Detection of Fraudulent Emails by Authorship Extraction

A. Pandian
Department of MCA
SRM University, Chennai, India

Mohamed Abdul Karim
College of Applied Sciences, Sohar,
Ministry of Higher Education, OMAN

ABSTRACT

Fraudulent emails can be detected by extraction of authorship information from the contents of emails. This paper presents information extraction based on unique words from the emails. These unique words will be used as representative features to train Radial Basis function (RBF). Final weights are obtained and subsequently used for testing. The percentage of identification of email authorship depends upon number of RBF centers and the type of functional words used for training RBF. One hundred and fifty authors with over one hundred files from the sent folder of Enron email dataset are considered. A total of 300 unique words of number of characters in each word ranging from three to seven are considered. Training and testing of RBF are done by taking different lengths of words. Our simulation shows the effectiveness of the proposed RBF network for email authorship identification. The accuracy of authorship identification ranges from 95% to 97%.

Keywords: email authorship identification, spam, word frequency, radial basis function

1. INTRODUCTION

As the volumes of emails on the net increases, spam and hoax mails have to be detected. The principal objective of author identification is to classify [Koppel et al, 2002] the emails belonging to an author. This approach is used in forensic for author identification in malicious emails. Certain commercial software such as AntConc, Copy Catch Gold, Lexico3, Signature Stylometric System, T-lab, Yoshikoder, and WordSmith [2]Department of Information Technology
Tools use statistical methods to identify an author. These systems use parameters such as the number of unique words, number of content words used in the list, total number of words in the text or vocabulary items used, vocabulary richness, mean sentence length, mean paragraph length, mean of 2-3 letter words, mean of words starting with vowels, cumulative summation method, and bigrams. The users who intend to utilize the software for their email author identification need to choose the type of statistical analysis options that best identify author of an email and obtain the characteristics that remain constant for large number of emails written by the author. Each author follows a certain style, which is based on functional words. By using these functional words and their frequencies, identification of the author is possible [Madigan et al, 2005].

2. RELATED WORK

By and large, research has focused on different aspects of text. There are two different properties of the texts that are used in classification: the content of the text and the style of the author. Stylometry [Goodman 2007] is the statistical analysis of literary. Style complements traditional literary scholarship since it offers a means of capturing the often elusive character of an author's style [Zheng 2006] by quantifying some of its features. Most stylometry [Pavelec et al. 2007] [Diederich and Chen 2008] studies employ items of language and most of these are lexically based.

The usefulness of function words in Authorship attribution has been examined [Diederich et al. 2003]. Experiments were conducted with support vector machine classifiers in twenty novels and success rates above 90% were obtained. The use of functional words is a valid and good approach in attribution of authorship [Koppel 2006]. A success rate of 65% and 72% has been measured in the study for authorship recognition, which is an implementation of multiple regression and discriminant analysis [Stamatatos et al, 2000]. Concurrently experiments conducted with support vector classifiers [Diederich et al. 2003] detected authors with 60-80% success rates using different parameters.

The effect of word sequences in authorship [Abbasi 2005] attribution has been studied. The researchers aimed to consider both stylistic and topic features of texts. In this work, the documents are identified by the set of word sequences that combine functional and content words. The experiments are conducted on a dataset consisting of poems using naïve Bayes classifier [Peng et al, 2004].

Later authorship studies (Farkhund Iqbal 2010) contain lexical, syntactic, 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.

3. APPROACH OF INFORMATION EXTRACTION

Different types of words are used for filtering and as templates. Words indicating work, action, different categories of prepositions, pronouns, adjectives, adverbs, conjunctions and interjections are listed in Table 1. While analyzing an email for uniqueness, the extracted features are categorized based on the list of words presented. Hence, unnecessary words are eliminated and the number of unique words that represent an email is minimal.

