Authorship Identification for Tamil Classical Poem using Subspace Discriminant Algorithm

A Pandian¹, V V Ramalingam², K Manikandan³ and R P Vishnu Preet⁴

Department of Computer Science and Engineering, S.R.M. Institute of Science and Technology, Chennai, India
³Department of IT, S.R.M. Institute of Science and Technology, Chennai, India
pandianmtech@gmail.com

Abstract. The Development of extensive perceiving confirmation of a creator’s work consolidates stylometry examination that joins various fascinating issues. Extraction of specific kind of highlights from the substance draws in us with the chance to perceive the producers of obscure works. Center of this paper is to briefly recognize the creators of unidentified Tamil dataset in context of crafted by known creators. Content preparing is the technique for getting amazing data from the dataset that joins quantifiable highlights from the dataset. This paper proposes content preparing method to concentrate features and perform grouping on the same. Crafted by a unidentified sonnet or content can be discovered in light of performing arrangement on potential creators’ past known work and building a classifier to characterize the obscure lyric or content in any dialect. This procedure can be additionally reached out to every single provincial dialect around the globe. Numerous writing analysts are thinking that it’s hard to sort ballads as the writers of them are not recognized. By playing out this procedure, creators of different lyrics in Tamil vernacular can be perceived which will be significant to the general public.

1. Introduction
Creators of different local dialect writings are not yet seen. For example, in Tamil vernacular different lyrics still stay authorless. Remembering them would be having more utilize. In light of different explores, for reasons unknown an immense piece of the authorless lyrics can be related with one of the writers, whose name and works is as of now known. In this way, by utilizing a sensible calculation, creators for the unidentified work can be seen. Thomas Bayes (1871) was the first to utilize quantifiable hypothesis for fathoming origin issues in the federalist papers. Auguste de Morgan as perfect on time as in 1851 has recommended the mean length of words as a measure to choose origin issue. Seeing the creators of a sonnet on the premise of complex characters is the creator attribution issue in etymological research. Tallying highlight extraction can contribute more to this initiation issue, which contains extraction of a critical piece of the from time to time used words, length of sentence, exceptional characters utilized and so on. In 1, the creator shows a summary of parts that can be utilized to perform include extraction from datasets. Order was performed on Enron E-mail dataset using desire – amplification calculation and bisecting K-implies calculation that gives 90 % accuracy. In 2, the creator utilizes the highlights recorded in the table – 1 for the origin ID issue. The creator utilizes 800 occasions of a tamil dataset for the same. C
4.5 calculation is utilized to characterize the creators in view of their past work. An exactness of 84.66% is accomplished by the creator by utilizing a similar calculation.

In 3, the creator utilizes mukkoodar pallu dataset which comprises of 800 lyrics that are in Tamil dialect. Utilizing bayes net calculation, the creator has accomplished a classifier exactness of around 94.1%. The highlights utilized by the creator are recorded in the table – 1 on which a choice tree was built to choose the most important highlights.

In 4, 456 events having a place with 7 creators of Arabic works are utilized to perform grouping utilizing bolster vector machines, neural frameworks and markov chains. Bolster Vector Machine performs order by building a classifier display that does out each case to either portrayal, making a non-probabilistic parallel direct classifier. Neural frameworks are virtual impression of neuron cells that are accessible in human identity. These neural frameworks work in a vague course from neurons in cerebrum are impelled in the human personality. Markov chain estimation performs gathering precisely when the markov property is fulfilled. A pinnacle accuracy of 82% is master.

In 5, the creator shows the system to separate highlights from Tamil dataset that includes 28420 characters and 5000 words with an exactness of 72% to 82%. It utilizes FLD and RBF calculations to beat the covering issue. Fisher's Linear Discriminant calculation performs gathering by making a straight blend of parts that detaches no under two classes of things. Spiral Basis Function figuring is vague neural framework systems. It works in context of the neuron parameters. In 6, the creator has used Arabic dataset to remove highlights from the dataset and perform arrangement processon the dataset utilizing markov chain calculation with an accuracy of 96.96%. The creator clears up succinctly the best way to deal with remove highlights applicable to Arabic tongue and perform order on it. Every single one of the parts that are identified with the Arabic dataset and that fulfill the markov property is considered for arrangement. These components are picked and are utilized to manufacture the classifier.

An exactness of 82% is expert on Arabic sonnets featured in 7, which uses bolster vector machine, neural framework and markov chain. In 8, the creator shows how to extricate highlights from old Tamil contents that are digitalized, perform grouping on them utilizing bolster vector machine and bi-gram to accomplish a precision of 83%. N-grams are routinely accumulated from talk or substance corpus. A n-gram of size one is called unigram and a n-gram of size two is called bi-gram.

