Journal of Computer Science 10 (6): 1003-1014, 2014
ISSN: 1549-3636
© 2014 Science Publications
doi:10.3844/jcssp.2014.1003.1014 Published Online 10 (6) 2014 (http://www.thescipub.com/jcs.toc)

AUTHORSHIP CATEGORIZATION IN EMAIL INVESTIGATIONS USING FISHER'S LINEAR DISCRIMINANT METHOD WITH RADIAL BASIS FUNCTION

[1]Pandian, A. and [2]Md. Abdul Karim Sadiq

[1]Department of MCA, SRM University, Chennai, India-603203, India
[2]Ministry of Higher Education, College of Applied Sciences, Sohar, Sultanate of Oman

Received 2012-02-10; Revised 2012-05-12; Accepted 2014-01-31

ABSTRACT

Email plays a vital role in faster communication. Lots of mails are sent to common public with falsified information that appears to be a realistic. It is mandatory to trace the origin of the email and the authors/systems responsible for generating such emails. Representative signatures of email are to be generated using lexical and syntactic based methods. The signature of each email has huge dimensions and is called a vector/pattern. In order to make it convenient for subsequent processing, the huge dimension of the signature is converted into 2-dimensional pattern using Fisher's Linear Discriminant Function (FLD). The 2-dimensional patterns of the signatures of emails under consideration are used as training data for the Radial Basis Function (RBF) network which can learn non-linear data. The classification of email is very well achieved due to transformation by FLD and training by RBF. The proposed method helps in building signature database for accurate categorization in email forensics. The proposed combination of algorithms helps in clustering the different emails generated by an author or by a system.

Keywords: Projection Vectors, Email, Lexical Features, Syntactic Features, Fisher's Linear Discriminant Function, Radial Basis Function

1. INTRODUCTION

Previous authorship studies (Zheng *et al*., 2006; Stamatatos, 2009) contain lexical, syntactic (David, 1992; Grieve, 2007; Luyckx and Daelemans, 2008), structural and content-specific features. Lexical features are used to learn about the preferred use of isolated characters and words of an individual. Word-based features including word length distribution, words per sentence and vocabulary richness were very effective in earlier authorship studies. Syntactic features, called style markers, consist of all purpose functional words such as 'though', 'where', 'your' and punctuations like '!' and ':'.

The objective of this study is to create signatures for each email using lexical, syntactic methods. The signature represents uniqueness for each email and hence grouping of emails of an author is enhanced. The information in the email is based upon the thoughts an author. If it were handwritten, then it is still more easier to identify the author. However, the same behavior is reflected in the email created by the same author except the non availability of the handwritten signature in the email. This property really helps in identifying the author using the unique signature of the email.

2. MATERIALS AND METHODS

2.1. 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 from enron database.
Step 2: Tokenize the information of the enron emails. Create a dictionary of information. The template contains functional words like preposition,

Corresponding Author: Pandian, A., Department of MCA, SRM University, Chennai, India-603203, India Tel: 00-91-9150354754

conjunctions, interjections, pronouns, verbs, adverbs, adjectives. This template has been used for filtering out 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 to T (20 features) (RWLF)
- Average Sentence Length in terms of Characters (ASLC)
- Ratio of characters in words to N (RCW)
- Word which occurs only once in the email document (SWO)
- 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 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 is taken as a separate class. In this study, emails of 100 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 (50 emails*100 authors) signatures are 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×100 authors = 1000 signatures) to get final weights. Many neural networks are available, however, we preferred RBF as it learns non linear data effectively.

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 RBF and final weights are obtained
Testing the proposed system	Receive email of an author not used for training the proposed system. Adopt step 2, step3, step 4 and process with final weights obtained in step 5. Compare the output with template to categorize the author	

Step 6: Testing the proposed system is done by using 80% of 50 emails per author (Total of 40 emails×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.

2.2. Methods

2.2.1. Linear Discriminant

Linear Discriminant Analysis (LDA) (Sambasiva *et al.*, 2009) 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 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 Equation (1 and 2):

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

where, X is the signatures and:

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

φ_1 is a projection vector (also called a discriminant vector) and φ_2 is 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:

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

The vector set ‘yi’ , is obtained by projecting the original signatures ‘X’ of the 5000 signature patterns onto the space spanned by φ1 and φ2 by using Equation (3).

2.3. Radial Basis Function

The radial basis function is a supervised neural network which uses 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 is processed with the labeled values (targets) assigned to obtain the final weights which will be used for testing.

3. RESULTS

The plots in **Fig. 1-13** 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. 14**. The brown color plot shows the difference among the successive authors. The average difference is 0.3511 that indicates that the author can be categorized.

