AUTHORSHIP ANALYSIS FOR REGIONAL LANGUAGES USING MACHINE LEARNING APPROACH

A. Pandian[1], J. Venkata Subramanian[2], V. V. Ramalingam[1] and K. M. Uma Maheswari[1]
[1]Department of Computer Science Engineering, SRM University, Kattankulathur, Tamil Nadu, India
[2]Department of CA, SRM University, Kattankulathur, Tamil Nadu, India
E-Mail: pandian.a@ktr.srmuniv.ac.in

ABSTRACT

The old classical poems, written in various regional languages, many of them the authors were not identified. For example, in Tamil language, Agananuru, Purananuru and Paripadal, still we didn't know many of the authors. Hence, if we identify these, it will be more helpful to the society to know and identify the author of various valuable old poems. The Author Identification study is useful to identify the most plausible authors and best suited for authorship verification whereas it can be applied to authorship characterization and profiling.

Keywords: classical poems, text processing, authorship analysis, authorship identification, machine learning.

1. INTRODUCTION

Since the start of Research in the area of the author identification, several researchers in the field have created and given to the scientific community their own corpus, which today serves as a framework of standardized tests. In this sense, the different corpus available together with a description, details of its format, content, structure and number of poems written. They can be tested with all possible features that were used in the translator attribution experiments with the same machine learning methods by training on the original English writing samples of thirteen authors.

The highest accuracy yielding feature set was again the 'Translation Feature Set'. There are differences in terms of subject matter, the distribution and legitimate pre-processing performed. The advent of non-traditional authorship attribution techniques can be traced back to 1887, when Mendenhall first created the idea of counting features such as word length. His work was followed by work from Yule [13] with the use of sentence lengths to judge authorship.

By and large, research has focused on different aspects of the 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 [14] 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 [15] by quantifying some of its features.

Most stylometry [16, 17] studies employ items of language and most of these are lexically based. The usefulness of function words in Authorship attribution has been examined [18]. Experiments were conducted with Support Vector Machine classifiers in twenty novels and success rates above 90% were obtained. The use of function words is a valid and good approach in attribution of authorship [19]. 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 [20].

Concurrently experiments conducted with support vector classifiers [18] detected authors with 60-80% success rates using different parameters. The effect of word sequences in authorship [21] 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 function and content words.

The experiments are conducted on a dataset consisting of poems using naïve Bayes classifier [22]. Later authorship studies [23] 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.

1.1 Applications of authorship identification

To analyze anonymous or disputed documents and books such as the ancient articles and poems written by various authors.

Plagiarism detection - to establish whether claimed authorship is valid.

Criminal Investigation - to determine the source of unauthorized or unsolicited Emails

Forensic investigations - verifying the authorship of spam mails newsgroup messages, or identifying the basis of a piece of intelligence.

1.2 Key Features for identification of authors

- When an author writes they use certain words unconsciously.
- Find some underlying 'fingerprint' for an author's style.
- The fundamental assumption of authorship attribution is that each author has habits in wording that make their writing unique.
- It is well known that certain writers can be quickly identified by their writing style.
- Extract features of the given text that differentiate an author from another
- Applying certain statistical or machine learning techniques on giving training data.

- Showing examples and counterexamples of an author's work

1.3 Issues involved in the process

Identification of authors needs expertise in linguistics, statistics, text authentication, literature, etc. Hence, this is an interdisciplinary area. Too many style measures have to be applied and style markers need to be determined. Although statistical methods may be complicated or simple, too many exist in the literature. The features are extracted only after parsing all the documents thoroughly. The results have to be combined in order to obtain certain characteristics about the authors. Apply each of the statistical or machine learning approaches to assign a given document to the most likely author.

1.4 Current techniques

Computerized analysis of documents was developed in 1980's, from the previous statistical analysis of literary style. This is termed "Stylometry". In order to quantify some of the features of an author's style, the following measures are explored.

Word or sentence length: This is a method developed in the origin of stylometry. Because of the naïve quantification, it is not a reliable method.

Function words: This method relies on word usage and context-free words. Using this method, we can analyze word frequency, position, and immediate context of words. This is a criticized method, and cannot reliably distinguish between certain literature types.

Vocabulary distributions: In this method, we measure the richness or diversity of an author's vocabulary. It analyzes the frequency profile of word usage to glimpse the author's extent of vocabulary.

Content analysis: This method tabulates the frequency of types of words in a text. It aims to reach the denotative or connotative meaning of the text.

2. RELATED WORK

2.1 Authorship analysis

Authorship attribution is particularly concerned with the identification of the real author of a disputed anonymous document. In the literature, authorship identification is considered as a text categorization or text classification problem. The process starts by data cleaning followed by feature extraction and normalization. Each suspected document is converted into a feature vector [42]; the suspect represents the class label. Feature values are calculated by using Stylometric features. The extracted features are classified into two groups: training and testing sets. The training set is used to develop a classification model whereas the testing set is used to validate the developed model by assuming the class labels are not known. Common classifiers include decision trees, neural networks and Support Vector Machine [42].

