International Journal of Engineering & Technology, 7 (4.19) (2018) 8-12

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Comparative Studies of Author Identification algorithms for Telugu Classical Poems

A. Pandian[1], K.Manikandan[2], V.Ramalingam[3], Pavuluri Sai Krishna[4], V. Jeevan Reddy[5]

[1]Associate Professor, [2,3]Asst Prof (SG), [4,5]B.Tech Student
[1,3,4,5] Dept. of CSE, SRMIST, Chennai, [2]Dept. of IT, SRMIST, Chennai

Abstract

Creator recognizable proof is the assignment of distinguishing the creator of a given test from an arrangement of suspects. The fundamental worry of this assignment is to characterize a fitting portrayal of test that catches the written work styles of writers. In this task, weka based machine learning apparatuses are utilized for ID of creator for include extraction of reports spoke to utilizing variable size character n-grams. We composed our own java program to extricate the highlights like number of words, sentences and so on. From, the ballad which thusly sustained as contribution to weka device for the recognizable proof of creator then in the wake of testing the contribution with all the calculation all the exactness rates are noted down to see which calculation is given us the best precision rate. Presently to discover the creator name for a mysterious sonnet the lyric highlights are extricated utilizing the java code and the yield is taken in the java record given to the weka instrument and tried with the calculations and after that the creator name is given to the unknown ballads.

Keywords: *weka, portrayal, stylometry, ballads, creator.*

1. Introduction

Author distinguishing proof is the errand of recognizing who composed a given bit of content from a given arrangement of applicant creators (suspects). From machine learning point of view, it can be seen as multiclass single- content order assignment where creator speaks to a class (mark) of a given content. The investigation of stylometry and origin backpedals to the nineteenth century, with Mendenhall leading the pack by describing the style of various creators through the recurrence dissemination of expressions of different lengths. Amid the principal half of the twentieth century, numerous factual examinations were taken after presenting measures for composing styles including Zipfs circulation and Yules K measure.

Present day creation ID began by Mosteller and Wallace take a shot at the federalist papers, where they connected Bayesian measurable examination on the frequencies of a little arrangement of capacity words (e.g "and", "to", "the"), as expressive highlights of test. In the writing numerous highlights have been proposed to catch complex highlights including vocabulary abundance measures, linguistic highlights, work words frequencies and character n-gram frequencies. Profound learning has been effectively connected to different common dialect handling assignments creating execution comes about beating already best in class system. For instance, connected profound learning on the area adaption of feeling investigation by utilizing abnormal state highlight portrayal extricated utilizing profound neural systems and beat the condition of craftsmanship strategies on the order undertaking. Additionally, profound with the fast improvement of data, more correspondence and capacity of records is performed carefully.

An extraordinary extent of business documentation and correspondence, in any case, still takes puts in physical shape and the fax machine stays key device of correspondence around the world. Along these lines, optical character acknowledgment (OCR) is winding up increasingly vital. In any case, all the current takes a shot at OCR make a vital understood presumption that the content and dialect of the archive to be handled is known.

Human intervention in identifying the script and language of document in dealing with massive images cannot satisfy the requirement of speed and automation. Hence, script identification, by way of the front processing technology of OCR system, is essential and significant.

Writer ID can be viewed as a characterization issue of texts:"Given an arrangement of records composed by a same writer, set can be substantial or made out of just a single component, we need to choose if another archive has been composed by an indistinguishable writer from the others". We need to take care of an issue of order having as reaction a twofold esteem ("yes" or "not") r a likelihood to have a place with the arrangement of known reports. Be that as it may, one of the specificities of this issue is that lone components having a place with one of the two conceivable classes are given: the records having a similar creator, however the below average are not expressly depicted. Additionally, now and then the quantity of positive illustrations is diminished to just a single archive and, the assignment turns out to be significantly more troublesome. To moderate the nonattendance of negative illustrations, one can attempt to deliver some of them. Thusly is investigated by various creator among which Seidman who fabricates a class of impostors arbitrarily picked on the web based on ten more successive words in the accessible archives.