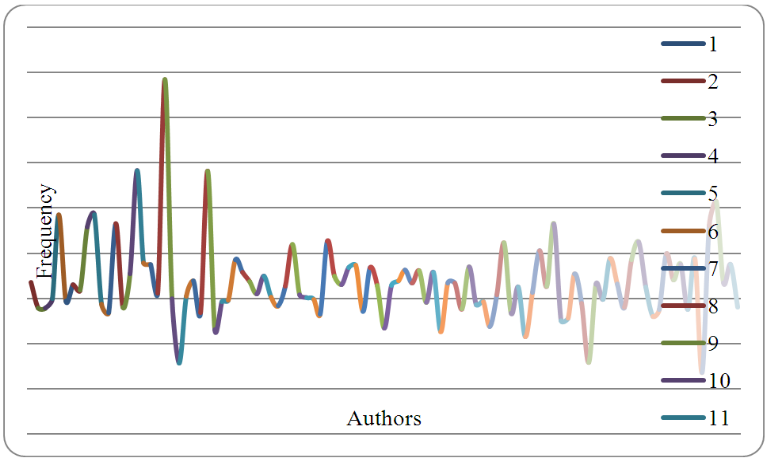

Fig. 1. Frequency of actual characters count in a line

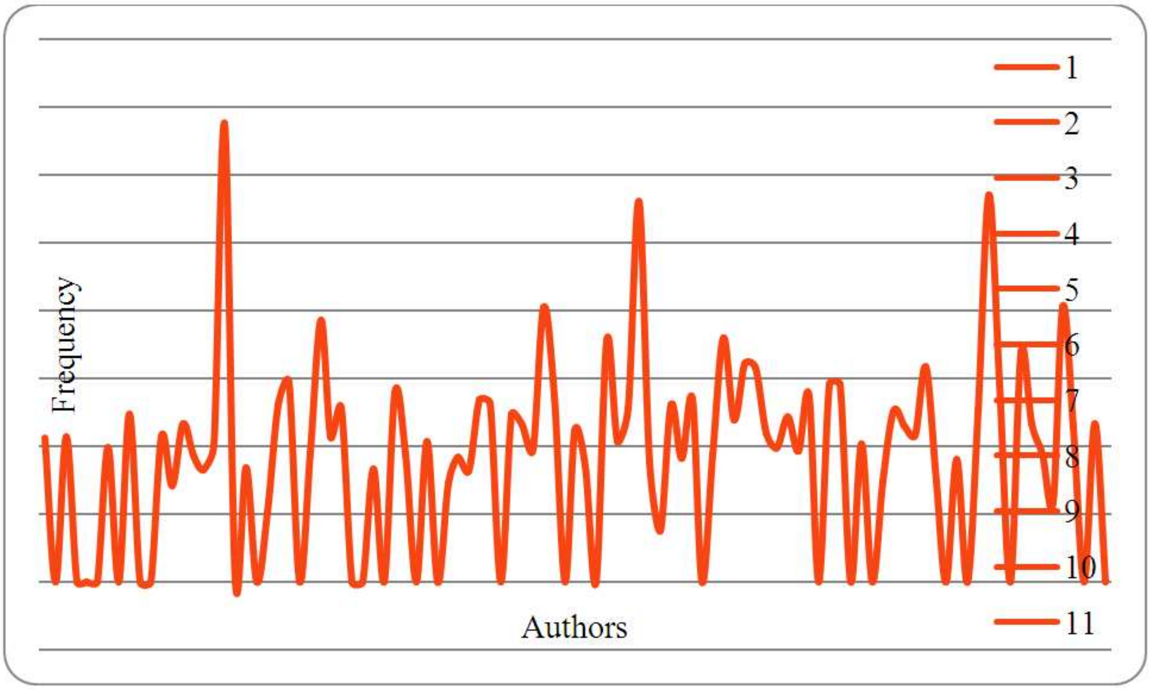

Fig. 2. Frequency of ratio of digits to total characters

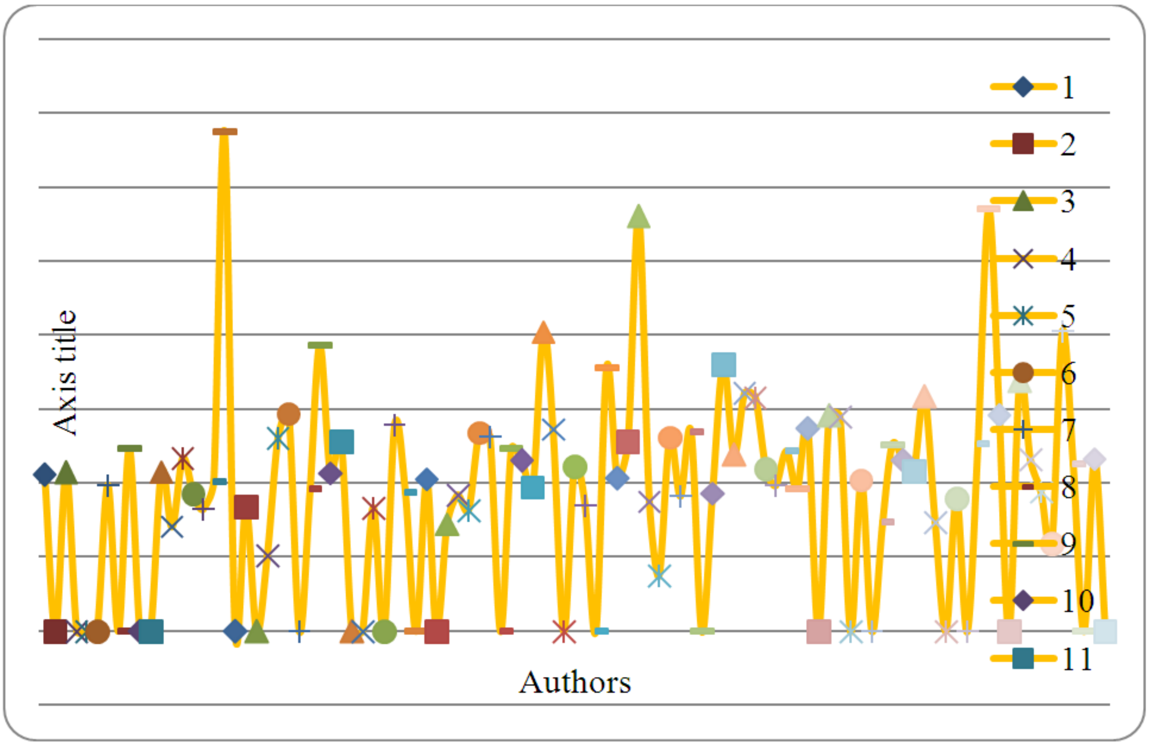

Fig. 3. Frequency of ratio of letters to total characters

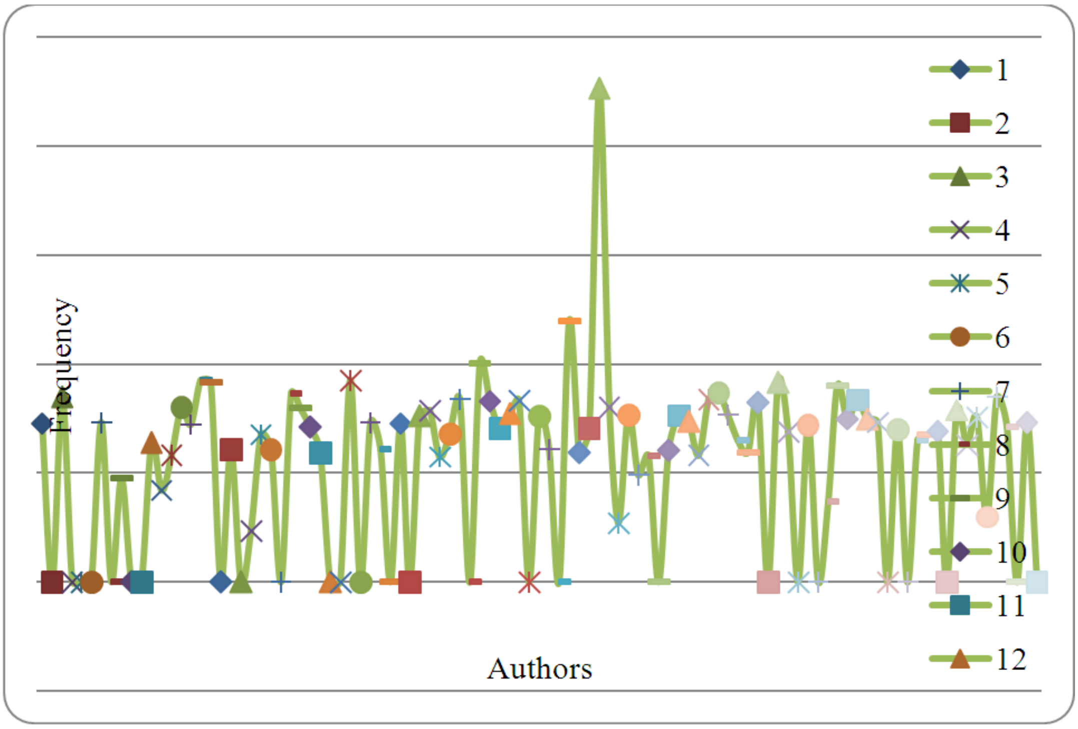

Fig. 4. Frequency of ratio of upper case letters to total characters

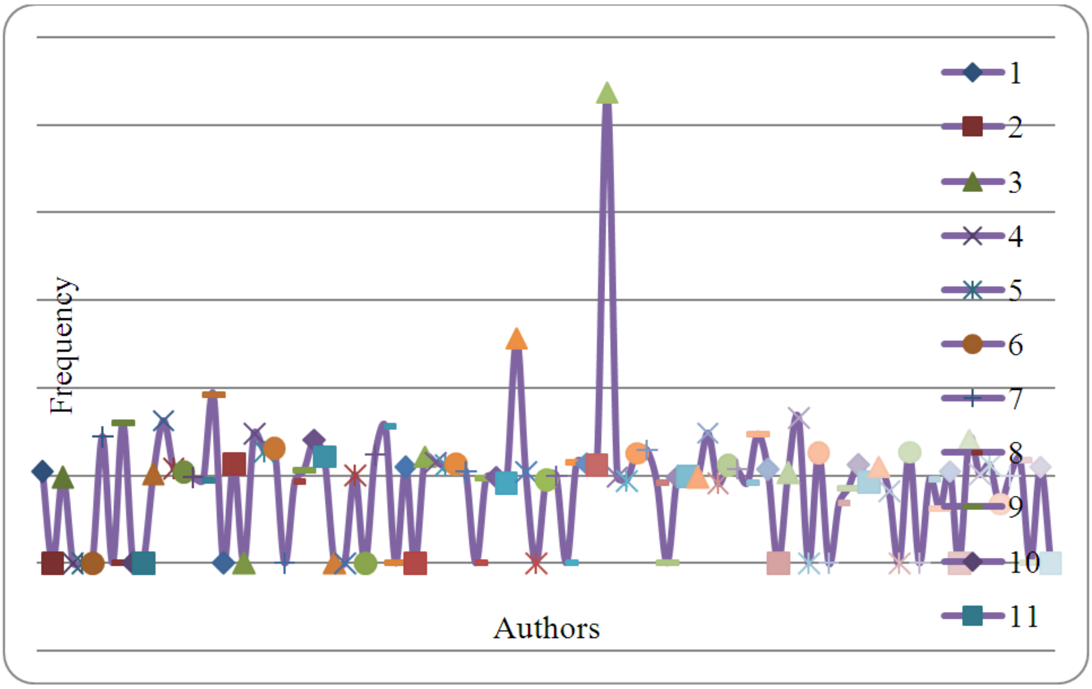

Fig. 5. Frequency of ratio of spaces to total characters

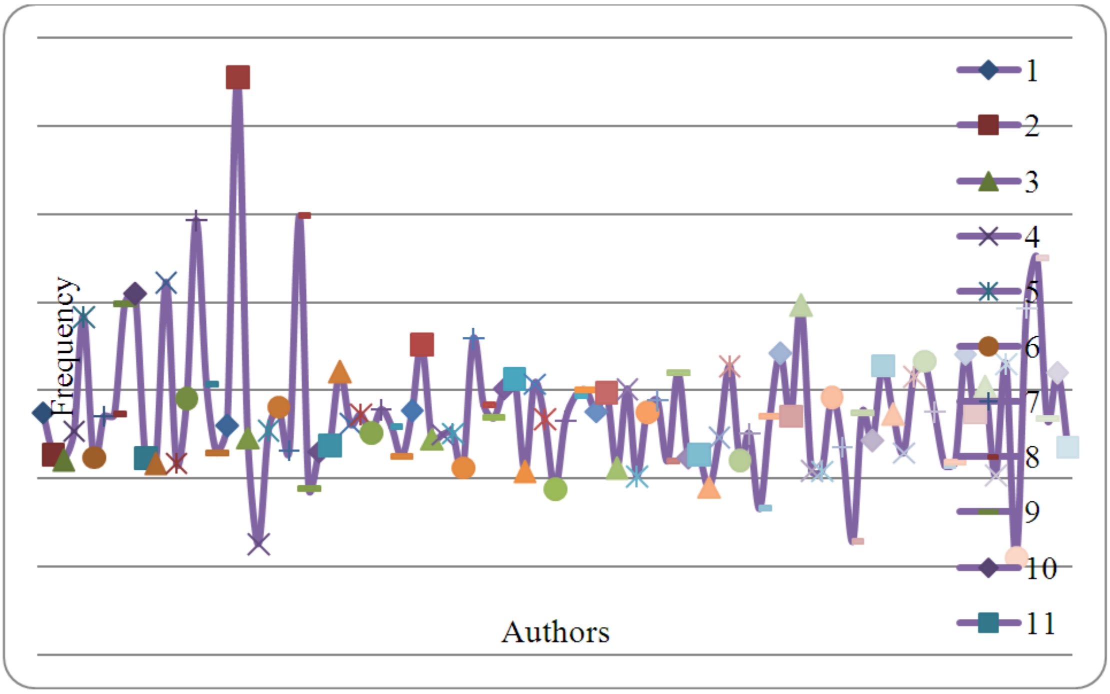

