Authorship Identification for Tamil Classical Poem (Mukkoodar Pallu) using Bayes Net Algorithm

Pandian* , V. V. Ramalingam and R. P. Vishnu Preet
Department of Computer Science and Engineering, SRM University, Kattankulathur, Chennai - 603203, Tamil Nadu, India; pandian.a@ktr.srmuniv.ac.in, ramalingam.v@ktr.srmuniv.ac.in, vishnupreet rp@gmail.com

Abstract

Objective: To classify the authors of unknown Tamil dataset based on the work of known authors. Methods/Analysis: Text processing is the method of deriving high quality information from text that includes statistical patterns from the text. This paper proposes text processing method to extract features and perform classification on the same. Findings: The accuracy of the classifier turns out to be 94.1%. Classifier accuracy is improved from 88.23% to 94.1% by varying the classification algorithm (Bayes Net). Novelty/Improvement: This method can be further extended to all regional languages. By doing this, authors of various other poems in Tamil language can be identified which will be helpful to the society.

Keywords: Authorship, Classification, Feature Selection, Tamil Articles

1. Introduction

Authors of many regional language poems are not yet identified. For instance, in Tamil language many poems are still anonymous. Identifying them would be of more use. Based on various researches, it turns out that most of the authorless poems can be associated with one of the authors, whose name and work is already known. So by using a suitable algorithm, authors for the unknown work can be identified. Thomas Bayes (1871) was the first to use statistical theory for solving authorship issues in the federalist papers. Auguste de Morgan as early as in 1851 has suggested the mean length of words as a measure to resolve authorship problem.

Identifying the writer of an article on the basis of stylistic character is the author attribution problem in linguistic research. Feature extraction can contribute more to this authorship problem, which consists of extraction of frequently used words, length of sentence, special characters used etc.

In, explains what features can be extracted from the dataset and find the accuracy of the classifier model. Enron dataset for e-mail was used with 6 different algorithms and it turns out that the accuracy is 90.08%. Adaptive metropolis algorithm gives an accuracy of 68.19%, N Bayes algorithm gives an accuracy of 79.07%, 79.86% accuracy by using Bayes Net algorithm, 88.47% accuracy by using CMAR algorithm, CBA algorithm gives an accuracy of 84.18% and 90.08% accuracy by using CMARAA algorithm.

In proposed the usage of sonnets to perform authorship identification extracted from the text. Identification was performed by normalizing the sonnets that were
collected from various federalist papers. An easy way to extract features from regional languages is depicted elaborately\(^3\), which uses Slovakian dataset and compaction algorithm.

In\(^4\) explained how to use AT and SAT models to perform authorship attribution on the given dataset. The authors use random kitchen algorithm to perform the classification on the dataset to find the author of unknown texts\(^5\).

In\(^6\) developed a method to identify Tamil words from online datasets, extraction of features from them and performing classification on them is explained. In\(^2\) explained the way to perform authorship attribution on Enron email dataset by performing classification on it by using random forest algorithm.

In\(^7\) explained how to identify and extract features both local and global to identify numerals and Tamil words from online articles. In\(^8\) provided a way to identify and extract features that are pertinent to Kannada language to identify the authors of unknown articles.

In\(^9\), the author attribution is done on Tamil datasets using techniques like reevaluation and pattern analysis to find the authors of unknown texts while reference\(^10\) shows how to recognize Tamil texts from online datasets and perform classification on them based on random kitchen algorithm.

### 2. Materials and Method

The present authorship identification methods are limited to English only. Other regional languages are not supported. Authorship identification can be performed in two ways: Natural Language Processing and Text Processing. In this paper, text processing method is proposed to perform authorship identification process. Text processing method uses features like character count, sentence count; functional words count and perform classification to find the authors of unknown texts. A list of features that can be used to perform classification is mentioned in\(^1\) Table 1.

| Features type | Features |
|---------------|----------|
| Lexical:      |          |
| character-based|         |
| 1. Character count (N) |
| 2. Ratio of digits to N |

These features are extracted from the dataset and used for performing classification. These features define the stylometry of the author. Stylometry is the application of study of written styles from handwritten articles that can be used in authorship identification. Stylometry includes extraction of lexical, syntactic and semantic features pertinent to the language considered. Above table shows the lexical and syntactic features that are extracted from the dataset. By using the Bayes Net algorithm, an accuracy of 94.1% was attained.

![Architecture](figure1.png)

**Figure 1.** Architecture.

Bayes Net algorithm uses Simple Estimator A0.5 for estimating the instances and K2 for searching through the instances. Decision tree algorithm gives an accuracy of 88.23% by varying the confidence factor and number of objects, while Bayes Net algorithm gives an accuracy of 94.1%. The confusion matrix, Figure 1 shows the correctly classified instances which includes the work of three authors.
2.1 Feature Extraction
As the classification process cannot be done directly on the dataset, features that are useful for building classifier have to extract from the dataset. These features cannot be extracted manually and so macro is written to perform feature extraction iteratively. The lists of features that are shown in Table 1 are considered and are used to perform classification. These features have to be extracted from the dataset first. In order to extract features, Tamil dataset is converted into Unicode first. By using Microsoft excel, the listed features are extracted by using macro. A macro is a small piece of code that is used to perform certain operations recursively.

