characters in words to N (RCW)
- A word which occurs only once in the email document (SWO)
- A word which occurs only twice in the email document (TWO)
- Syntactic features
- Occurrences of punctuations (OP)
- Occurrences of function words (OFW)

Find the number of words and the number of occurrences (frequencies) an email and all the emails of authors. Create a matrix with rows equivalent to the total number of unique words extracted from all emails of all authors. The number of columns is equivalent to number authors. Fill up the columns with frequencies of words corresponding to respective authors. Each column is treated as a signature, which is further transformed into 2-dimensional pattern. A labeling is done for each pattern.

Step 4: The emails of each author are taken as a separate class. In this study, emails of100 authors are grouped into 100 classes. Fishers linear discriminant method is used to create two projection vectors φ_1 and φ_2. These projection vectors transform 322 dimensional signature into 2 dimensional pattern. Fifty emails for each author has been considered and hence a total of 5000 (50emails*100authors) signatures is obtained.

Step 5: Radial basis function with 75 centers (any other value) is used to learn 20% of emails of each author (Total of 10 emails X 100 authors = 1000 signatures) to get final weights. Many neural networks are available, however, we preferred RBF as it learns non linear data effectively.

Step 6: Testing the proposed system is done by using 80% of 50 emails per author (Total of 40 emails X 100 authors = 4000 signatures) are used. Step 2 to step 4 are adopted to obtain two dimensional signatures of the testing emails. Each signature is processed with the final weights obtained in step 5. The output of the RBF is used for categorization of the authorship of an email.

Methods:

Linear discriminant: Linear Discriminant Analysis (LDA) and the related Fisher's linear discriminant are methods used in statistics, pattern recognition and machine learning to find a linear combination of features which characterize or separate two or more classes of objects or events. The resulting combination may be used as a linear classifier. This linear classification can be fine tuned by applying a radial basis function on it. The mapping of the original vector 'X' onto a new vector 'Y' on a plane is done by a matrix transformation, which is given by:

$$Y = AX \tag{1}$$

where, X is the signatures and:

$$A = \begin{bmatrix} \varphi_1^T \\ \\ \varphi_2^T \end{bmatrix} \tag{2}$$

where,

φ_1 = A projection vector (also called a discriminant vector)

φ_2 = Another projection vector

The 2-dimensional pattern from the original 322-dimensional vector is denoted by 'y_i'. The vector 'y_i' is given by:

$$y_i = (u_i, v_i) = \{X_i^T \varphi_1, X_i^T, \varphi_2\} \tag{3}$$

The vector set 'y_i', is obtained by projecting the original signatures 'X' of the 5000 signature patterns onto the space spanned by φ_1 and φ_2 by using Eq. (3).

Radial basis function: The radial basis function is a supervised neural network, which uses a distance measure between the input pattern and the centers of the RBF nodes (Pandian and Sadiq, 2011). The summation of the distance is passed over an exponential activation function. This forms the outputs of the hidden nodes in the RBF network. A bias value is appended to the outputs of nodes in the hidden layer. The outputs of the hidden layer are processed with the labeled values (targets) assigned to obtain the final weights which will be used for testing.

RESULTS AND DISCUSSION FOR ENGLISH EMAILS

The plots in Fig. 2 to 4 define the characteristics of the emails of 100 authors based on the information mentioned in step 3. The email can be categorized to an author by averaging the signatures of the emails as shown in Fig. 2. The brown color plot shows the difference between the successive authors. The average difference is 0.3511 that indicates that the author can be categorized.