Table 1 Sample Words Used For Filtering

Work (70)	Action (524)	Preposition _1 (94)	Preposition _2 (30)
analyze	Accelerate	Aboard	according to
annotate	Accommodate	About	ahead of
ascertain	Accomplish	Above	as of
attend	Accumulate	Absent	as per
audit	Achieve	Across	as regards
build	Acquire	After	aside from
calculate	Act	Against	because of
consider	Activate	Along	close to
construct	Adapt	Alongside	due to
control	Add	Amid	except for

Table 2 Sample Words Used For Filtering

Preposition _3 (16)	Preposition _4 (9)	Pronoun (77)	Adjectives (395)
as far as	apart from	All	early
as well as	but	Another	abundant
by means of	except	Any	adorable
in accordance with	plus	anybody	adventurous
in addition to	save	Anyone	aggressive
in case of	concerning	anything	agreeable
in front of	considering	Both	alert
in lieu of	regarding	Each	alive
in place of	worth	each other	amused
in point of		Either	ancient

Table 3 Sample Words Used For Filtering

Adverbs (331)	Conjunctions (25)	Interjections (77)
Abnormally	And	Absolutely
absentmindedly	But	Achoo
Accidentally	For	Ack
Acidly	Nor	Agreed
Actually	Or	Aha
Adventurously	So	Ahem
Afterwards	Yet	Ahh
Almost	after	Ahoy
Always	although	Alack
Angrily	as	Alas

Words indicating work are used to analyze how an author writes an email. These avoid misinterpretation and elicit clarity in the email. The number of work-words will point out performance task requirements in a neat, unambiguous manner. These work-words decipher exactly what an author has in his mind. Action words indicate some actions being expressed in the email. Prepositions, adjectives, adverbs, conjunctions and interjections have their standard part-of-speech.

The total number of words used as basic dictionary is 1648 (work, action, prepositions, adjectives, adverbs, conjunctions and interjections). The numbers mentioned in the parenthesis are the total in each category whereas; only few words are shown in the tables for brevity.

In the training phase, a variety of emails are input to the system. Tokenization is the process of extracting all the words in the email as tokens. These words are matched with the filter words given in Table 1. Subsequently, the word frequencies are calculated for each category. The author matrix is built with authors along the columns and their corresponding word frequencies. These entries are used as labeled training patterns. In the testing phase, new emails are input. Authors are identified based on the weights learnt. A schematic diagram for implementation of the proposed work is presented in Figure 1.

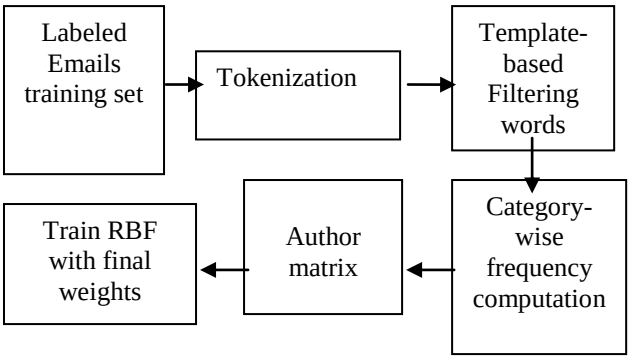

Fig.1 (a) Training phase

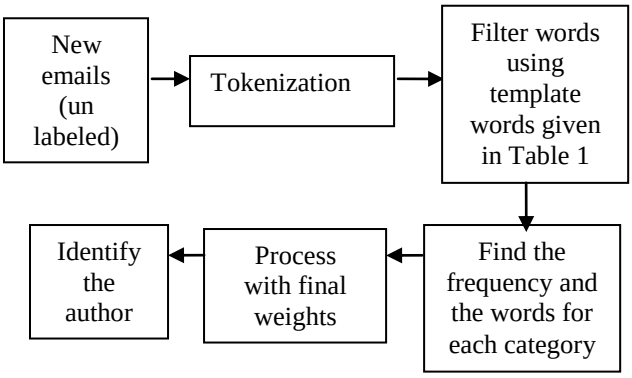

Fig.1 (b) Testing phase

The concept of distance measure is used to associate the input and output pattern values. Radial Basis Function (RBF) is suitable for producing approximations to an unknown function f from a set of input data abscissa. The approximation is produced by passing an input point through a set of basis functions, each of which contains one of the RBF centers.

An exponential function is used as an activation function for the input data. Distance between input data and set of centers chosen from the input data are found and passed through an exponential activation function. A bias value of f is used along with the data. These data are further processed to get a set of final weights between radial basis function and the target value.

The topology of RBF network is twelve nodes in the input layer, four nodes in the hidden layer and one node in the output layer. The difference in input data and a center is passed through exp(-x) and is called the radial basis function. A rectangular matrix is further obtained for which inverse is found. The resultant value is processed with the entire inputs and target values to obtain final weights.