Different creators, as Zhang et al and Halvani change this issue of arrangement with two classes into issue with a few classes, either

by including outer classes or by separating the underlying classes into a few. These same creators increment the span of the class containing the know archives when this last one is lessened to just a single. Along these lines, these methodologies permits to change the issue into a traditional from of arrangement, however amid the development of the arrangement of negative cases there is a hazard to take a few reports altogether different from the known records. It us broadly recognized that individuals around the globe are progressively utilizing the PC advances and PC intervened correspondences to interface with each other. The web's consistent availability and easy to use stage have changed the sharing of data and correspondence, encouraging a worldwide web of virtual groups.

2. Material and Methods

Finding the creators for anonymous poems in Telugu find the opportunity to be especially troublesome as there is no framework to remember them curiously. By separating these highlights vital to Telugu compositions and by utilizing reasonable estimations, essayists for these dark works can be seen. Gathering is done by utilizing content giving strategy. Content taking care of is the framework for getting top notch data from substance that solidifies genuine cases from the substance.

Data set is nothing but a collection of related sets of information that is composed of separate elements but can be manipulated as signals unit by the computer . In this project the data set is the most important part because it is the source for training the machine and finding the author for unauthorised poem. We have collected poems of 8 different authors and for each authors we have collected more than 101 poems.In this dataset (418) poems used for training the machine and (418) poems for testing the machine. The data sets we have used in the project are as follows

S.no	Author	No. of poems collected
1	Pakki Venkata Narasimha	103
2	Venkaya kavi	104
3	Swami Parmanandha	106
4	Buchana	110
5	Dasi Sree Ramulu	100
6	Kancharla Gopana	103
7	Sadanandha Yogi	102
8	Bharthru Hari	108

By extracting lexical, syntactic and semantic elements as clarified in the classification process is performed. The rundown of features that are considered is shown in table-1.

These highlights are extricated from the informational index and used for performing grouping. These highlights describe the stylometry of the maker. Stylometry is the utilization of examination of created styles from physically composed articles that can be utilized as a major aspect of origin recognizable proof. Stylometry consolidates extraction of lexical, syntactic, measurable highlights that ate separated from the dataset. By utilizing J48 calculation, an exactness of 88.69% was accomplished.

The J48 algorithm consists of two parameters, confidence factor and minimum number of objects. These two factors have to be varied in order to obtain some difference in the accuracy. The confidence factor have to be varied from 0.1 to 1.0 while the minimum number of objects have to varied from 1 to 23 actual number of features considered in the Data set. After performing the tweaks, the final accuracy achieved is 88.6% confidence factor being 0.2 and minimum number of objects being 4, the peak accuracy was achieved.

3. Feature Extraction

Feature extraction handle assembles an arrangement of derived qualities from the underlying arrangement of information that is planned to human translation. Dataset cannot be specifically utilized as a part of the tool to perform arrangement. Just the features that are extricated from the dataset from the dataset can be utilized to assemble the classifier. This classifier that is built is then used to perform the classification process on the dataset in hand.

Three types of features, lexical, syntactic and semantic are extracted. Lexical features include categories such as noun, verb, adjective, and pronoun. Syntactic features include noun phrase, verb phrase and prepositional phrase. Semantic features are those that include a set of features that intensifies the meaning of a word.

In addition to these features, statistical features are also extracted from the dataset. Statistical features account to a major part of the classifier accuracy. The classifier accuracy has increased from 86% to 90% by including statistical features to the features set and performing some tweaks in the algorithm used. Statistical features include minimum, maximum, sum and mean.

The attributes recorded in table- I are extracted from the dataset. The dataset is initially changed over into Unicode format so it can perused in Microsoft excel. Computers cannot comprehend Telugu characters. They bargain just with numbers in their memory. Unicode gives an encoding framework that covers all the regional languages and gives an approach to computers to comprehend them.

