



Authorship Identification Techniques in Tamil Articles – A Survey

A.Pandian¹, Dr.Abdul Karim Sadiq²

¹Department of MCA, SRM University, Kattankulathur, Chennai, India

²College of Applied Sciences, Sohar, Ministry of Higher Education, Sultanate of Oman

Abstract: Authorship Identification is a way of determining who wrote a text when it is unclear who wrote it. It is useful when two or more people claim to have written something or when no one is willing (or able) to stay that he or she wrote the text makes use of all linguistic domains: semantics, syntax, lexicography, phonology (orthography) and morphology. Each of these domains is rule governed, yet, within these rules and among the components, the grammar offers the writer choices. In this paper we compared the various techniques used to identify the corresponding authors in Tamil Articles.

Keywords: Authorship, features extraction, tamil articles, techniques used.

I. Introduction

Text as an end product is an outcome of the particular choices taken by its author. This is why each specific text carries the fingerprints of its creator. While trying to find authorship, the following assumptions are arises, they are,

- there is a specific single author
- there are choices to be made
- the author is consistent in his/her preferred choices
- these choices are present and could be detected in all end products of that creator

Author Identification study is useful to identify the most plausible authors and to find evidences to support the conclusion.

Authorship analysis problem is categorized as follows,

- 1) Authorship identification (authorship attribution): It determines the likelihood of a piece of writing to be produced by a particular author by examining other writings by that author.
- 2) Authorship characterization: It summarizes the characteristics of an author and generates the author profile based on his/her writings like Gender, educational, cultural background, and writing style
- 3) Similarity detection: It compares multiple pieces of writing and determines whether they were produced by a single author without actually identifying the author like Plagiarism detection. To extract unique writing style from the number of online messages various features need to be considered are Lexical features, content-free features, Syntactic features, Structure features, Content-specific features

Although authorship attribution problem has been studied in the history but in the last few decades, authorship attribution of online messages has become a forthcoming research area as it is confluence of various research areas like machine learning, information Retrieval and Natural Language Processing. Initially this problem started as the most basic problem of author identification of anonymous texts, now has been grown for forensic analysis, electronic commerce etc. This extended version of author attribution problem has been defined as needle-in-a-haystack problem.

When an author writes they use certain words unconsciously and we should able to find some underlying pattern for an authors style. The fundamental assumption of authorship attribution is that each author has habit of using specific words that make their writing unique Extraction of features from text that distinguish one author from another includes use of some statistical or machine learning techniques.

II. Brief History

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 (1938) and Morton(1965) with the use of sentence lengths to judge authorship

A. Applications of Authorship Attribution

- To analyze anonymous or disputed documents/books, such as the ancient articles and poems written by various authors.
- Plagiarism detection - it can be used to establish whether claimed authorship is valid.
- Criminal Investigation – Un Authorized Emails

- iv) Forensic investigations - Verifying the authorship of spam mails, newsgroups messages, or identifying the source of a piece of intelligence.

B. *How can be Identify?*

- 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 from the given text that distinguish one author from another
- Applying some statistical or machine learning technique given training data
- Showing examples and counterexamples of an author's work

C. *Problems in Identification*

- ❖ Highly interdisciplinary area
 - Expertise in linguistics, statistics, text authentication, literature?
- ❖ Too many style measures have to apply
- ❖ Statistical method – complicated or so simple? Also too many exist in the literature as well
- ❖ Determine style markers.
- ❖ Parse all of the documents and extract the features
- ❖ Combine the results in order to get certain characteristics about the authors
- ❖ Apply each of the statistical/machine learning approaches to assign a given document to the most likely author.

Rest of the Paper is organized as follows. Section 2 Reviews existing techniques used for Authorship Analysis along with their classification. Section 3 explains basic procedure for authorship analysis. Section 4 summarizes Comparisons of various techniques since year 2004 till 2014. Section 5 Reviews performance evaluation parameters required for Authorship Analysis Techniques followed by section 6 which is conclusion.

III. State of the Art of Current Techniques

A. *Computerized Analysis*

This Analysis was developed in 1980’s, from the previous statistical Analysis of literary Style named, “Stylometry”. In this Stylometry, to quantify some of the features of an author’s Style. They are, Word or Sentence Length: This is a method, developed in the origin of Stylometry. But it is not a reliable method.