In 9, the writer scatters the covering issue using fisher's straight discriminant and outspread premise calculation by utilizing Enron email dataset, while in 10, the writer reveals how to think segments to discover the root of an article by utilizing winding reason figuring for gathering in Enron email dataset with an exactness of 80% to 90%. In 11, the creator displays to see tamil letters from their old contents by using LabVIEW device and performs order on the dataset by utilizing division computation. The enron email dataset was amassed by CALO (Cognitive Assistant that Learns and Classifies), which contains information collected from around 150 customers.

In 12, the creator discloses how to remove highlights and discover the precision of the classifier that shows up. Utilizing Enron dataset for email and utilizing 6 sorts of highlights, creators have accomplished an exactness of 90.08%. 68.19% exactness was capable by utilizing flexible city calculation, 79.07% exactness was accomplished by utilizing NBayes calculation, 79.86% precision by utilizing Bayes Net calculation, 88.47% precision by utilizing CMAR calculation, 84.18% accuracy by utilizing CBA calculation and 90.08% exactness by utilizing CMARAA calculation.

In 13-16, the creators portray particular components used to perform characterization and their differentiating exactness. The desire - maximization calculation is an iterative method to perform arrangement. This calculation performs cycle between two stages E and M. The want step (E) makes a quick overview of probability and the growth step (M) opens up the run of the mill probability recorded in the want step.
2. Materials and Methods
The present origin recognizable proof methods reinforce just English compositions. They don’t support Tamil compositions. Finding the writers for un-wrote Tamil compositions find the opportunity to be especially troublesome as there is no framework to remember them strangely. By isolating features imperative to Tamil compositions and by utilizing appropriate estimation, scholars for these dark works can be seen. Gathering is done by utilizing content taking care of system. Content dealing with is the framework for getting top notch data from substance that merges genuine cases from the substance.

Figure 1. Architecture

Figure-1 exhibits the design that is followed in this strategy of grouping. The dataset considered here is Paripadal that includes 70 events. By removing lexical, syntactic and semantic components as illuminated in 16, the order procedure is performed. The once-over of highlights that are considered is appeared in Table-1.

| Features type          | Features                                      |
|------------------------|-----------------------------------------------|
| Lexical:               |                                               |
| character-based        |                                               |
| 1. Character count (N) |                                               |
| 2. Ratio of digits to N|                                               |
| 3. Ratio of letters to N|                                              |
| 4. Ratio of uppercase letters to N|                      |
| 5. Ratio of spaces to N |                                               |
| 6. Ratio of tabs to N  |                                               |
| 7. Occurrences of uyir, mei and uyirmei letters (246 features)| |
| 8. Occurrences of special characters |                        |
| Lexical:               |                                               |
| word-based             |                                               |
| 9. Token count(T)      |                                               |
| 10. Average sentence length in terms of characters |                |
| 11. Average token length|                                          |
| 12. Ratio of characters in words to N |                      |
13. Ratio of short words (1 to 3 characters) to T  

| Syntactic features | 14. Occurrences of punctuations, . ? ! ; ”’(8 features) |
|--------------------|--------------------------------------------------------|
|                    |                                                        |

**Table 1. list of Features**

These features are extracted from the dataset and utilized for performing classification. These features characterize the stylometry of the creator. Stylometry is the application of investigation of composed styles from manually written articles that can be used as a part of authorship identification. Stylometry incorporates extraction of lexical, syntactic and semantic elements applicable to the language considered. Table - 1 demonstrates the lexical, syntactic and statistical features that are extracted from the dataset. By using the C 4.5 algorithm, an accuracy of 70% was achieved. The subspace discriminant algorithm, which is similar to linear discriminant algorithm, is used to perform the classification on the specified data set.

2.1 Feature Extraction

Feature extraction handle assembles an arrangement of derived qualities from the underlying arrangement of information that is planned to human translation. Dataset can't be specifically utilized as a part of the tool to perform arrangement. Just the features that are extricated 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 70% to 90% by including statistical features to the features set and performing some tweaks in the algorithm used. Statistical features include Standard Deviation, Minimum, Maximum, Sum, Mean, Median, Mode,
Kurtosis, Skewness, Standard error, Variance, Count and Range. Kurtosis is a statistical feature that refers to the peak in a frequency – distribution curve. Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The skewness value can be positive or negative, or even undefined. Variance is a statistical measure that tells us how measured data vary from the average value of the set of data. The features recorded in table-1 are extracted from the dataset. The dataset is initially changed over into Unicode format so it can be perused in Microsoft excel. Computers can't comprehend Tamil characters. They bargain just with numbers in their memory. Unicode gives an encoding framework that covers every one of the regional languages and gives an approach to computers to comprehend them. UTF-8, UTF-16 and UCS-2 are the accessible Unicode encoding positions out of which UCS-2 is presently out of date. The encoding utilized as a part of this procedure is UTF-16 which can be perused in Microsoft Excel. 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. Figure 2 shows the queries used in feature extraction.