The extraction procedure is done by utilizing sql commands, which can extricate the predetermined features consequently. Sqlite browser is utilized to make a database with every one of the poems and components. The extracted features are in numeric format.

These numeric features that are extracted are all used in the classification process as all of these features play a vital role in improving the classifier accuracy to a great extent.

Attribute Set

Attributes type	Attributes
	• Count Word
	• Sentence count
	• Character count
	• Paragraph count
	• White space count
	• Occurrence of achulu, halulu, gunithalu, vothulu
Statistical features	• Mean of Word Count, Median of Word Count, Mode of Word Count
	• Ratio of Count Word TowardsA
	• Ratio of Sentence Count TowardsA
	• Ratio of Character Count TowardsA
	• Ratio of Paragraph Count TowardsA
	• Ratio of White Space Count TowardsA
	• Ratio of Count Towards Lines
	• Ratio of Sentence Count TowardsB
Syntactic features	• Ratio of Character Count TowardsB
	• Ratio of Paragraph Count TowardsB
	• Ratio of White Space Count TowardsB

We have added a special to the feature set i.e frequency of words . The frequency of words feature has been the stand out feature for many algorithms. Because of this feature In the dataset the accuracy rates for all the algorithms have been raised.

Table1: Best attributes ranges from (1-9) are selected from j48 decision tree algorithm range and achieved accuracy as below.

Features	Accuracy
Mean of word count	61.79%

Ratio of sentence count towardsA	83.62%
Ratio of count word towards lines	84.9%
Ratio of count towardA	88.304%
Count word	87.13%
Hallulu count	85.57%
Sentence count	85.57%
White space Count	87.13%
Guninthalu count	87.91%

Table2: Varying features in the dataset over the best chosen attributes accuracy as follows.

Minimum Number of Objects	Accuracy
1	88.10%
2	87.914%
3	88.10 %
4	88.69%
5	87.9%
6	86.35%
7	86.15%
8	86.74%
9	86.74%
10	86.35%
11	85.77%
12	85.18%
13	85.18%
14	84.99%
15	84.79%
16	84.99%
17	84.99%
18	84.79%
19	84.79%
20	84.99%
21	84.40%
22	84.40%
23	84.79%

Table3: Highest Minimum number of objects Vs changing confidence factor from 0.1-1.0.

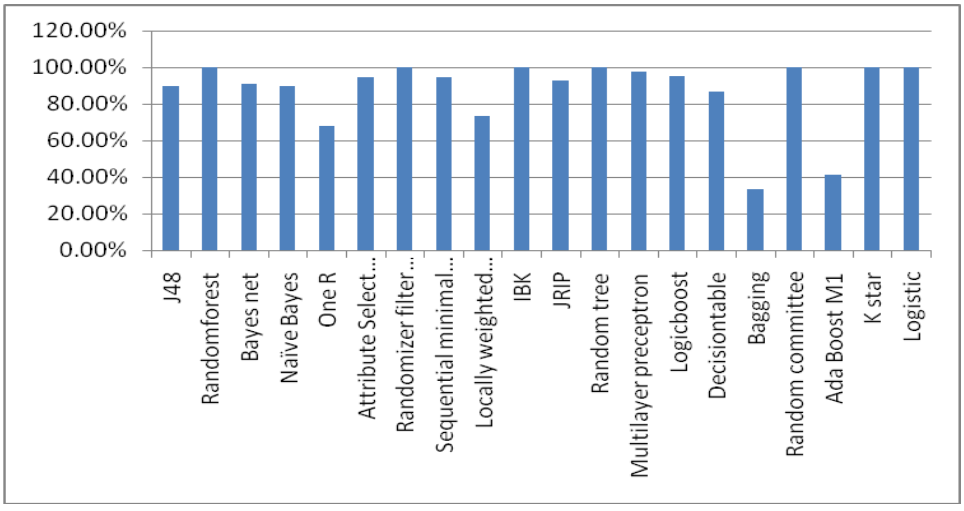
Confidence Factor with constant MNO(minimum no of object)	Accuracy
0.1	88.30%
0.2	88.69%
0.3	88.30%
0.4	88.49%
0.5	88.49%
0.6	88.69%
0.7	88.69%
0.8	88.69%
0.9	88.69%
1.0	88.69%

After finding confidence factor over greatest minimum no of objects we have achieved outstanding accuracy 88.69% using J48 algorithm.