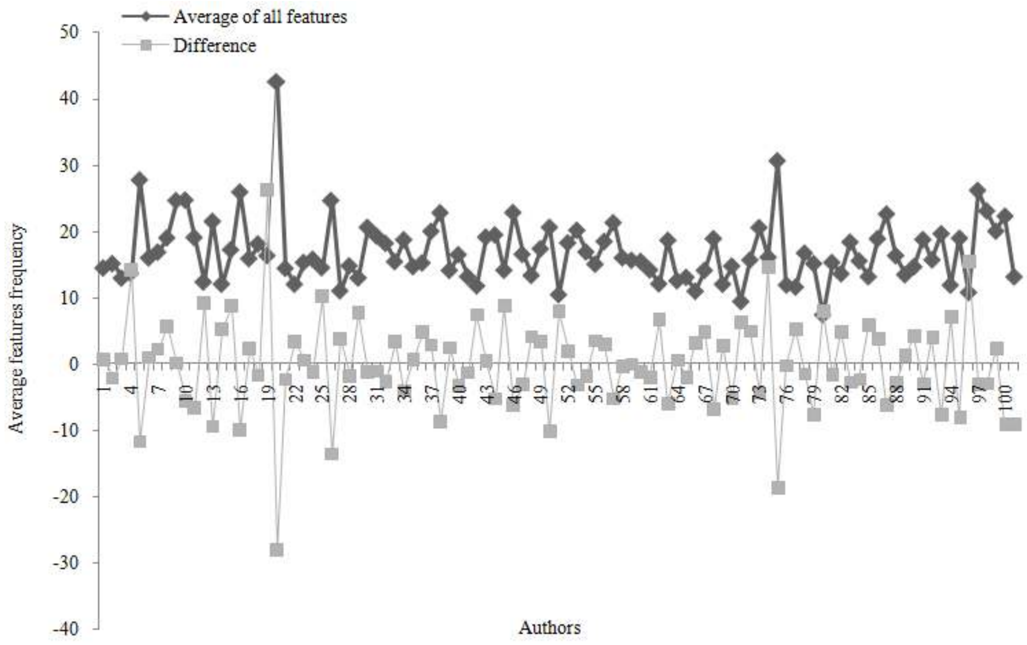

Fig. 2: Average frequency of all features

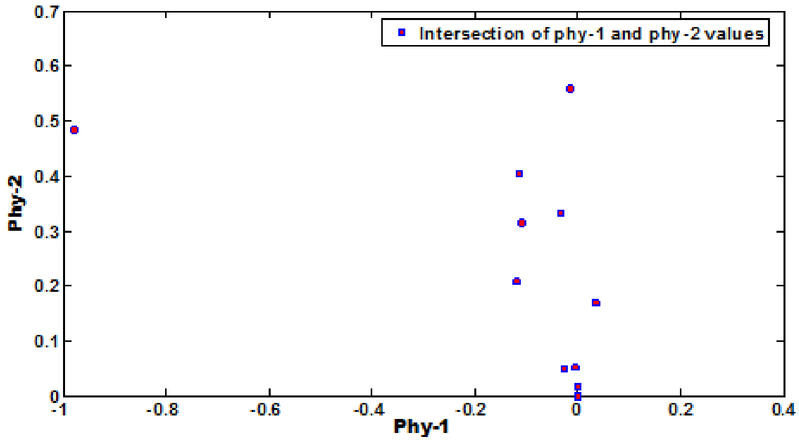

Fig. 3: φ_1 and φ_2 intersections

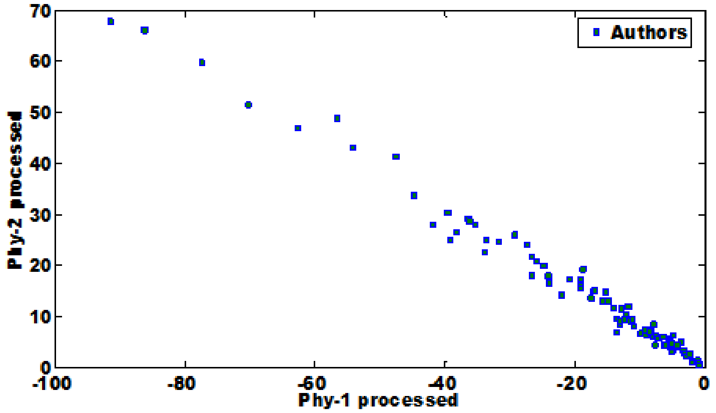

Fig. 4: Projected author patterns

Figure 3 presents the intersections of φ_1 and φ_2 projection vectors. In Fig. 3, signatures of 100 authors are projected using φ_1 and φ_2 vectors into 2-dimension. From this plot, very few authors' signatures overlap and the remaining authors' signatures are visible distinctly. In order to overcome the overlapping, RBF is used for correct categorization.