The procedure is shown in the form of flowchart in Figure 2. The input pattern is read with the columns of the author matrix being used as training patterns. The number of patterns is equal to number of authors. Then, one hundred training patterns are created as centers. The distance between these centers and training patterns is

calculated to create RBF. The resultant values are passed through activation function, exp(-x) to produce outputs of RBF nodes in the hidden layer of the network. The number of training patterns and the number of centers will produce a rectangular matrix. This is converted into square matrix by singular value decomposition (SVD). Inverse of this matrix is found and processed with labeling to get final weights.

Details of the Figure 2 is given below:

Read input pattern: The columns of the author matrix are used as training patterns. The number of patterns is equal to number of authors.
Create center: One hundred training patterns are used as centers.
Create RBF: Calculate distance between training patterns and one hundred centers. The resultant values are passed through activation function, exp(-x) to produce outputs of RBF nodes in the hidden layer of the network.
The number of training patterns and the number of centers will produce a rectangular matrix. This is converted into square matrix and inverse of the same is found and processed with labeling to get final weights.

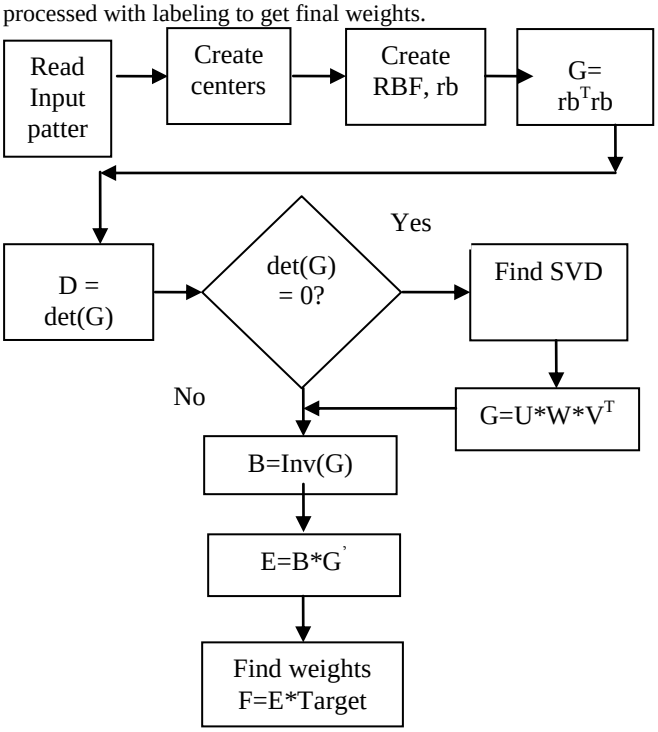

Fig 2 Radial basis function flow chart

4. EXPERIMENTAL PROCEDURE

Enron email dataset has been used for evaluating the efficiency of RBF in email authorship identification. This email dataset was made public by the Federal Energy Regulatory Commission during its investigation [Klimt and Yang, 2004]. It contains all types of emails: personal and official. This contains around 5,17,431 emails from 151 users. Each mail in the folders contains the senders and the receiver email addresses, date and time, subject, body, text and some other email specific technical details. It is available in the form of MySql database. The Enron database contains four tables. The first table contains information of each of the 151 employees. The second one contains the information of the email message, the sender, subject, text and other information. The third contains the recipient's information. The fourth table contains information whether it was sent as a forward or reply. Table 2 presents names of few folders under each author. We considered 146 authors for this study.

There are fifteen unique words that are identified in all the emails under consideration by using the filtering words given in Table 1. The unique words are our, out, plan, please, that, to, we, what, when, which, with, you, your, yours, and zip.

Person	Sent _mail	All documents	contacts	Deleted _items	Discussion _threads	inbox	Notes inbox	sent	sent_ items
allen-p	602	628	2	361	412	66	48	562	345
arnold-j	814	1047	X	723	401	142	84	816	723
arora-h	X	65	X	197	57	79	X	9	68
badeer-r	52	299	2	13	277	3	115	X	7
bailey-s	X	16	X	434	X	4	10	X	14
bass-e	1409	2037	X	415	1386	310	601	1363	258
Baughman-d	X	389	4	431	384	383	X		96
beck-s	1093	3137	7	309	2630	751	190	1099	482
benson-r	X	84	X	203	77	274	75	7	9
blair-l	39	2	X	662	X	291	X	X	929
X represents no information									

Table 2 Details Of Enron Folder

5. CHARACTERIZATION MEASURES

Based on certain grammatical rules followed and the words beginning with vowels in the emails by an author, characterization and feature extraction is done for training RBF. Figure 3 plots the authors' id along x-axis and number of words with vowels at the beginning of words along the y-axis. Each stem is the average number of words beginning with vowels considering all the emails by an author. Figure 4 plots the number of "work" words used by each author. Figure 5 shows the number of "action" words. Figures 6 to 11 present the numbers of parts-of-speech – prepositions, pronouns, adjectives, adverbs, conjunctions and interjections – in the same order.