3.2. Algorithms Accuracy

S.No	Algorithms	Percentage
1	J48	90.05%
2	Randomforest	100%
3	Bayes net	91.228%
4	Naïve Bayes	90.044%
5	One R	67.836%
6	Attribute Select Classifier	94.34%
7	Randomizer filter classifier	100%
8	Sequential minimal Optimiation	94.73%
9	Locally weighted learning	73.294%
10	IBK	100%
11	JRIP	92.78%
12	Random tree	100%
13	Multilayer preceptron	97.85%
14	Logicboost	95.51%
15	Decisiontable	86.54%
16	Bagging	33.1384%
17	Random committee	100%
18	Ada Boost M1	41.33%
19	K star	100%
20	Logistic	100%

3.3. Achieved Accuracy Graph



4. Results and Discussions

4.1. Training Set

The outcome of the comparison of twenty related algorithms to their corresponding accuracies are listed in Table .These are found using weka explorer by training the data set here we got accuracy as follows.

The Random Forest algorithm which has given its best accuracy on certain datasets has given an peak accuracy of 100% on the dataset at hand. The Naïve Bayes algorithm has also performed well on various other datasets while on the dataset at hand it has given a accuracy of 90.04%. The KStar algorithm has produced an accuracy of about 100% while OneR algorithm has performed to produce an accuracy of 67.83% and SMO algorithm producing 94.73%. J48 algorithm has produced an outstanding 90.05% on the

dataset at hand. The Multilayer Perceptron algorithm which is considered to perform well on almost all datasets has given an accuracy of 97.85%.

The LWL and Logit Boost algorithms have given a similar accuracy of 73.29% respectively, while the Random Tree algorithm and Logistic has produced an accuracy of 100% on the dataset. The Randomizable Filter Classification algorithm and Random Committee algorithm produced the similar accuracy of 100% respectively. The IBK algorithm has produced an accuracy of 100% whereas the JRip algorithm has produced an accuracy of 92.78%. The OneRand AdaBoost M1 have all produced the least accuracy of 67.83% and 41.33% respectively. By adding frequency of words accuracy gain in each algorithm .for training the dataset we use every features(23) in the dataset after we got accuracy as mentioned in the graph. In that we will choose the peak accuracy got for our datset and the model is saved for testing the dataset.