Authorship attribution studies differ in terms of the Stylometric features used and the type of classifiers employed. References [43] and [44] describe two approaches which attempt to mine e-mail authorship for the purpose of computer forensics. Authors extract various e-mail document features including linguistic features, header features, linguistic patterns and structural characteristics. All these features are used with the Support Vector Machine (SVM) learning algorithm to attribute authorship of e-mail messages to an author.

The framework for authorship identification in online messages to deal with the identity-tracing problem. In this framework, four types of writing style features (lexical, syntactic, structural and content-specific features) are extracted from English and Chinese online-newsgroup messages. Comparison has been made between three classification techniques: decision tree, SVM and back-propagation neural networks. Experimental results showed that this framework is able to identify authors with a satisfactory accuracy of 70 to 95% and the SVM classifier outperformed the two others.

The Function words and applies five classifiers (Naive Bayesian, Bayesian networks, Nearest-neighbour method, Decision Trees, SVM). The data analyzed is a collection of newswire articles from the AP (Associated Press) sub-collection.

2.2 Authorship characterization

Authorship characterization is used to detect sociolinguistic attributes like gender, age, occupation and educational level of the potential author of an anonymous document [24]. The studies about the effects of gender attributes on authorship analysis. Other studies discussed the educational level, age and language background. The studies collected information about gender, age and occupation of the writer of an anonymous chat segment.

2.3 Authorship verification or similarity detection

Studies consider the problem of authorship verification as a similarity detection problem: to determine whether two texts are produced by the same person without knowing the real author of the document [24].

A new algorithm to identify when two aliases belong to the same individual, while preserving privacy. The technique has been successfully applied to postings of different bulletin boards, achieving more than 90% accuracy.

References [25] and [26] present a novel technique called write prints for authorship identification and similarity detection. Authors used in the experimentation extended feature list, including idiosyncratic features. Authors take an anonymous entity, compare it with all other entities, and then calculate a score. If the score is above a certain predefined value, the entity is clustered with the matched entity.

Reference [27] proposes an approach called linguistic profiling. In this study [27] proposed some distance and scoring functions for creating profiles for a group of example data. The average feature counts for each author was compared with a general stylistic profile built from the training samples of widely selected authors. The study focused on detecting similarity between student essays for plagiarism and identity theft.

3. PROCEDURE TO IMPLEMENT

Data collection

Collect Materials written by potential authors from various sources and Digitized.

Feature extraction

After extraction, each unstructured text is represented as a vector of writing-style features.

Model generation

Dataset should be divided into training and testing set. Classification techniques should be applied. An iterative training and testing process may be needed

Authorship identification

Developed model can be used to predict the authorship.

3.1 Machine learning classifiers and clustering algorithms

The use of machine learning classifiers and clustering algorithms marked an important turning point in authorship attribution studies. The application of such methods is straightforward: training texts are represented as labeled numerical vectors and learning methods are used to find boundaries between classes (authors) that minimize some classification loss function.

To predict the performance of a particular algorithm, accuracy measure, precision and recall are used. They are defined as:

$$\text{Accuracy} = \frac{\text{Number of messages whose author was correctly identified}}{\text{Total number of messages}}$$

$$\text{Precision} = \frac{\text{Number of messages author correctly assigned to the author}}{\text{Total number of messages assigned to the author}}$$

$$\text{Recall} = \frac{\text{Number of messages author correctly assigned to the author}}{\text{Total number of messages written by the author}}$$

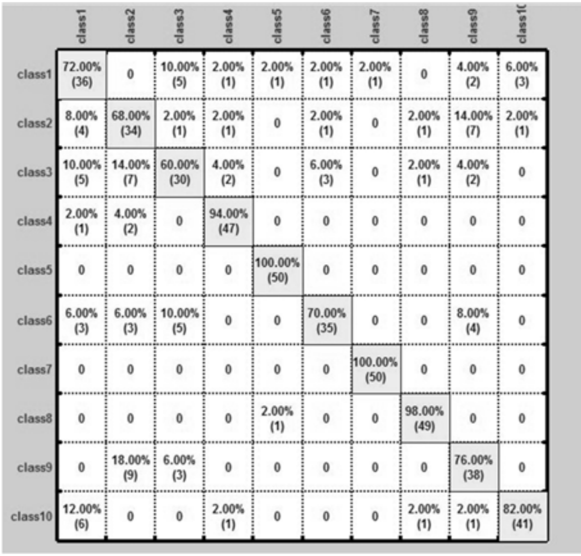
4. EXPERIMENTAL RESULTS

In this heuristics on Tamil text strings of various parameters such as spacing, punctuation mark or other induction marks are not prioritized. In this proposed method, the Tamil text is manipulated into images, and the algorithm has the potential to analyze with 100 various image fields (simplified into 10 classes) that includes magazines, OCR extracted images, newspaper and printed colour advertisement. It consists of signature based Tamil font style, non- structured layout and it's background texture, patterns are analyzed in the stage-wise process. In the test Tamil images, there are 28420 characters and 5000 words are analyzed with trained images. Around 25262 characters and 3700 words are successfully analyzed through this proposed method. It is very reliable and robust and stable.