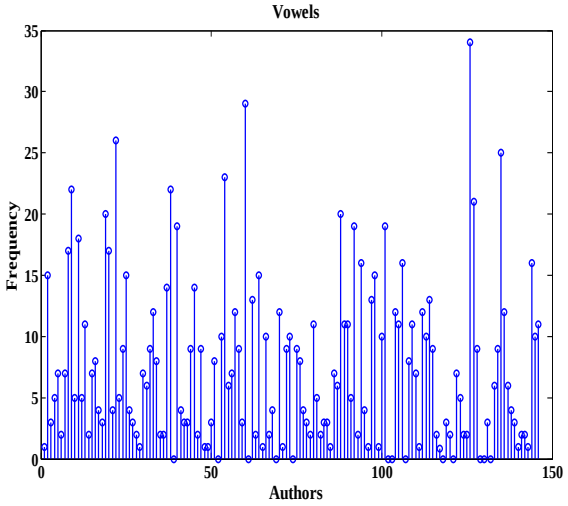

Fig.3 Number of words with vowel in the beginning of words

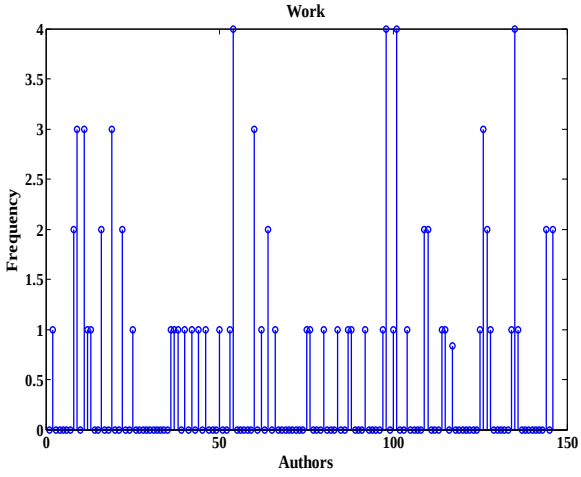

Fig4. Work words for each author

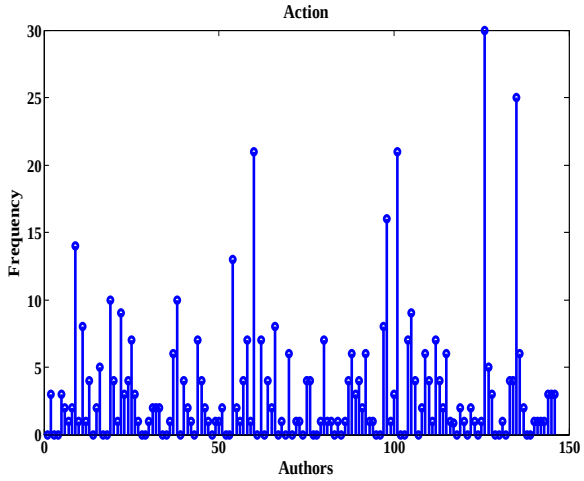

Fig.5 Action words for each author

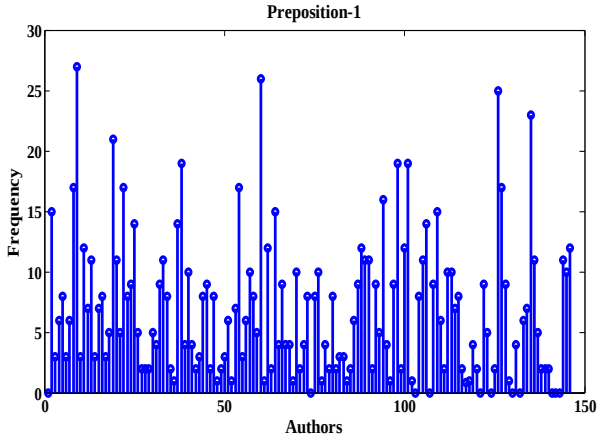

Fig.6 Preposition 1 for each author

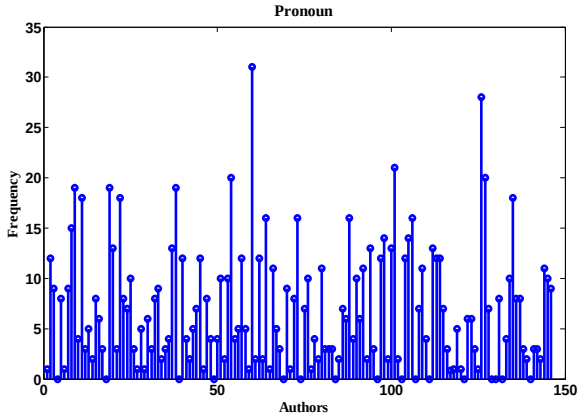

Fig.7 Pronoun for each author

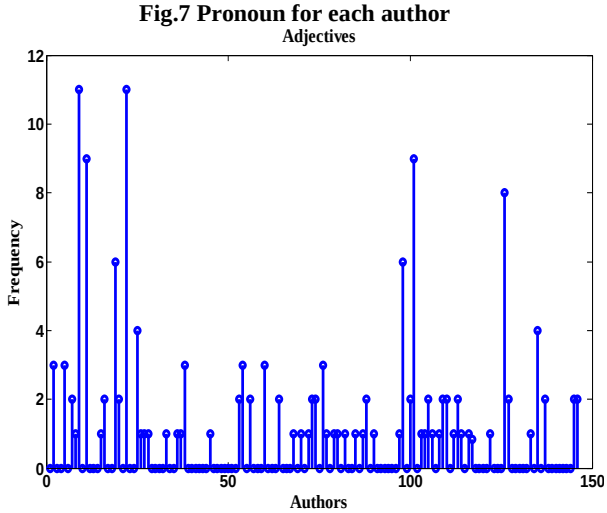

Fig.8 Adjectives for each author