4.2. Testing

4.2.1. Newdata.arff

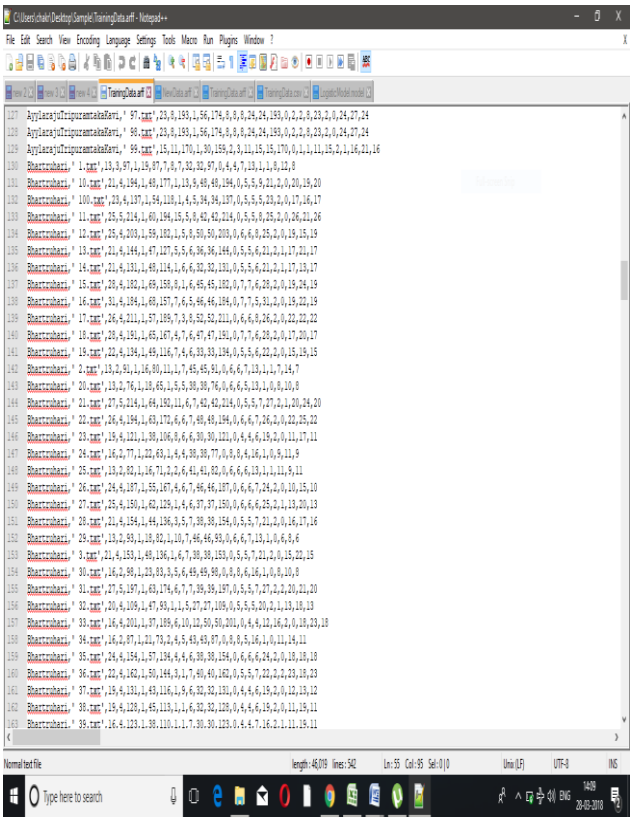
```
@attribute ' whitespaceCount' numeric
@attribute ' meanOfWordCount' numeric
@attribute ' medinOfWordCount' numeric
@attribute ' modeOfWordCount' numeric
@attribute ' ratioOfCountWordTowardsA' numeric
@attribute ' ratioOfSentenceCountTowardsA' numeric
@attribute ' ratioOfCharacterCountTowardsLines' numeric
@attribute ' ratioOfParagraphCountTowardsA' numeric
@attribute ' ratioOfWhitespaceCountTowardsA' numeric
@attribute ' ratioOfCountWordTowardsLines' numeric
@attribute ' ratioOfSentenceCountTowardsB' numeric
@attribute ' ratioOfCharacterCountTowardsB' numeric
@attribute ' ratioOfParagraphCountTowardsB' numeric
@attribute ' ratioOfWhitespaceCountTowardsB' numeric
@attribute ' achuluCount' numeric
@attribute ' halluluCount' numeric
@attribute ' guninthaluCount' numeric
@attribute ' vothkhuluCount' numeric

@data
?,' 1.txt',21,13,196,1,44,179,4,11,9,15,15,196,0,1,1,9,21,2,0,19,18,19
?,' 10.txt',28,8,193,1,65,169,5,4,6,24,24,193,0,3,3,6,28,2,0,17,19,17
?,' 100.txt',22,8,196,1,53,178,9,7,8,24,24,196,0,2,2,8,22,2,0,15,15,15
?,' 1.txt',13,3,97,1,19,87,7,8,7,32,32,97,0,4,4,7,13,1,1,8,12,8
?,' 10.txt',21,4,194,1,48,177,1,13,9,48,48,194,0,5,5,9,21,2,0,20,19,20
?,' 100.txt',23,4,137,1,54,118,1,4,5,34,34,137,0,5,5,5,23,2,0,17,16,17
?,' 1.txt',15,15,169,1,31,157,20,11,11,11,11,169,0,1,1,11,15,2,0,18,22,18
?,' 10.txt',23,5,196,1,54,176,2,2,8,39,39,196,0,4,4,8,23,2,1,16,16,16
?,' 100.txt',22,5,201,1,46,182,9,9,9,40,40,201,0,4,4,9,22,2,1,20,15,20
?,' 1.txt',18,5,103,1,44,89,8,2,5,20,20,103,0,3,3,5,18,2,0,13,13,13
?,' 10.txt',13,5,94,1,27,85,5,7,7,18,18,94,0,2,2,7,13,2,0,13,11,13
?,' 98.txt',13,4,113,1,31,104,15,5,8,28,28,113,0,3,3,8,13,2,0,12,13,12
?,' 99.txt',13,4,125,1,32,116,5,10,9,31,31,125,0,3,3,9,13,2,0,15,18,15
```

For testing the dataset none of the features is selected for classification because it is already trained.for testifying the dataset it should be choosed as supplied test set and have to set test instances by opening file in newdata Attribute-Relational File Format (ARFF) is for testing purpose each author some poems are taken by replacing question mark (?) in place of author name as shown in above. And next thing is to open classifier evaluation options

In that choose output predictions as plain text out will be saved and unmark except preserve order for % Split and load the saved model in result list and next choose the field as author and right click on result list we will get options click on the Re-evaluate model on current test set we will get author names as kept in new data as question mark.Hence by using greatest accuracy one in testing author identification is done as displayed below with after evaluation.