Fig. 6. Frequency of occurrences of alphabets

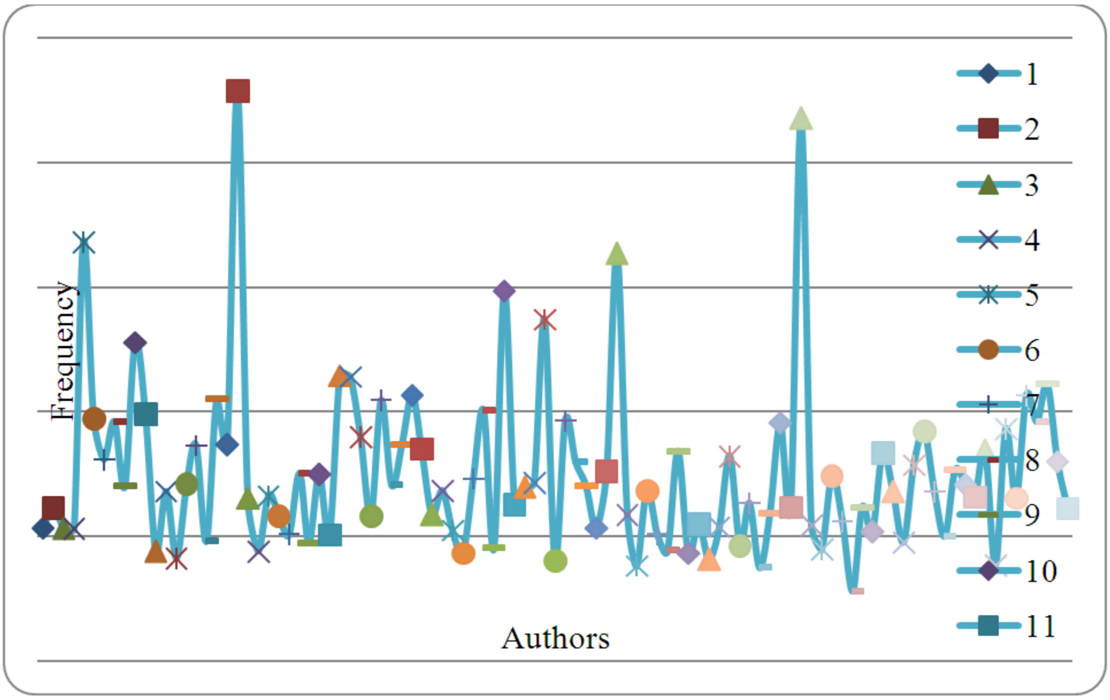

Fig. 7. Frequency of number of words

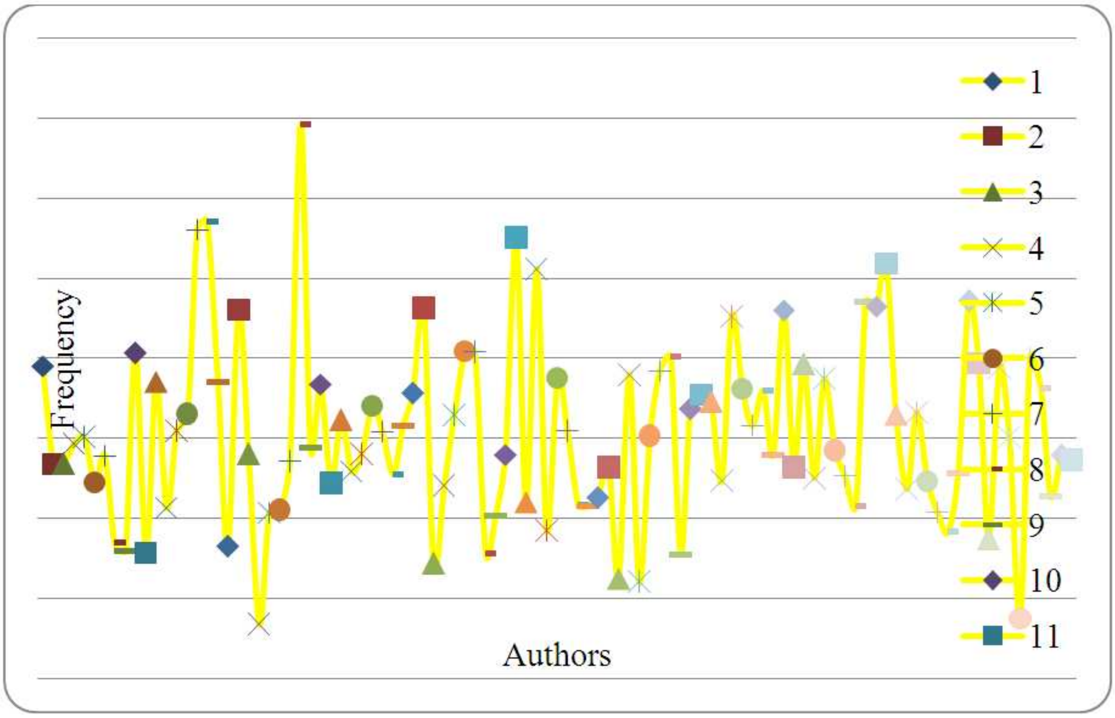

Fig. 8. Frequency of average sentence length

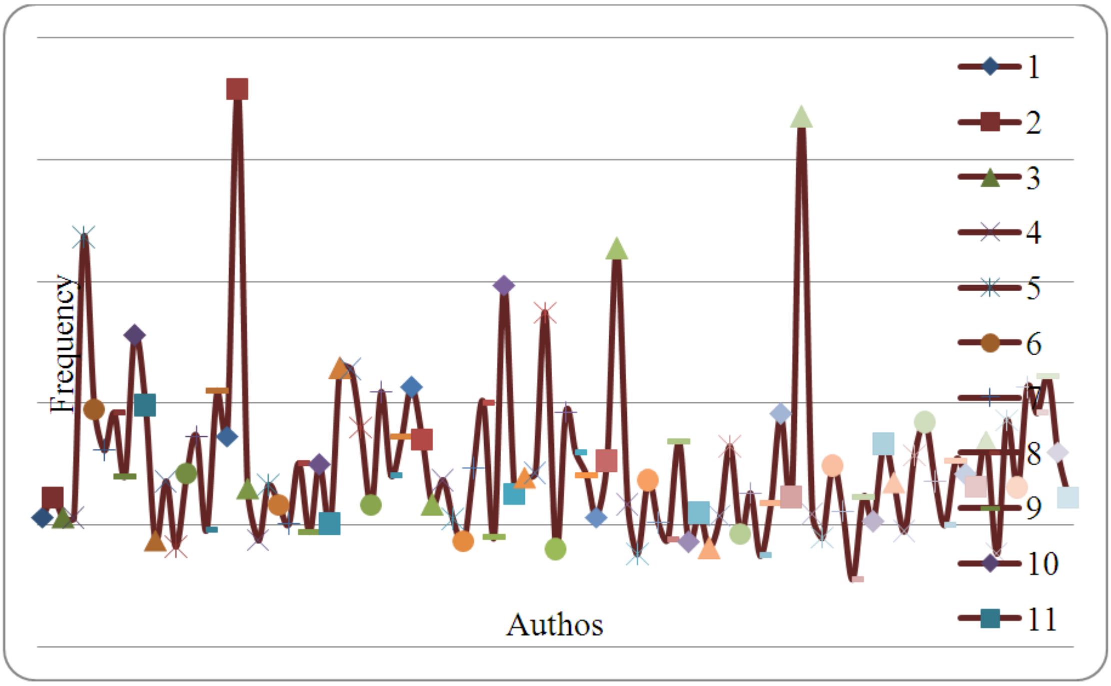

Fig. 9. Frequency of average token length

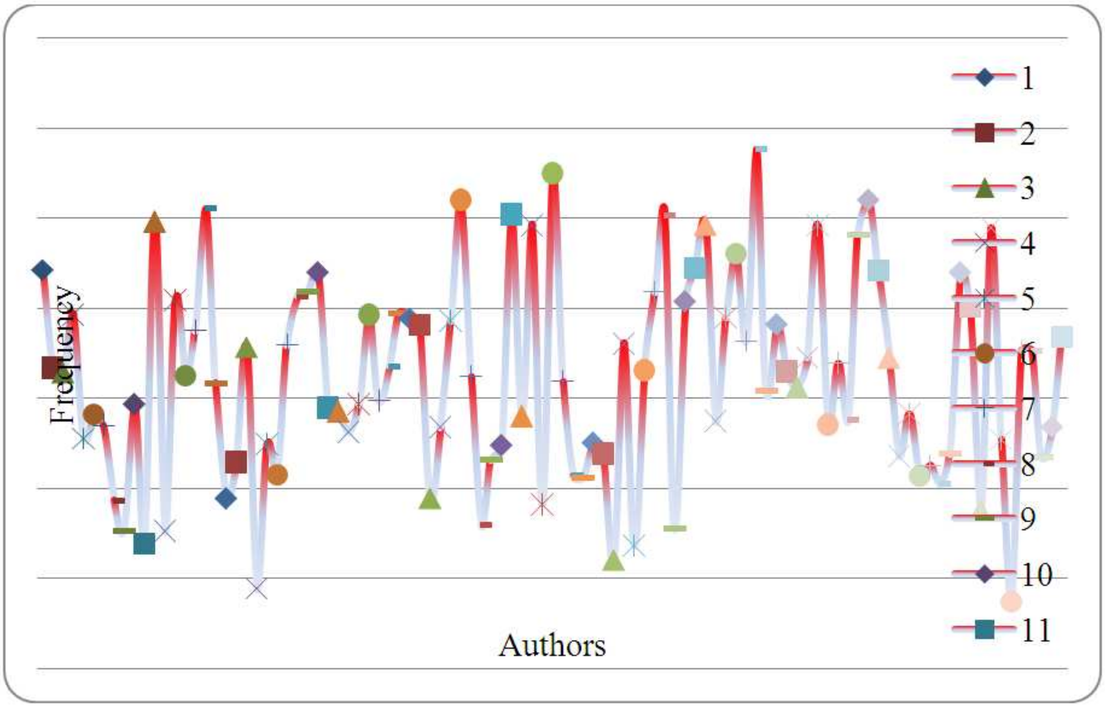

Fig. 10. Frequency of ratio of characters in words in N

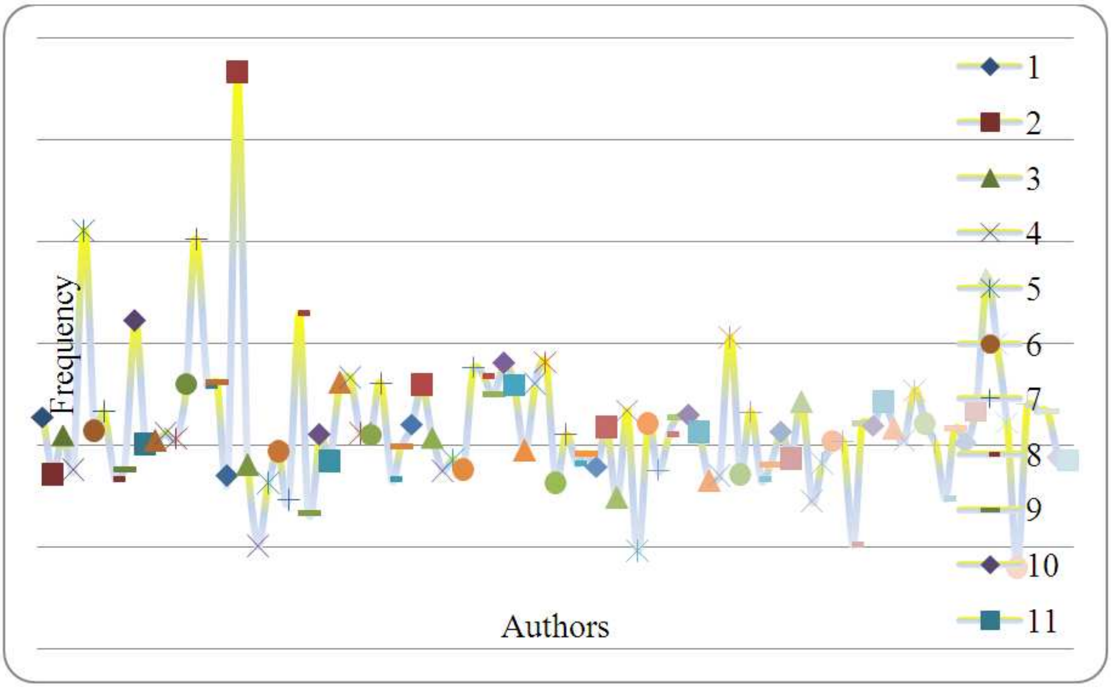

Fig. 11. Frequency of occurrences of punctuations

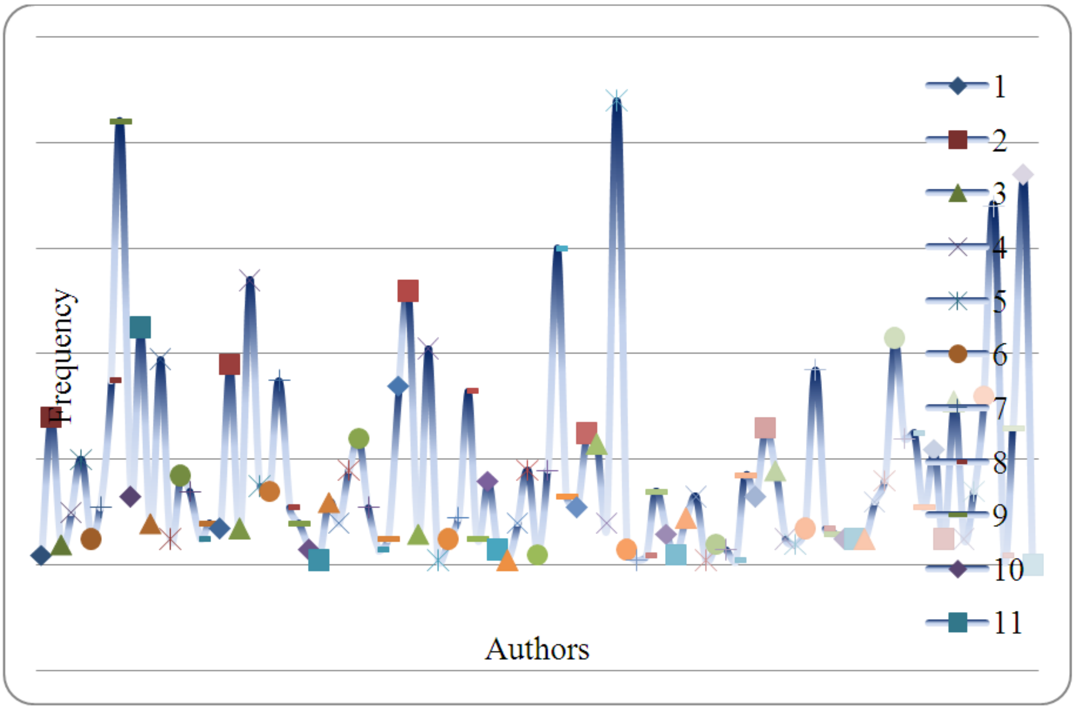