RBF network is trained with projected signature patterns along with labeling. A final weight matrix is obtained which is further used to test the untrained emails. The outputs of RBF are categorized to a trained authors database else, the email is categorized to some other author outside the database.

Problem statement and objectives for Tamil emails:

The problem focused in this study is as follows:

- Suspicious Tamil email is under consideration. The Writing Style (WS) in this suspicious email has WS of one author or more than one author. The number of suspects can be (S_1, S_2, …S_n).
- The $WS_{1\text{-}N}$ is available in the database repository (R).

Initial approach: Extract the WS of N authors using lexical, syntactic methods.

Cluster the WS of emails of each author and check for separability among the authors.

To enhance the identification of an anonymous author of the suspected email, apply reduction of the WS signature of the authors of higher dimension to 2 dimensions.

Subsequently, use Radial Basis Function (RBF) and Echo state Neural network (ESNN) for identification of the authors.

The objectives of the second part of this study are to present:

- Better representable feature extraction of Tamil characters from the Tamil email.
- Reducing the size of signature pattern using Fishers linear discriminant function.
- Implementation of radial basis function neural network and Echostate neural network for AA

MATERIALS AND METHODS FOR TAMIL EMAILS

Materials: Table 2 to 4 present words used for filtering the Tamil email and analyze for unique information.

Work words will analyze how an author writes email and what clarity is present in the email. The number of work words will indicate performance task requirements in an unambiguous manner. Action words indicate some actions present in the email.

Preposition-1, preposition-2, preposition-3, preposition-4, adjectives, adverbs and conjunctions have their standard meanings.

The total number of words used as basic dictionary is 1571 (work+action+prepositions+adjectives+ adverbs+conjunctions). The numbers mentioned in parenthesis are the total in each category, whereas, only few words are given in the Table 2 to 4.

Table 2: Sample words used for filtering

Work (70)	Action (524)	Preposition_1 (94)	Preposition_2 (30)
Analyze (ஆராய்தல்)	Accelerate (துரிதப்படுத்து)	Aboard (கப்பல் மீது)	According to (அதன்படி)
annotate (உரை எழுதிச் சேர்)	accommodate (தகுதியான படி எற்பாடு செய்)	about (கிட்டத்தட்ட)	ahead of ((முன்னால்)
ascertain (உறுதிசெய்)	accomplish (நிறைவேற்று)	above (மேலே)	as of (அப்படியே உடைய)
attend (உடனிரு)	accumulate (சிறுகச் சிறுகச் சேர்)	absent (வராத)	as per (அப்படியே ஒன்றிற்கு)
audit (தணிக்கை)	achieve (செய்து முடி)	across (எதிர்ப்பக்கத்தில்)	as regards (அப்படியே அக்கறைக் காட்டு)
build (கட்டு)	acquire (கையகப்படுத்து)	after (பின்னர்)	aside from (அப்பால்இருந்து)
calculate (கணக்கிடு)	act (சட்டம்)	against (எதிராக)	because of (காரணத்தால்)
consider (பரிசீலனை செய்)	activate (ஊக்குவி)	along (எப்போதும்)	close to (அருகில்)
construct (கட்டுதல்)	adapt (பொருந்தச்செய்)	alongside (எப்போதும் பக்கம்)	due to (காரணமாக)
control (கட்டுப்பாடு)	add (கூட்டு)	amid (இடையில்)	except for (தவிர பதிலாக)