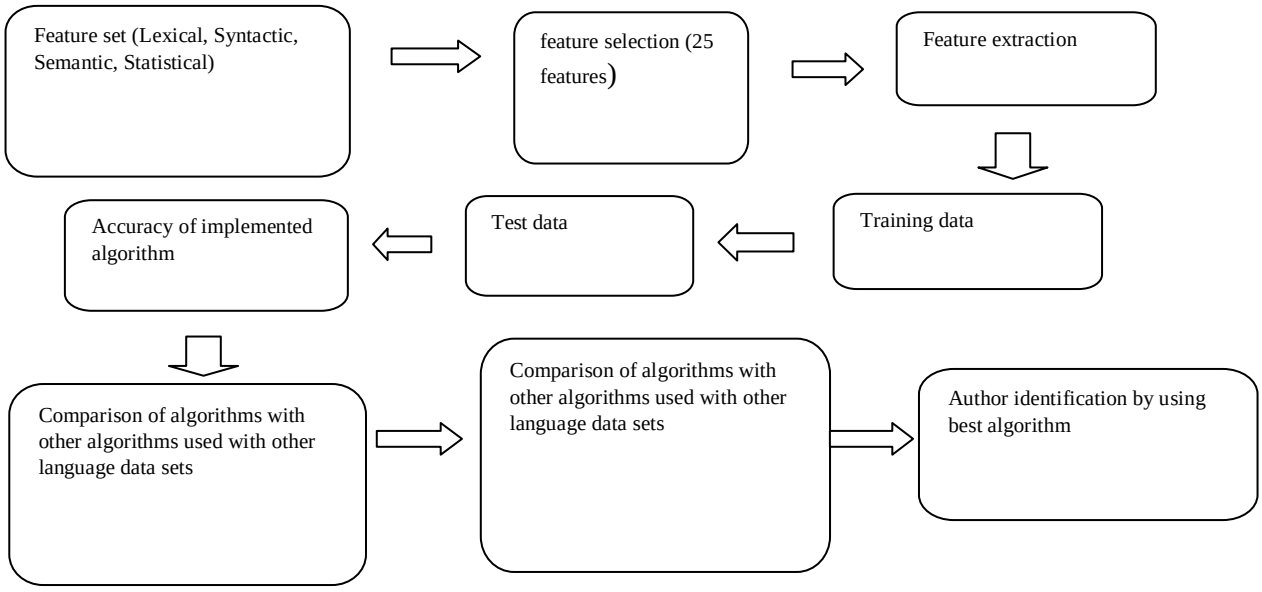
4.2.2. Identification of Author Using Our Dataset



4.3. Architecture Diagram

5. Conclusion

In our work we examined twenty algorithms for classification, the C4.5 algorithm has accomplished satisfactory and has given an maximum peak accuracy of 90.66% on the dataset. Other algorithms like Multilayer Perceptron and IBK have also provided a decent accuracy ranging from 97.85%-100%.Algorithms like Bagging and AdaBoost M1 have given the least accuracy of 33.18% and 41.336% respectively. Out of the 20 algorithms used for comparison, the J48 algorithm has performed well with an accuracy of 90.66%. By adding frequency of words in our dataset we got greatest accuracy compare to previous works. And it is tested with weka for finding an author successfully identified. For future works it is easy extension for finding an author for all Indian languages.



References

[1] Dr.Pandian.A, Ramalingam.V.V and R.P.Vishnu Preet 2016," Author identification for Tamil Classical Poem (Mukkoodar pallu) using Bayes Net Algorithm", Indian Journal of Science and Technology,Vol9(47),DOI:10.17485/ijst/2016/v9i47/107910,December 2016.