B. *Function Words*

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

C. *Vocabulary Distributions*

In this method, 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.

D. *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.

IV. Authorship Attribution Techniques

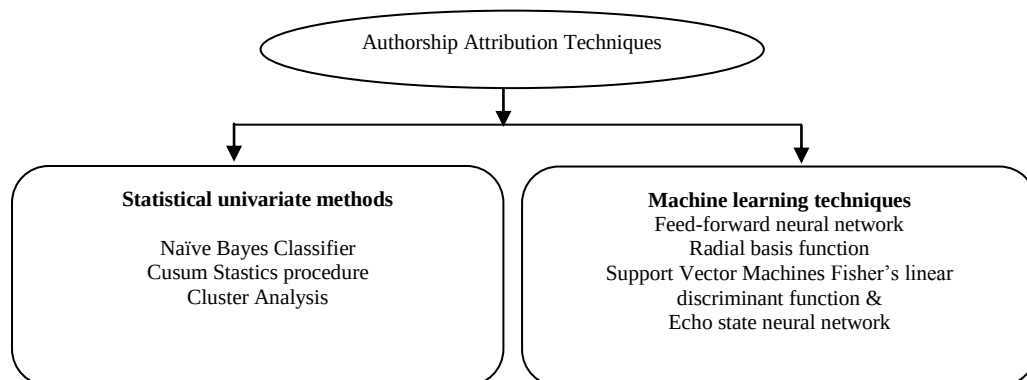


Figure 1: Authorship Attribution Techniques

V. Typical Procedure for Authorship Identification

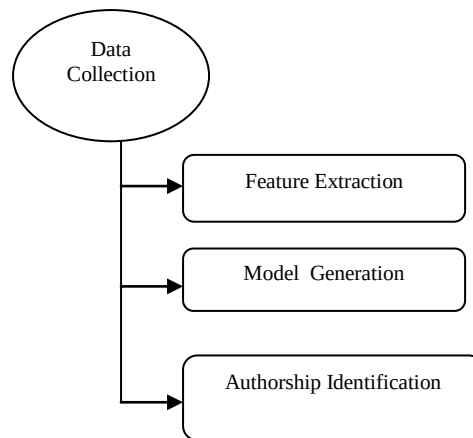


Figure 2: Typical Procedure for Authorship Identification

Step1: Data collection:-Collect Materials written by potential authors from various sources and Digitized.

Step2: Feature Extraction:-After extraction, each unstructured text is represented as a vector of writing-style features

Step3: 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

Step4: Author Identification:-Developed model can be used to predict the authorship.

This section summarizes the various techniques used for authorship identification research forum since 2004 to 2014. History of studies on authorship attribution problems presented in tabular format and year wise. For each method, we identify the corpus on which methods were tested, the feature types used and the categorization method used, size of Training set. Table 1 represented the comparative study of all authorship techniques.

VI. Conclusion and Future Implications

Now days many people use modern technology like computers, tablets, mobiles, in their regional languages. So, authorship attribution in each regional language has major impact. Similarly in many old articles, poems in their regional languages authors were not identified. For example, in tamil language, Agananuru, Purananuru, Mukkoodar pallu ,etc., many others were not identified. Hence, if we go in these directions it will be more helpful to the society to know the author of various valuable old articles and poems.