Table 3: Sample words used for filtering

Preposition_3 (16)	Preposition_4 (9)	Pronoun (77)	Adjectives (395)
As far as (சாத்தியமான இயலும் முடியும் அளவில்)	Apart from (தனியாக இருந்து)	All (அனைத்தும்)	Early (முன்னதாக)
As well as (அத்துடன்)	but (ஆனால்)	another (மற்றொன்று)	abundant (எக்கச்சக்கமான)
by means of (அதன்மூலம்)	except (தவிர)	any (ஏதேனும்)	adorable (வணங்குவது)
in accordance with (விதிப்படி)	plus (கூட்டல்)	anybody (யாராவது)	adventurous (துணிவான செயல்)
in addition to (அதோடு சேர்த்து)	save (சேமி)	anyone (எவரும்)	aggressive (பகைமை உணர்வுடன் தாக்க முற்படும்)
in case of (ஒரு வேளை)	concerning (அக்கறை)	anything (எதுவும்)	agreeable (சம்மதி)
in front of ((முன்னால்)	considering (ஆழ்ந்து ஆராய்)	both (இருவரும்)	alert (எச்சரிக்கையான)
in lieu of (பதிலாக)	regarding (அக்கறைக் காட்டு)	each (ஒவ்வொன்று)	alive (உயிருடன்)
in place of (பதிலாக)	worth (மதிப்பு)	each other (ஒருவருக்கொருவர்)	amused (பயன்படுத்தப்பட்ட)
in point of (இடத்தில் மையம் உடைய)		either (இது அல்லது அது)	ancient (பழமையான)

Table 4: Sample words used for filtering

Adverbs (331)	Conjunctions (25)
Abnormally (அசாதாரணமான)	and (மற்றும் மேலும்)
absentmindedly (கவனக் குறைவான)	but (ஆனால்)
accidentally (எதிர்பாரா நிகழ்ச்சி)	for (பதிலாக)
(acidly) (அமிலம்)	(nor) (அன்றியும்)
actually (உண்மையில்)	or (அல்லது)
Adventurously (துணிவான)	so (எனவே)
afterwards (பிற்பாடு)	yet (இன்னும்)
almost (கிட்டத்தட்ட)	after (பின்னர்)
always (எப்பொழுதும்)	although (ஆயினும்)
angrily (கோபமாக)	As (அப்படியே)

The sequence of operations of the proposed system in this study for Tamil email authorship association is as follows:

Step 1: Tamil emails of 50 authors are considered. Ten emails of each author has been considered.

Step 2: Tokenize the information of the emails. Create a dictionary of information. The template contains function words like prepositions (முன்னிடைச்சொல்), conjunctions (உசாத்துணை), interjections (கூட்டுச்சொற்கள்), pronouns (இடப்பெயர்கள்), verbs (சாரியை), adverbs, adjectives (உரிச்சொற்கள்). Use this template for filtering out irrelevant information that will not be used for AA.

Step 3: Created signature for each Tamil email by extracting features based on lexical characters, lexical words and syntactic properties. The total number of features for each email signature is 322. The details of the features (Farkhund *et al*., 2008; 2010) are as follows (Table 5)

Obtain the number, of words and the number of occurrences (frequencies) of information in emails.

Create a matrix with rows equivalent to the total number of unique words extracted from all emails of all authors. The number of columns is equivalent to number authors.

Fill up the columns with frequencies of words corresponding to respective authors.

Treat each column as a signature. Do a labeling for each 2-dimensional pattern.

Step 4: Take the emails of each author as a separate class. In this study, we group emails of 50 authors into 50 classes. Create two projection vectors φ_1 and φ_2 using Fishers linear discriminant method. These projection vectors transform 322 dimensional signature into 2

Table 5: Features used in this study

Feature	Abbreviation	Tamil characters
Lexical analysis based on characters		
Total characters in email (TC)		
Ratio of Vowels (V) /TC	R_V_TC	அ, ஆ, இ
Ratio of Consonants (CO) /TC	R_CO_TC	க், ங், ச்
Ratio of Compound form (CF) /TC	R_CF_TC	க, கா, கி, கீ
Ratio of [digits/TC]	R_D_TC	
Ratio of [letters/TC]	R_L_TC	
Ratio [spaces/TC]	R_S_TC	
Ratio of [compound type 1/TC]	R_C1_TC	
Ratio of [compound type 2/TC]	R_C2_TC	
Ratio of [compound type 3/TC]	R_C3_TC	
Occurrences of special characters	OSC_T	< > j { }
Lexical word based analysis		
Number of words	NW	
Average token length (word_length)	ATL	
Ratio short words (1 to 3 characters) to T	RSWT	
Ratio of word length frequency distribution of T (20 features)	RWLF	
Average sentence length in terms of characters	ASLC	
Ratio characters in words to N	RCW	
A word which occurs only once in the email document	SWO	
A word which occurs only twice in the email document	TWO	
Syntactic features		
Occurrences of punctuations	OP	
Occurrences function words	OFW	