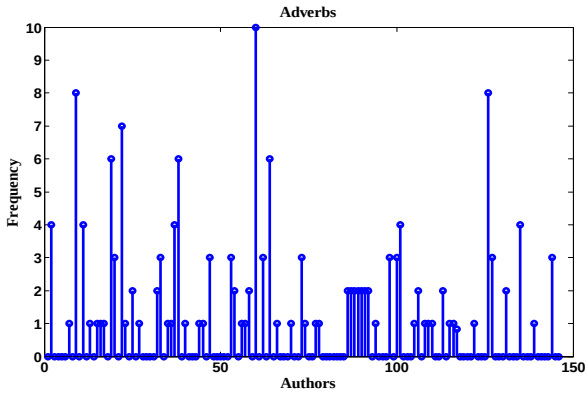

Fig.9 Adverbs for each author

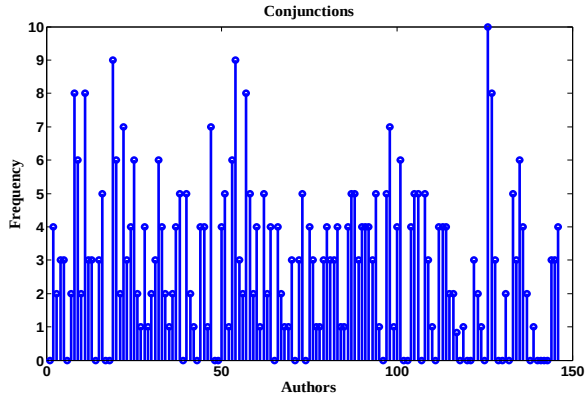

Fig.10 Conjunctions for each author

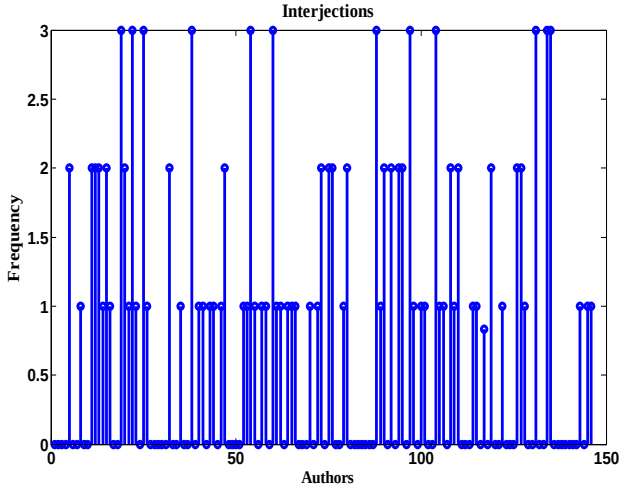

Fig.11 Interjections for each author

We use the following algorithm for email identification by neural network training and testing:
For each author, find the number of words and their frequency of occurrences in every email by using the filtering words listed in Table 1.