[2] Farkhund Iqbal, Hamad Binsalleeh, Benjamin C.M. Fung, Mourad Debbabi, 2015," E-mail authorship attribution using customized associative classification",Digital investigation (Elsevier), Vol.7,pp.56-64

[3] "A Computational Framework for Tamil Document Classification using Random Kitchen Sink", Sanjanasri J.P and Anand Kumar M, IEEE, 2015, International Conference on Advances in Computing, Communications and Informatics (ICACCI)

[4] "An Evaluation of Authorship Attribution Using Random Forests", Mahmoud Khonji, Youssef Iraqi, Andrew Jones, IEEE, 2015, International Conference on Information and Communication Technology Research (ICTRC2015)

[5] "Towards Author Identification of Arabic Text Articles", by Ahmed Fawziotoom, Emad E Abdullah, Shifaa Jaafar, Aseer

[6] Hamdellh, Dana Amer, 2014 IEEE, 5th International Conference on Information and Communication Systems (ICICS)

[7] Pandian, A., and Md. Abdul Karim Sadiq, 2014, "Authorship Categorization In Email Investigations Using Fisher's Linear Discriminate Method With Radial Basis Function", International Journal of Computer Science, Vol.10,No.6,pp.1003-1014 (SNIP: 0.874)

[8] Al-Falahi Ahmed, Ramdani Mohammad, Bellahfkimustafa, Al-Sarem Mohammad, "Authorship Attribution in Arabic Poetry", 78-1-4799-7560-0/15, 2015, IEEE

[9] Ahmed Fawzi Otoom, Emad E. Abdullah, Shifaa Jaafer, Aseel Hamdallh, Dana Amer "Towards Author Identification of Arabic Text Articles", 2014, IEEE, 5th International Conference on Information and Communication Systems (ICICS)

[10] Bhargava Urala k, A.G.Ramakrishnan and Sahil Mohammad, "Recognition of Open Vocabulary, Online Tamil Handwritten Pages in Tamil Script", 2014 IEEE, Vol.42, No.3, pp.6-9.

[11] Pandian, A., and Md. Abdul Karim Sadiq, 2012, "Detection of Fraudulent Emails by Authorship Extraction", International Journal of Computer Application Vol.41, No.7, pp.7 – 12.

[12] Pandian, A., and Md. Abdul Karim Sadiq, 2013, "Authorship Attribution In Tamil Language Email For Forensic Analysis", International Review on Computers and Software, Vol. 8, No. 12 , pp.2882-2888, (SNIP: 1.178).

[13] M.Mahalakshmi, MalathiSharavanan, "Ancient Tamil Script Recognition and Translation Using LabVIEW", IEEE, 2013, International conference on Communication and Signal Processing, April 3-5.

[14] Farkhund Iqbal, Hamad Binsalleeh, Benjamin C.M. Fung, Mourad Debbabi, 2010,"Mining writeprints from anonymous e-mails for forensic investigation",Digital Investigation(Elsevier), Vol.7, pp.56-64

[15] Bagavandas, M., Hameed, A., Manimannan G, 2009, "Neural Computation in Authorship Attribution: The Case of Selected Tamil Articles", Journal Quantitative Linguistics, Vol.16, No.2, pp.115-131.

[16] R Chandrasekaran and G Manimannan, 2013, "Use of Generalized Regression Neural Network in Authorship Attribution", International Journal of Computer Applications, Vol.62,No.4,pp.7-10.

[17] Pandian, A., and Md. Abdul Karim Sadiq, 2014, "A study of Authorship Identification Techniques in Tamil Articles", International Journal of Software and Web Sciences, Vol. 7 No.1, pp.105-108.

[18] Farkhund Iqbal, Hamad Binsalleeh, Benjamin C.M. Fung, Mourad Debbabi, 2010,"Mining writeprints from anonymous e-mails for forensic investigation",Digital Investigation (Elsevier), Vol.7, pp.56-64.