Fig. 12. Frequency of number of vowels

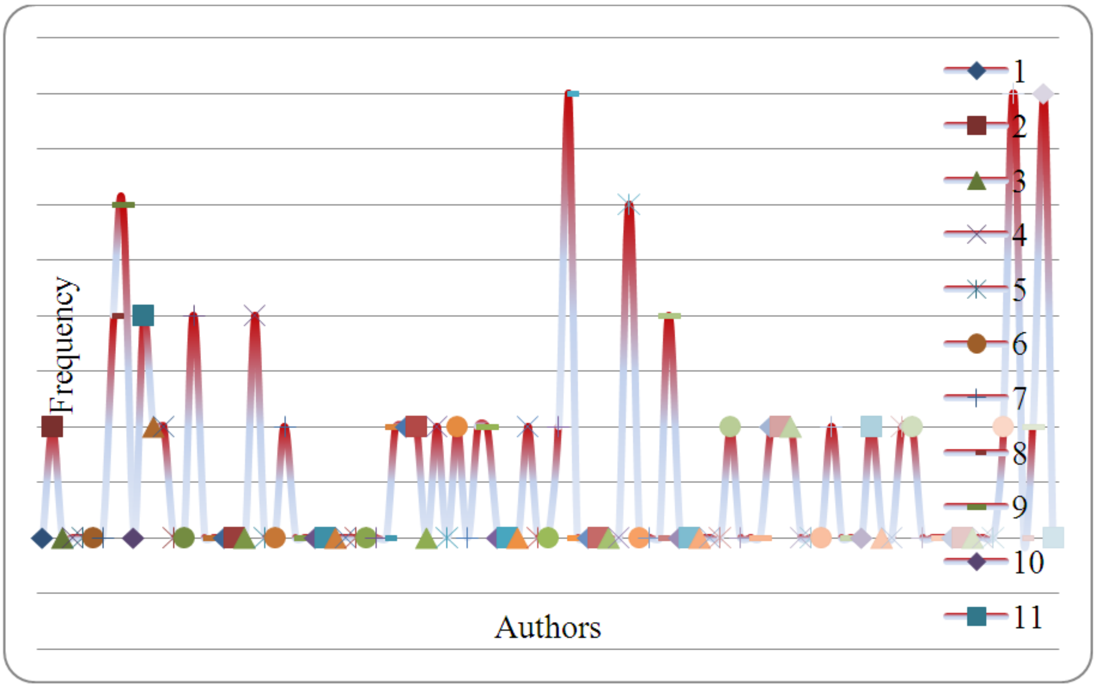

Fig. 13. Frequency of words that define work

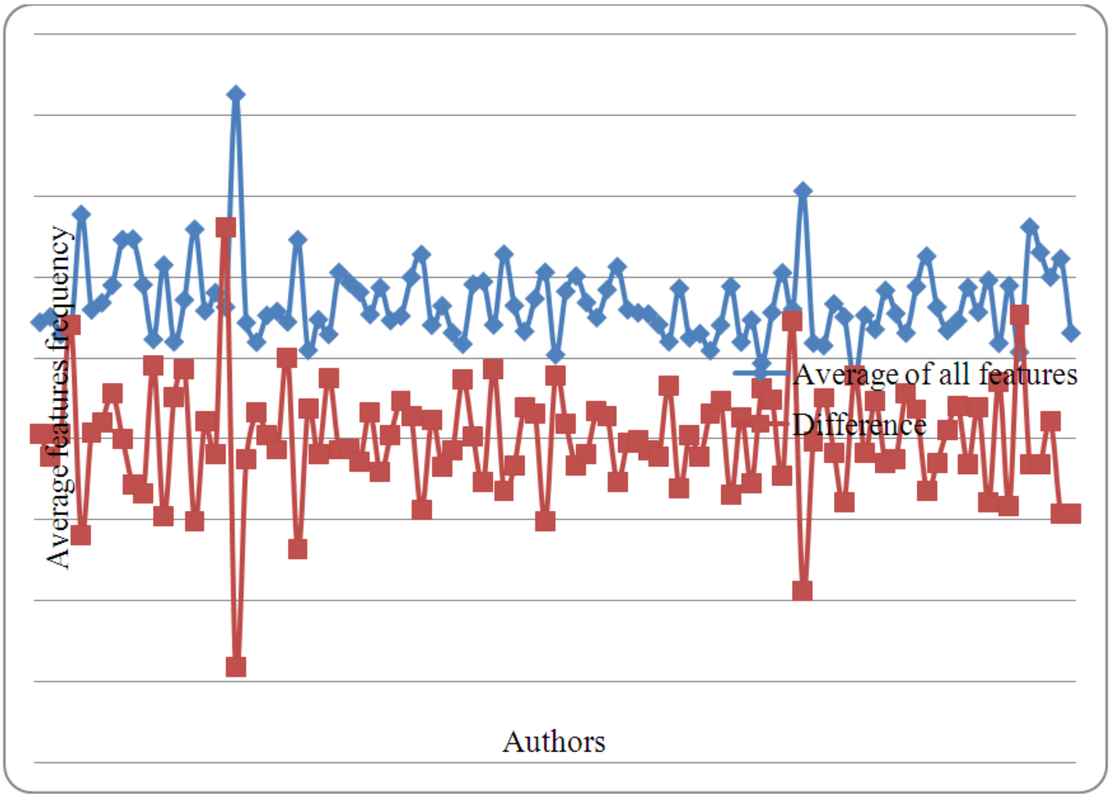

Fig. 14. Average frequency of all features

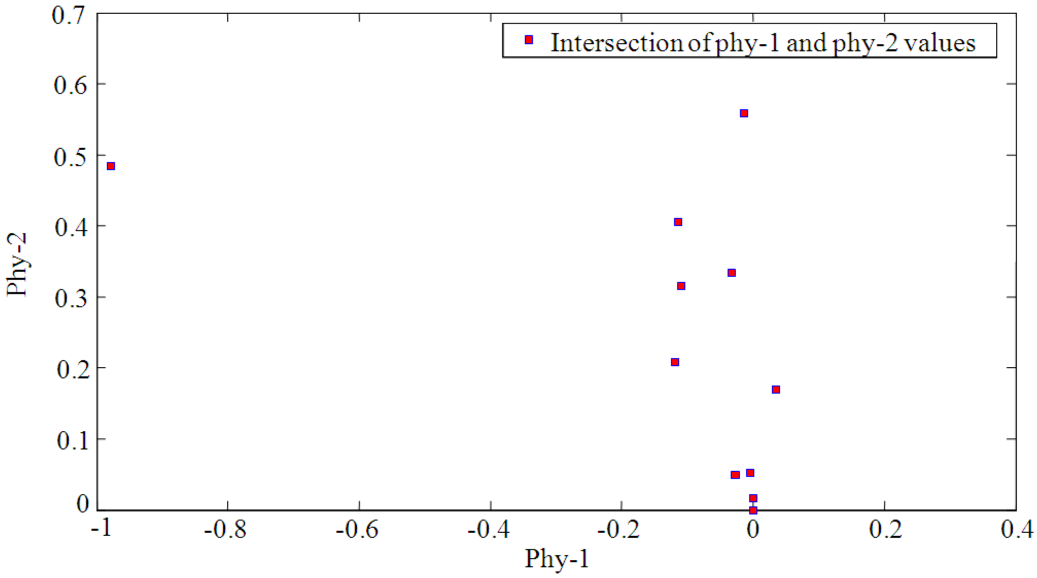

Fig. 15. φ_1 and φ_2 intersections

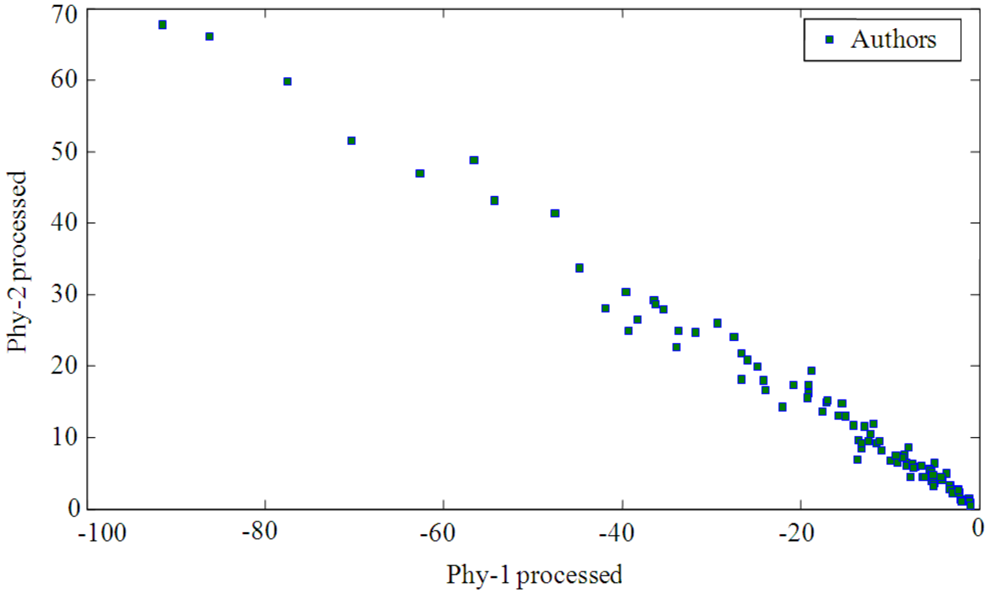

Fig. 16. Projected author patterns

Figure 15 presents the intersections of φ_1 and φ_2 projection vectors. In **Fig. 16**, signatures of 100 authors are projected using φ_1 and φ_2 vectors into 2-dimension.