1. Construct a matrix with rows indicating total number of unique words considering all emails of the authors. Each column is for an author.
2. Lexicon is built from the email corpus.
3. A column of the zero matrix is filled based on the availability of the words in a document with their frequencies. Each column will be treated as a pattern for training. Labeling is done for each pattern.
4. Train the RBF network with patterns considered for training. A final weight matrix is obtained which is further used to test the incoming mails that belong to existing authors else, the mail can belong to some other person other than these existing authors considered in this experiment.

6. RESULTS AND DISCUSSIONS

Figure 12 plots the performance of RBF in training the patterns. When the number of centers used is less than half of the total number of input patterns, the performance of author identification is minimal. As the number of centers increase, the author identification improves. The legend shows the number of centers varied. Figure 13 presents the performance of the RBF in terms of emails versus author identification. In this plot, output obtained from RBF overlaps target outputs. With 146 centers, the RBF identifies maximum number of authors.

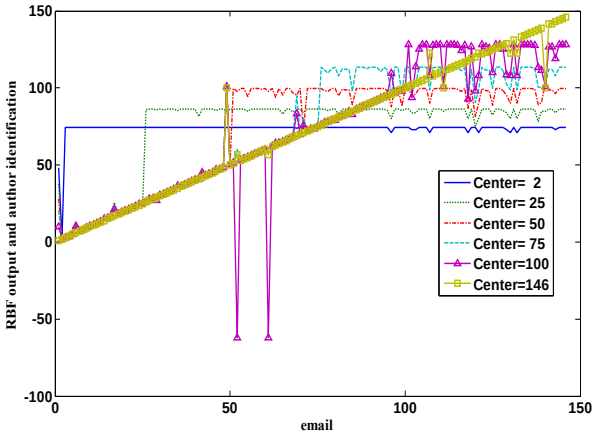

Fig.12 Performance of RBF center selection

We presented a novel method of identifying email authorship using RBF patterns of data. The training data has been collected by averaging the frequencies of words used by each person and fixing a target value for the person. Testing pattern has been created by modifying the existing contents of an email. A new word has been considered while testing. If the new word does not fit into the patterns used for training, that word is excluded in testing. As we are unaware to which author the email belongs, all the training patterns are treated as test patterns after adding the frequencies of the new mail. Since 146 authors are considered, 146 outputs are obtained after testing

Receiver Operating Characteristics (ROC) of the authorship identification reveals the following analysis:

Has the author of a document been identified correctly? This is termed true positive.

Has the author been wrongly classified, though the document does not belong to the same? This is termed false positive.

Has a new document (not from the training corpus) been classified appropriately (not to the authors in this experiment). This is termed true negative

Has a new document (not from the training corpus) been wrongly identified as belonging to one of the authors in the experiment? This is termed false negative.

Sensitivity = Number of True Positives / (Number of True Positives + Number of False Negatives)

Specificity = Number of True Negatives / (Number of True Negatives + Number of False Positives)

False Positive Rate = 1 – Specificity

Table 3 presents the confusion matrix values and the ROC values. The author emails have been considered that belong to the training corpus and that do not belong to training corpus. All the emails that belong to (sent / sent_mail) folders are used for training. The emails of the remaining folders of all authors have been considered for testing. The performance of RBF has been calculated using confusion matrix. The plot (Figure 14) indicates that the proposed RBF system suits the author identification from the given emails. This is inferred from the points obtained above the diagonal of the ROC curve.