Table-1. Comparison of the various evaluation metrics with the proposed system.

Evaluation metrics		SVM_LP boosting + DCT	SVM_LP boosting + DST	SVM_LP boosting + DWT
Input text strings texture images for various classes	TP	3700	3500	3800
	TN	800	800	900
	FP	200	200	100
	FN	300	500	200
	Sensitivity	0.925	0.875	0.95
	Specificity	0.73	0.62	0.9
	Accuracy	0.9	0.86	0.94
	Total error (%)	10	14	6

In this analysis, the text strings texture images of various sets are taken and it divided into class 1 to class 10. The text strings image classification accuracy of the proposed system is evaluated using the evaluation metrics, such as sensitivity, specificity and accuracy that based Zhu et al. (2010) is defined. Based on the confusion matrix, the error in the LP boosting classifier is clearly shown in various Text string texture image classes. It is noted that the performance of the algorithm efficiently improved when the machine classifier analyze the Text strings texture images in the "class 5". The similarities of “class 5” to compare with other class image are significantly reduced during the process of retrieving the "class 5" images.



	class1	class2	class3	class4	class5	class6	class7	class8	class9	class10
class1	72.00% (36)	0	10.00% (5)	2.00% (1)	2.00% (1)	2.00% (1)	2.00% (1)	0	4.00% (2)	6.00% (3)
class2	8.00% (4)	68.00% (34)	2.00% (1)	2.00% (1)	0	2.00% (1)	0	2.00% (1)	14.00% (7)	2.00% (1)
class3	10.00% (5)	14.00% (7)	60.00% (30)	4.00% (2)	0	6.00% (3)	0	2.00% (1)	4.00% (2)	0
class4	2.00% (1)	4.00% (2)	0	94.00% (47)	0	0	0	0	0	0
class5	0	0	0	0	100.00% (50)	0	0	0	0	0
class6	6.00% (3)	6.00% (3)	10.00% (5)	0	0	70.00% (35)	0	0	8.00% (4)	0
class7	0	0	0	0	0	0	100.00% (50)	0	0	0
class8	0	0	0	0	2.00% (1)	0	0	98.00% (49)	0	0
class9	0	18.00% (9)	6.00% (3)	0	0	0	0	0	76.00% (38)	0
class10	12.00% (6)	0	0	2.00% (1)	0	0	0	2.00% (1)	2.00% (1)	82.00% (41)

Figure-1. Confusion matrix analyses for the proposed Tamil text extraction in dictionary.

In the field of machine learning, a confusion matrix, also known as a contingency table or an error matrix , is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one (in unsupervised learning it is usually called a matching matrix). Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. The name stems from the fact that it makes it easy to see if the system is confusing two classes.

The performance evaluations of the proposed texture classification system are identified. From each original image, 128x128 pixel sized images are extracted with an overlap of 32 pixels between vertical and horizontal direction. The performance is measured through the SAS (Sensitivity, Specificity and Accuracy) parameters

$$\begin{aligned}&Sensitivit\ y = TP/(TP + FN)\\&Specificity = TN/(TN + FP)\\&Accuracy= (TN+TP)/(TN+TP+FN+FP)\end{aligned} \qquad (35)$$

Where TP stands for True Positive, TN stands for True Negative, FN stands for False Negative and FP stands for False Positive. As suggested by above equations, sensitivity is the proportion of true positives that are correctly identified by a diagnostic test. It shows how good the test is at detecting a texture features based on the classifier. Specificity is the proportion of the true negatives correctly identified by a trained test. It suggests how good the test is at identifying normal (negative) condition. Accuracy is the proportion of true results, either true positive or true negative, in a population. It measures the degree of veracity of a diagnostic test for a condition that analyzed in [18].

Table-2. Error analysis of various transform method based on SVM_LP boosting classifier.

Methods	Mean error
SVM_LP boosting+DWT	7.5
SVM_LP boosting+DCT	12.5
SVM_LP boosting+DST	17.5

In the above Table-2, the sensitivity of the proposed SVM_LP boosting + DWT approach is better compared to other methods SVM_LP boosting + DST and SVM_LP boosting + DCT. The specificity for the proposed design LP boosting + DWT leads by 0.17% and 0.28% of the existing SVM_LP boosting + DST and SVM_LP boosting + DCT method respectively. Similarly, the accuracy of SVM_LP boosting +DWT is extremely higher than all other approaches.