dimensional pattern. We consider ten emails for each author and hence obtain a total of 500 (10 Tamil emails X 50 authors) signatures.

Step 5: Training of Radial basis function is done separately with 75 centers (any other value) in the hidden layer. Similarly, training of ESNN is done separately with 21 reservoirs in the hidden layer. In each case, 20% of the emails are used (Total of 2 emails X 50 authors = 100 signatures) to get final weights.

Step 6: Testing RBF and ESNN is done separately. Eighty percent of 10 emails per author (Total of 8 emails X 50 authors = 400 signatures) are used. Adopt step 2 to step 4 to obtain two dimensional signatures of the testing emails. Process each signature with the final weights obtained in step 5. Use the outputs of the RBF/ESNN for AA

Methods:

Echo State Neural Network (ESNN): The echo state neural network is a recurrent network (Jaeger, 2001a, b; Purushothaman and Suganthi, 2008). The echo state condition is the spectral radius (the largest among the absolute values of the eigenvalues of a matrix, denoted by (|| ||) of the reservoir's weight matrix (||W|| <1). This condition states that the input controls the dynamics of the ESNN and the effect of the initial states vanishes. The current design of ESNN parameters relies on the selection of the spectral radius. There are many possible weight matrices with the same spectral radius. They do not perform at the same level of mean square error (MSE) for functional approximation.

RESULTS AND DISCUSSION FOR TAMIL EMAILS

The implementation of FLD, RBF and ESNN is done using Matlab 10. The plots in Fig. 2 to 8 define