Table 3 Confusion Matrix For Receiver Operating Characteristics

Instances	True Positive	False Negative	Sensitivity	False Positive	True Negative	Specificity	True Positive Rate
1	80	20	0.80	10	40	0.80	0.20
2	82	18	0.82	8	42	0.84	0.16
3	90	10	0.90	5	45	0.90	0.10
4	85	15	0.85	7	43	0.86	0.14
5	92	8	0.92	8	42	0.84	0.16
Sensitivity=True Positive Rate=True Positive/Total words True Positive Rate=1-Specificity							

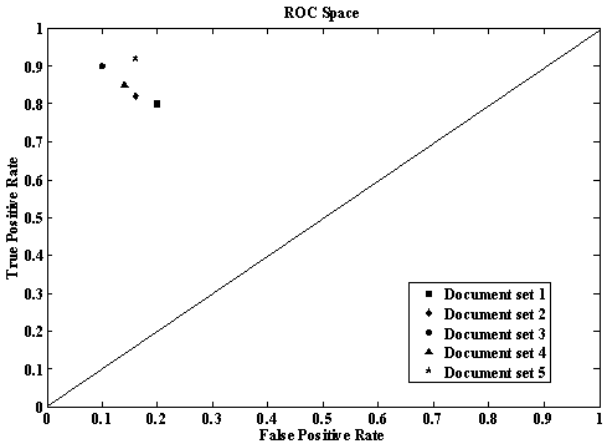

Fig.17 Receiver Operating Characteristics

7. CONCLUSION

The RBF approach has been presented for detecting authors of emails, thereby reducing fraudulent emails. Different RBF centers and their effectiveness in author identification are presented. The receiver operating characteristics curve has shown that the proposed RBF network performance is acceptable. In future, the huge amount of words can be meaningfully filtered such that they are more specific to an author. This can further enhance the detection of unsolicited fraud emails. Therefore, the number of spam emails will decrease.

8. REFERENCES

[1] Abbasi A. And Chen H, "Applying Authorship Analysis to Extremist-Group Web Forum Messages" IEEE INTELLIGENT SYSTEMS, pp. 67–75, 2005.

[2] David Madigan, Alexander Genkin, David Lewis, Shlomo Argamon, Dmitriy Fradkin, and Li Ye, "Author Identification on the Large Scale", *Proc. of The Meeting Of The Classification Society of North America,2005*.

[3] Diederich, J., and Chen, H. 2008. Writeprints, "A stylometric approach to identity-level identification and similarity detection", ACM Transactions on Information Systems (26:2),pp. 7.

[4] Diederich, J., Kindermann, J., Leopold, E. and Paass, G. (2003), "Authorship Attribution with Support Vector Machines", Applied Intelligence 19(1), pp. 109-123.

[5] Goodman R., Hahn M., Marella M., Ojar C., And Westcott S., "The Use Of Stylometry For Email Author Identification: A Feasibility Study", Proc. Student/Faculty Research Day, CSIS, Pace University, White Plains, NY, pp.1-7, May 2007.

[6] Klimt B. & Yang Y., (2004). The Enron corpus: A new dataset for email classification research, In Proceedings of ECML'04, 15th European Conference on Machine Learning, pages 217-226,(2004

[7] Koppel, M., Schler, J., Argamon, S. and Messeri, E., "Authorship Attribution with Thousands of Candidate Authors", in Proc. 29th ACM SIGIR Conference on Research & Development on Information Retrieval, 2006.

[8] Moshe Koppel, Shlomo Argamon, And Anat Rachel Shimoni, "Automatically Categorizing Written Texts By Author Gender", *Literary And Linguistic Computation*. 17(4):pp.401-412, 2002.

[9] Pavelec, D., Justino, E., And Oliveira, L. S., "Author Identification Using Stylometric Features", Inteligencia Artificial (11:36), pp. 59-65, 2007.

[10] Peng, F., Schuurmans, D., ,Wang, S., "Augumenting Naive Bayes Text Classifier With Statistical Language Models , Information Retrieval", 7 (3-4), Pp. 317 – 345, 2004.

[11] Stamatatos, E., Fakotakis, N., & Kokkinakis, G., (2000). Automatic text categorization in terms of genre and author. Computational Linguistics, 26(4), 471-495.

[12] Zheng R., Li J., Chen H., Huang Z., "A Framework For Authorship Identification Of Online Messages: Writing-Style Features And Classification Techniques", Journal of the American Society for Information Science and Technology, 57(3):378–93.

[13] Farkhund Iqbal , Hamad Binsalleeh, Benjamin C. M. Fung, Mourad Debbabi , " Mining writeprints from anonymous e-mails for forensic investigation, Digital Investigation,1 – 9 (2010) .