VII. References

- [1] Mappillairaju Bagavandas, G. Manimannan, (2004), Journées internationales d'Analyse statistique des Données Textuelles,
- [2] M. Bagavandas and G. Manimannan, (2006), A Study on Authorial Consistency and Attribution : A Canonical Discriminant Analysis Approach, Journées internationales d'Analyse statistique des Données Textuelles
- [3] M. Bagavandas & G. Manimannan, (2008), Style Consistency and Authorship Attribution: A Statistical Investigation, Journal of Quantitative Linguistics, Vol - 15, Issue 1, pages 100-110
- [4] M. Bagavandas & S. Nazreen Begum, (2008), An analysis of structure of an epic: A statistical approach, Journal of Quantitative Linguistics, Vol - 15, Issue - 3 pages 221-242
- [5] M. Bagavandas, Abdul Hameed & G. Manimannan, (2009), Neural Computation in Authorship Attribution: The Case of Selected Tamil Articles, Journal of Quantitative Linguistics, Vol - 16, Issue - 2 pages 115-131
- [6] Menaka S., Pattabhi R. K. Rao, Sobha Lalitha Devi, (2011), Automatic Identification of Cause-Effect Relations in Tamil Using CRFs, Computational Linguistics and Intelligent Text Processing Lecture Notes in Computer Science Volume 6608, 2011, pp 316-327
- [7] Chandrasekaran, R. and Manimannan, G. (2013), Use of Generalized Neural Networking Authorship Attribution, International Journal of Computer Application, Volume 6, Number 4, ISSN 0975-8887, New York.
- [8] Subasree Venkatsubramaniyen, Subha Rashmi and Rajeswari Sridhar, (2013), Classification of Tamil Poetry using context free grammar using tamil grammar rules, ccsl, sipp, aisc, pdcta, pp. 159-167
- [9] Arbind Kumar Singh and G. Manimannan, (2013), Literary Analysis using CUSUM Technique on Bharathiar Writings, IOSR Journal of Mathematics, Volume 8, Issue 4, PP 42-50, e-ISSN: 2278-5728, p-ISSN: 2319-765X
- [10] A. Pandian, Md. Abdul Karim Sadiq, (2013), Authorship Attribution in Tamil Language Email For Forensic Analysis, International Review on Computers and Software, Vol. 8 N. 12, ISSN: 1828-6003
- [11] R. Lakshmi Priya and G. Manimannan (2014), A Study of Ambiguous Authorship in Tamil Articles using Multivariate Statistical Analysis, International Journal of Computer Applications Volume 86 – No , ISSN 0975-8887.

Table 1: Comparison of Techniques

S.No	Authors / Year	Features	Techniques Used	Corpus	No. of Authors	Used Training Set
1	Mappillairaju Bagavandas, & G. Manimannan 2004	18 stylistic Features	ANOVA technique, two sample t-statistic and Factor analysis	Tamil Language Magazine “Bharathi Dharisanam” 1975	3	32 articles
2	M. Bagavandas & G. Manimannan 2006	18 linguistic variables, 11 morphological variables, 4 habitual words and 3 function words	Canonical discriminant analysis	Tamil Language Magazine “India” 1906	3	55 articles
3	M. Bagavandas & G. Manimannan <i>Jan 2008</i>	24 commonly occurring function words	canonical discriminant analysis	Tamil Language Magazine	3	32 attributed and 23 disputed articles
4	M. Bagavandas & S. Nazreen Begum <i>Jun 2008</i>	frequencies of occurrence of characters	Statistical methods	Cilappathikaaram	1	30 chapters of chilappathikaaram
5	M. Bagavandas, Abdul Hameed & G. Manimannan <i>Apr 2009</i>	24 function words	Learning vector quantization	Tamil Language Articles	3	32 articles
6	Menaka S., Pattabhi R. K. Rao, & Sobha Lalitha Devi <i>2011</i>	cause-effect relations in a given Tamil text.	Conditional Random Fields	8648 sentences of Tamil Language	Many	-
7	R. Chandrasekaran & G. Manimannan <i>January 2013</i>	24 function words and 18 morphological variables	ANN	Tamil Language Magazine “India” 1906	3	32 Attributed and 23 un-attributed articles
8	Subasree Venkatsubhramaniyen, Subha Rashmi & Rajeswari Sridhar <i>2013</i>	Venpaa, Asiriyapaa, Vanchipaaof Tamil Poems	CFG	Various Ancient Poems	Many Authors	Various Poems of Ancient Authors
9	Arbind Kumar Singh & G. Manimannan <i>Sep. - Oct. 2013</i>	(i) use of 2, 3 and 4 letter words, (ii) words starting with a vowel (iii) combination of these two together	CUSUM analysis	Tamil Language Magazine “India” 1906	Many Authors	13 Articles
10	A.Pandian & Md. Abdul Karim Sadiq <i>December 2013</i>	322 features of Emails	Fisher’s linear discriminant function, Radial basis function & Echo state neural network	Emails of 50 Authors	50 Authors	500 Emails
11	R. Lakshmi Priya & G. Manimannan <i>January 2014</i>	morphological and function words	Principal Component Analysis (PCA) and Multivariate Discriminant Analysis (MDA)	Tamil Language Magazine “India” 1906	3	92 Articles