2.2 Feature Selection
Feature selection is done by building a decision tree based on C4.5 algorithm. Decision tree is built based on two factors, namely Information Gain and Entropy. Information gain and Entropy are inversely proportional. That is if entropy is more, information gain is less. Decision tree is built from root node based on information gain. From the root node, further nodes are built based on information gain. Node which has got highest information gain tends to be the next node and forms a subset of nodes based on the homogeneity of information gain.

Decision tree is used to prune the number of features to a minimum number without affecting the classifier accuracy. Pruning is done based on the information gain and entropy difference. Feature selection is done in order to reduce the number of unwanted features that are initially considered. Problem with lot of features is that it could either affect the classifier accuracy or increase the number of unwanted features. Selection of relevant features lead to increased classifier accuracy. The lists of selected features are listed in the table-1.

2.3 Classification Algorithm
The classification algorithm used is Bayes Net. Bayes Net or Bayesian Network algorithm works is a probabilistic graphical model which represents random variables and their corresponding conditional dependencies in a directed acyclic graph. Bayes Net works on the basis of probabilities of events. It simply creates a network based on the probabilities of all the events specified. These probabilities are especially useful to predict if some other similar event could occur.

The classifier is built based on the architecture shown in figure-1. All the steps are followed one by one to build the classifier.

3. Results and Discussion
The Bayes Net algorithm has produced an accuracy of 94.1% on Mukkoodar Pallu dataset using Weka tool. Initially the classifier accuracy obtained was 88.23% using C4.5 algorithm. The decision tree algorithm C4.5 produced an accuracy of 76.4% on the dataset. To improve the classifier accuracy, two parameters: confidence factor and minimum number of objects were varied. Suppose if you choose the confidence factor as 0.2 shown in Figure-2 and minimum number of objects as 4 shown in Figure-3, the classifier accuracy was increased to 88.23%.

![Figure 2. Confidence factor vs. accuracy.](image)

![Figure 3. Minimum number of objects vs. accuracy.](image)

Bayes Net algorithm has performed better than C4.5 algorithm by using simple estimator and K2 algorithm for estimation and has improved the classifier accuracy. Bayes Net uses two parameters: simple estimator and K2 algorithm for searching. Simple estimator works by increasing the gap by 0.5. The K2 algorithm is used to uncover the underlying structure from the pre-determined nodes by using greedy algorithm. Table 2 shows the confusion matrix obtained...
after classification. The poems of author X and Z are all correctly classified. The work of author Y is only 50% accurate as only one poem is classified correctly and the other is wrongly classified. This gives an overall classifier accuracy of 94.1%.

Table 2. Confusion matrix

|      | X  | Y  | Z  |
|------|----|----|----|
| X    | 11 | 0  | 0  |
| Y    | 1  | 1  | 0  |
| Z    | 0  | 0  | 4  |

4. Conclusion

The decision tree algorithm C4.5 produced an accuracy of 76.4% on the dataset. To improve the classifier accuracy, two parameters: confidence factor and minimum number of objects were varied. By choosing the confidence factor as 0.2 and minimum number of objects as 4, the classifier accuracy was increased to 88.23%. The Bayes Net algorithm produces an accuracy of 94.1% by using simple estimator and K2 for searching. The authorship identification leads to an accuracy of 94.1% by choosing these two constraints. Thus by extracting general features that are common for all regional languages, an overall authorship identification system can be developed for all regional languages.

5. References

1. Iqbal F, Binsalleeh H, Benjamin CM. Fung, Debbabi M. Mining write prints from anonymous e-mails for forensic investigation. Digital Investigation. 2010; 7:56-64.
2. Meluch L, Tokarova I, Farkas P, Schindler F. Simple method based on complexity for authorship detection of text. MIPRO; 2016; 10(12).
3. Tokarova, Meluch L, Farkas P, Ruzicky E. Forensic author identification with help of compaction algorithm, (In Slovak). Slaboproudy obzor. 2015 Oct; 71(3):1-5.
4. Pratanwanich N, Lio P. Who wrote this? Textual modeling with authorship attribution in big data. IEEE International Conference on Data Mining Workshop; 2014.
5. Sanjana Anuradha. A computational framework for Tamil document classification using random kitchen sink. IEEE International Conference on Advances in Computing, Communications and Informatics (ICACCI); 2015.
6. Bhargava UK, Ramakrishnan AG, Mohammad S. Recognition of open vocabulary. IEEE Online Tamil Handwritten Pages in Tamil Script. 2014; 42(3):6-9.
7. Ramakrishnan AG, Arulmozhi. Language models in recognition applications: a new approach. Proceedings of the Tamil Internet Conference; Kuala Lumpur, Malaysia. 2013.
8. Khorji M, Iraqi Y, Jones A. An evaluation of authorship attribution using random forests. IEEE, International Conference on Information and Communication Technology Research (ICTRC’15); 2015.
9. Ramakrishnan AG, Urala B. Global and local features for recognition of online handwritten numerals and Tamil characters. Proceedings of International Workshop on Multilingual OCR (MOCR); 2013.
10. Sundaram S, Ramakrishnan AG. Performance enhancement of online handwritten Tamil symbol recognition with reevaluation techniques. Pattern Analysis and Applications. 2013; 1–23.
11. Ramalli R, Ramakrishnan AG. Fusion of complementary online and offline strategies for recognition of handwritten Kannada characters. Journal of Universal Computer Science. 2011; 17(1):81–93.