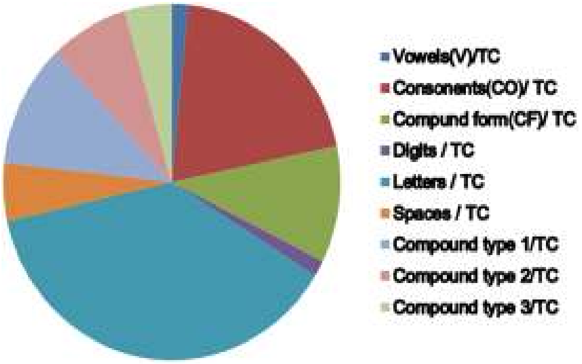

Fig. 5: Pi chart for distribution of Tamil letters

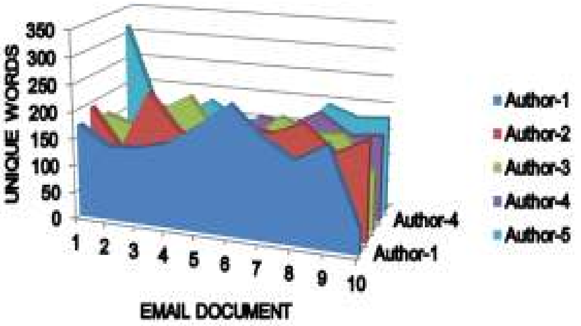

Fig. 6: Unique words in Tamil email

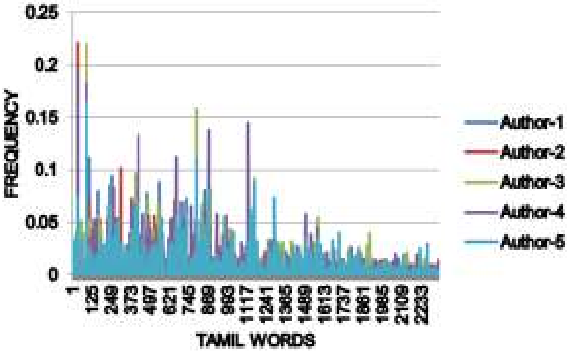

Fig. 7: Normalized word frequencies

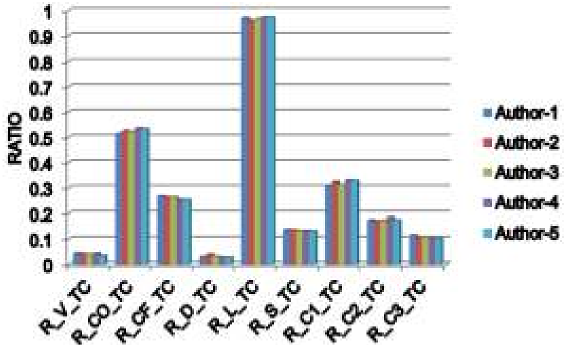

Fig. 8: Lexical analysis

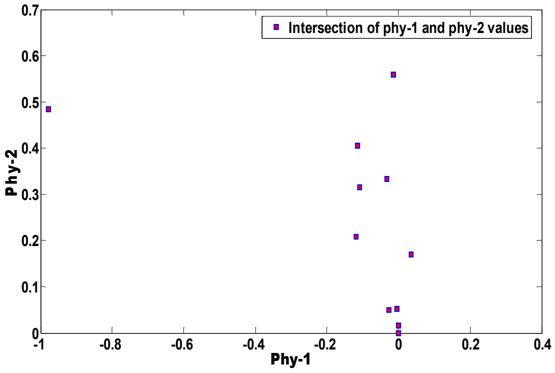

Fig. 9: φ_1 and φ_2 intersections

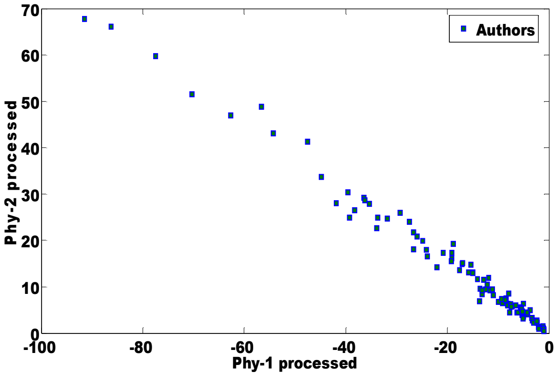

Fig. 10: Projected author patterns

the characteristics of the Tamil emails of 5 sample authors based on the information mentioned in step 3 of above section.

Figure 9 presents the plot φ_1 and φ_2 projection vectors obtained using Eq. (2). Figure 10 presents the plots of (u, v) using Eq. (3) for 50 authors (each 2 emails). There is overlap as many points as shown in Fig. 10.

CONCLUSION

In the first part of this study, there is overlapping of a few authors (Fig. 3), RBF has been used. Advantages of the proposed system are as follows:

- The size of the 322-dimensional signature pattern is reduced to 2-dimension.
- The training of the RBF is faster with less computational complexity.
- The size of the RBF topology is reduced from 322 to 2 in the input layer.
- Since, the activation function used in RBF is non-linear, the overlapping problem is solved.

The second part of this study presents Tamil email AA uses FLD with RBF and, FLD with ESNN. As there is overlapping of a few authors (Fig. 10), there is still some mismatching of results.

From the above results we have utilized the RBF and FLD in both English Emails and Tamil Emails. In addition to that, ESNN used in Tamil Emails. By this way in the future, the same methods we can try in the other languages having the most valuable ancient texts. In the future, we can try our above specified method to apply in a single bilingual document also.

REFERENCES

Abbasi, A. and H. Chen, 2005. Applying authorship analysis to extremist-group Web forum messages. IEEE Intell. Syst., 20(5): 67-75.

Argamon, S., M. Koppel, J. Fine and A. Shimoni, 2003. Gender, genre and writing style in formal written texts. Text Talk, 23(3).