Figure 2. SQL Queries

2.2 Feature Selection

Feature selection process is done by utilizing decision tree algorithm. A decision tree is made utilizing every one of the features that are listed in table-1 and the best features are chosen in view of the decision tree. The core algorithm to build decision tree is ID3, which is presently known as C4.5 algorithm. Decision tree is built based on two parameters: Entropy and Information Gain. The decision tree is developed from root hub and includes dividing of hubs into subsets that comprises of homogeneous items. Entropy is utilized to gauge the level of homogeneity between the hubs that are available in a subset. Data pick up increments as entropy abatements. Data pick up and entropy is contrarily corresponding to each other. Decision tree development depends on the trait that contains the most elevated data pick up. Decision tree with every one of the features is pruned down to a number which gives most extreme order exactness. Feature selection is done as it defeats the issues of computational cost and erroneous classifier accuracies because of insignificant information. The elements that are recorded in table-1 are altogether chosen by feature selection process as these elements give most extreme classifier accuracy.
2.3 Subspace Discriminant Classification Algorithm

C4.5 algorithm is developed by Ross Quinlon. This algorithm will be a development of the ID3 algorithm that might have been being used sooner times. C4.5 algorithm constructs a choice tree. From those set from claiming preparation information that is utilized In view of the entropy pick up. This calculation picks every hub In view of. The majority of the data gain, which is distinction On entropy What's more parts its subsets adequately. The hub with most elevated data. Get or most reduced entropy will be used to make choice. This technique is iterated for every last one of subsets until there would no further subsets will part. The steps of the algorithm are explained as follows:

1. Check for base cases.
2. For each attribute x, find the information gain by splitting on x.
3. Let x1 be the attribute with highest information gain.
4. Create a node that splits on x1.
5. Iterate on the subsets of x1 and add all the nodes as children of x1.

Following is the implementation of the algorithm:

```matlab
function [trainedClassifier, validationAccuracy] = trainClassifier(trainingData)

% trainClassifier(trainingData)
% returns a trained classifier and its accuracy.
% This code recreates the classification model trained in
% Classification Learner app.
% Input:
% trainingData: the training data of same data type as imported
% in the app (table or matrix).
% Output:
% trainedClassifier: a struct containing the trained classifier.
% The struct contains various fields with information about the
% trained classifier.
% trainedClassifier.predictFcn: a function to make predictions
% on new data. It takes an input of the same form as this training
% code (table or matrix) and returns predictions for the response.
% If you supply a matrix, include only the predictors columns (or
% rows).
% validationAccuracy: a double containing the accuracy in
% percent. In the app, the History list displays this
% overall accuracy score for each model.
% Use the code to train the model with new data.
% To retrain your classifier, call the function from the command line
% with your original data or new data as the input argument trainingData.
% For example, to retrain a classifier trained with the original data set
% T, enter:
% [trainedClassifier, validationAccuracy] = trainClassifier(T)
% To make predictions with the returned 'trainedClassifier' on new data
% T,
% use
% yfit = trainedClassifier.predictFcn(T)
% To automate training the same classifier with new data, or to learn how
% to programmatically train classifiers, examine the generated code.
% Auto-generated by MATLAB on 02-Mar-2017 16:45:58
```