4. DISCUSSION

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.

5. CONCLUSION

This study presents the email authorship categorization using Fisher's linear discriminant method combined with Radial basis function network. FLD transforms 322 dimensional signature pattern into 2-dimensional pattern. As there is overlapping of few authors (**Fig. 16**), RBF has been used. Advantages of the proposed system is as follows:

- The size of the 322-dimensional signature pattern is reduced to 2-dimension
- The training of 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

6. REFERENCES

David, I.H., 1992. A stylometric analysis of mormon scripture and related texts. J. Royal Stat. Soc. Series A, 155: 91-120. DOI: 10.2307/2982671

Farkhund, I., H. Binsalleeh, B.C.M. Fung and M. Debbabi, 2010. Mining writeprints from anonymous e-mails for forensic investigation. Digital Investigat., 7: 56-64. DOI: 10.1016/j.diin.2010.03.003

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. Digital Investigat., 5: S42-51. DOI: 10.1016/j.diin.2008.05.001

Grieve, J., 2007. Quantitative authorship attribution: An evaluation of techniques. Literary Linguist. Compu., 22: 251-270. DOI: 10.1093/llc/fqm020

Luyckx, K. and W. Daelemans, 2008. Authorship attribution and verification with many authors and limited data. Proceedings of the 22nd International Conference on Computational Linguistics, (CCL' 08), Association for Computational Linguistics Stroudsburg, PA, USA., pp: 513-520.

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

Sambasiva, R.B., S. Ramakrishna, M.S. Rao and S. Purushothaman, 2009. Implementation of radial basis function neural network for image steganalysis. Int. J. Comput. Sci. Security, 2: 12-22.

Stamatatos, 2009. A survey of modern authorship attribution methods. J. Am. Soc. Inform. Sci. Technol., 60: 538-556. DOI: 10.1002/asi.v60:3

Zheng, R., J. Li, Chen, H. and Z. Huang, 2006. A framework for authorship identification of online messages: Writing style features and classification techniques. J. Am. Soc. Inform. Sci. Technol., 57: 378-393. DOI: 10.1002/asi.20316