Based on the experimental results, the proposed system classification error rate is less than the other classifier; It is seen that the proposed method error ratio is only 7.5% for text strings image datasets whereas the SVM_LP boosting +DCT and SVM_LP boosting + DST methods have an error rate of 12.5% and 17.5% respectively. Compared to existing methods, the proposed SVM_LP boosting + DWT algorithm is much more sophisticated for the classification of text strings texture images.

5. CONCLUSIONS

Authorship Identification in each regional language has major impact. Various Authorship Identification commercial softwares are available only for English language. But, there is no algorithm or software for Tamil language or regional languages. We extend these works to Mukkoodarpallu and Muthollayiram in Tamil. (Because these poems are completely authorless) It. can be implemented for other Regional Languages also.

REFERENCES

[1] Bagavandas, M., Hameed, A., Manimannan G. 2009. Neural Computation in Authorship Attribution: The Case of Selected Tamil Articles. Journal of Quantitative Linguistics. 16(2): 115-131.

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

[3] Pandian A. and Md. Abdul Karim Sadiq. 2012. Detection of Fraudulent Emails by Authorship

Extraction. International Journal of Computer Applications. 41(7): 7-12.

[4] Pandian A. and Md. Abdul Karim Sadiq. 2013. Authorship Attribution In Tamil Language Email For Forensic Analysis. International Review on Computers and Software. 8(12): 2882-2888, (SNIP: 1.178)

[5] R Chandrasekaran and G Manimannan. 2013. Use of Generalized Regression Neural Network in Authorship Attribution. International Journal of Computer Applications. 62(4): 7-10.

[6] Pandian A. and Md. Abdul Karim Sadiq. 2014. Authorship Categorization in Email Investigations Using Fisher's Linear Discriminate Method with Radial Basis Function. Journal of Computer Science. 10(6): 1003-1014 (SNIP: 0.874).

[7] Pandian A. and Md. Abdul Karim Sadiq. 2014. A study of Authorship Identification Techniques in Tamil Articles. International Journal of Software and Web Sciences. 7(1): 105-108.

[8] Manimannan G and Lakshmi Priya r. 2014. Identification of Disputed Writings in Tamil Articles Using Multivariate Statistical Techniques. IOSR Journal of Mathematics. 10(2): 01-07.

[9] Pandian A. and Md. Abdul Karim Sadiq. 2014. A Study of Authorship Attribution in English and Tamil Emails. Research Journal of Applied Sciences, Engineering and Technology. 8(2): 203-211 (ISI Thomson Indexed & SNIP: 0.569).

[10] Pandian A. and Md. Abdul Karim Sadiq. 2014. Innovative Methods in Identifying Authors of Documents. International Journal of Engineering and Technology. 6(6): 2512-2520 (SNIP: 0.581).

[11] Michael R. Schmid a, Farkhund Iqbal b, Benjamin C.M. Fung c. 2015. E-mail authorship attribution using customized associative classification. Digital Investigation (Elsevier). 14: 116-126.

[12] Pandian A and Md. Abdul Karim Sadiq. 2015. A Novelty Approach On Tamil Spam Text Extraction By Using Texton Template Based Support Vector Machine And LPBoosting Classifier. ARPN Journal of Engineering and Applied Sciences. 10(12): 5265-5276 (SNIP: 0.401).

[13] G. Yule. 1938. On sentence-length as a statistical characteristic of style in prose, with application to two cases of disputed authorship. Biometrika. 30: 363-390.

[14] Goodman R., Hahn M., Marella M., Ojar C. and Westcott S. 2007. 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.

[15] 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.

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

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

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

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

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

[21] Abbasi, H. Chen. 2005. Applying authorship analysis to extremist-group web forum messages. IEEE Intelligent Systems. 20(5): 67-75.

[22] F. Peng, D. Shuurmans, S. Wang. 2004. Augmenting naive Bayes classifiers with statistical language models. Information Retrieval Journal. 7(1): 317-345.

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

[24] F. Iqbal. 2011. Messaging Forensic Framework for Cybercrime Investigation. A Thesis in the Dept. of Computer Science and Software Engineering-Concordia University Montréal, Canada.

[25] A. Abbasi, H. Chen. 2008. Writeprints: A Stylometric approach to identity-level identification and similarity detection in cyberspace. ACM Transactions on Information Systems. 26(2): 1-29.

[26] A. Abbasi, H. Chen, J. Nunamaker. 2008. Stylometric identification in electronic markets: Scalability and robustness. Journal of Management Information Systems. 5(1): 49-78.

[27] H. VanHaltern. 2007. Author verification by linguistic profiling: An exploration of the parameter space. ACM Transactions on Speech and Language Processing.