% Extract predictors and response
% This code processes the data into the right shape for training the
% classifier.
inputTable = trainingData;
predictorNames = {'uyir', 'mei', 'uyirmei', 'totalcharacter', 'totallines',
    'totalspace', 'totalwords', 'charspaceratio', 'avgsentlen', 'avgwordlen',
    'sum', 'min', 'max', 'mean'};
predictors = inputTable(:, predictorNames);
response = inputTable.authorname;
isCategoricalPredictor = [false, false, false, false, false, false, false, false,
    false, false, false, false, false, false, false, false, false];
% Train a classifier
% This code specifies all the classifier options and trains the classifier.
subspaceDimension = max(1, min(7, width(predictors) - 1));
classificationEnsemble = fitensemble(...
    predictors, ...
    response, ...
    'Subspace', ...
    30, ...
    'Discriminant', ...
    'Type', 'Classification', ...
    'NPredToSample', subspaceDimension, ...
    'ClassNames', {'a'; 'b'; 'c'; 'd'; 'e'; 'f'; 'g'});
% Create the result struct with predict function
predictorExtractionFcn = @(t) t(:, predictorNames);
ensemblePredictFcn = @(x) predict(classificationEnsemble, x);
trainedClassifier.predictFcn = @(x) ensemblePredictFcn
(predictorExtractionFcn(x));
% Add additional fields to the result struct
trainedClassifier.RequiredVariables = {'uyir', 'mei', 'uyirmei',
    'totalcharacter', 'totallines', 'totalspace', 'totalwords', 'charspaceratio',
    'avgsentlen', 'avgwordlen', 'sum', 'min', 'max', 'mean'};
trainedClassifier.ClassificationEnsemble = classificationEnsemble;
trainedClassifier.About = 'This struct is a trained classifier exported from
Classification Learner R2016a.';
trainedClassifier.HowToPredict = sprintf('To make predictions on a new
    table, T, use: \n    yfit = c.predictFcn(T) \n    replacing "c" with the name of
    the variable that is this struct, e.g. "trainedClassifier". \n    The table, T, must contain the variables returned by: \n    c.RequiredVariables \n    Variable formats (e.g. matrix/vector, datatype) must match the original training
    data. \n    Additional variables are ignored. \n    For more information, see
    <a href="matlab:helpview(fullfile(docroot, "stats", "stats.map"),
    "appclassification_exportmodeltoworkspace")">How to predict using an
    exported model</a>.';
% Extract predictors and response
% This code processes the data into the right shape for training the
% classifier.
inputTable = trainingData;
predictorNames = {'uyir', 'mei', 'uyirmei', 'totalcharacter', 'totallines', 'totalspace', 'totalwords', 'charspaceratio', 'avgsentlen', 'avgwordlen', 'sum', 'min', 'max', 'mean'};
predictors = inputTable(:, predictorNames);
response = inputTable.authorname;
isCategoricalPredictor = [false, false, false, false, false, false, false, false, false, false, false, false, false, false, false];
% Perform cross-validation
partitionedModel = crossval(trainedClassifier.ClassificationEnsemble, 'KFold', 5);
% Compute validation accuracy
validationAccuracy = 1 - kfoldLoss(partitionedModel, 'LossFun', 'ClassifError');
% Compute validation predictions and scores
[validationPredictions, validationScores] = kfoldPredict(partitionedModel);

3. Results and Discussions
The confusion matrix obtained by performing classification using subspace discriminant algorithm is shown in Table 2.

|   | A   | B   | C   | D   | E   | F   | G   |
|---|-----|-----|-----|-----|-----|-----|-----|
| A | 21  | 1   | 0   | 0   | 0   | 0   | 0   |
| B | 2   | 20  | 0   | 0   | 1   | 0   | 0   |
| C | 1   | 1   | 0   | 0   | 0   | 0   | 0   |
| D | 2   | 0   | 0   | 0   | 0   | 0   | 0   |
| E | 3   | 1   | 0   | 0   | 0   | 1   | 0   |
| F | 0   | 2   | 0   | 0   | 0   | 5   | 1   |
| G | 1   | 0   | 0   | 0   | 0   | 2   | 4   |

Table 2. Confusion Matrix

It can be seen that 21 instances of the author A are correctly classified and 1 instance is incorrectly classified. 20 instances of author B are correctly classified while 2 instances are not classified correctly. There are no correctly classified instances of three authors C, D and E while 5 instances of the author F and 4 instances of the author G are correctly classified. This leads to an accuracy of 72.5%.

4. Conclusion
The highlights recorded in table-1 were considered and the highlights were chosen by developing a choice tree utilizing the calculation specified. Certain highlights from the rundown of considered highlights list were chosen keeping in mind the end goal to overcome over fitting of the classifier. In this way the classifier was prepared and has delivered a precision of around 72.5%. In this manner by separating general highlights that are basic for every single provincial dialect, a general initiation recognizable proof framework can be created for every single local dialect.
References:

[1] Sara El Manar and Mohammed V Souissi 2014 Authorship Analysis Studies International Journal of Computer Applications 86

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

[3] Pandian A, Ramalingam V V and Vishnu Preet R P 2016 Authorship Identification for Tamil Classical Poem (Mukkoodar Pallu) using C4.5 Algorithm Indian Journal of Science and Technology, 9(46), DOI: 10.17485/ijst/2016/v9i46/107944

[4] Pandian A, Ramalingam V V and Vishnu Preet R P 2016 Authorship Identification for Tamil Classical Poem (Mukkoodar Pallu) using Bayes Net Algorithm Indian Journal of Science and Technology, 9(46) DOI: 10.17485/ijst/2016/v9i46/107910

[5] Ahmed Fawziotoom, Emad E Abdullah, Shifaa Jaafar, Aseer Hamdellh and Dana Amer 2014 Towards Author Identification of Arabic Text Articles, IEEE, 5th International Conference on Information and Communication Systems (ICICS)

[6] 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)

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

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

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

[10] 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.

[11] 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).

[12] M.Mahalakshmi, MalathiSharavanavan, 2013 Ancient Tamil Script Recognition and Translation Using LabVIEW, IEEE, 2013, International conference on Communication and Signal Processing, pp 3-5.

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

[14] 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.

[15] 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.

[16] 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.

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