Argamon, S., M. Koppel, J. Pennebaker and J. Schler, 2009. Automatically profiling the author of an anonymous text. Commun. ACM, 52(2): 119-123.

Baayen, H., H. van Halteran, A. Neijt and F. Tweedie, 2002. An experiment in authorship attribution. Proceeding of 6es Journ´ees Internationales d'Analyse Statistique Des Donn´Ees Textuelles (JADT, 2002).

Bagavandas, M., H. Abdul and G. Manimannan, 2009. Neural computation in authorship attribution: The case of selected Tamil articles. J. Quant. Linguist., 16(2): 115-131.

Binongo, J.N.G., 2003. Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution. Chance, 16(2): 9-17.

Corney, M., O. de Vel, A. Anderson and G. Mohay, 2002. Gender-preferential text mining of e-mail discourse. Proceedings of the 18th Annual Computer Security Applications Conference (ACSAC '02), pp: 282.

De Vel, O., A. Anderson, M. Corney and G. Mohay, 2001. Mining e-mail content for author identification forensics. SIGMOD Rec., 30(4): 55-64.

Diederich, J., J. Kindermann, E. Leopold and G. Paass, 2003. Authorship attribution with support vector machines. J. Appl. Intell. Arch., 19(1-2): 109-123.

Farkhund, I., H. Binsalleeh, B.C.M. Fung and M. Debbabi, 2010. Mining writeprints from anonymous e-mails for forensic investigation. Digit. Invest., 7: 56-64.

Farkhund, I., R. Hadjidj, B.C.M. Fung and M. Debbabi, 2008. A novel approach of mining write-prints for authorship attribution in e-mail forensics. Digit. Invest., 5: S42-51.

Genkin, A., D.D. Lewis and D. Madigan, 2007. Large-scale Bayesian logistic regression for text categorization. Technometrics, 49(3): 291-304.

Graham, N., G. Hirst and B. Marthi, 2005. Segmenting documents by stylistic character. Nat. Lang. Eng., 11(4): 397-415.

Grieve, J., 2007. Quantitative authorship attribution: An evaluation of techniques. Lit. Linguist. Comput., 22(3): 251-270.

Holmes, D.I., L. Gordon and C. Wilson, 2001a. A widow and her soldier: Stylometry and the american civil war. Lit. Linguist. Comput., 16(4): 403-420.

Holmes, D.I., M. Robertson and R. Paez, 2001b. Stephen Crane and the New-York tribune: A case study in traditional and non-traditional authorship attribution. Comput. Humanities, 35(3): 315-331.

Jaeger, H., 2001a. Short term memory in echo state networks. GMD Report 152, German National Research Center for Information Technology, German.

Jaeger, H., 2001b. The echo state approach to analyzing and training recurrent neural networks. GMD Report 148, German National Research Center for Information Technology, German.

Koppel, M. and J. Schler, 2003. Exploiting stylistic idiosyncrasies for authorship attribution. Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis, pp: 69-72.

Koppel, M., S. Argamon and A.R., Shimoni, 2002. Automatically categorizing written texts by author gender. Lit. Linguist. Comput., 17(4): 401-412.

Luyckx, K. and W. Daelemans, 2008. Authorship attribution and verification with many authors and limited data. Proceeding of the 20th Belgian-Netherlands Conference on Artificial Intelligence (BNAIC, 2008). Enschede, Netharlands.

Madigan, D., A. Genkin, D.D. Lewis, S. Argamon, D. Fradkin and L. Ye, 2005. Author identification on the large scale. Proceeding of the Meeting of the Classification Society of North America.

Pandian, A. and A.K. Sadiq, 2011. Email authorship identification using radial basis function. Int. J. Comput. Sci. Inform. Secu., 9: 68-75.

Purushothaman, S. and D. Suganthi, 2008. fMRI segmentation using echo state neural network. Int. J. Image Process., 2(1): 1-9.

Zhao, Y. and J. Zobel, 2005. Effective authorship attribution using function word. Proceeding of the 2nd Asian Information Retrieval Symposium (AIRS, 2005), AIRS, Springer, USA, pp: 174-190.