A Survey on Authorship Analysis

P. Vijayapal Reddy
Department of Computer Science and Engineering
Matrusri Engineering College
Hyderabad, India

Abstract— The paper discusses about the problem of Authorship analysis, different types of authorship analysis's such as authorship attribution, authorship identification, authorship profiling, plagiarism detection. It also addresses the issues in Indian language text.

Keywords— Authorship attribution, authorship profiling, plagiarism detection, text classification.

1. INTRODUCTION

Authorship Attribution is a kind of text classification (TC) problem but it is different from categorization. AA is different from text classification because the writing style is also important in AA apart from the text content which is the only factor used in text classification. The features in TC are deterministic where as in AA not deterministic. Based on the size of the data set and number of authors, classifiers and feature sets may behaves differently in AA [2]. Hence these differences make AA task more challenging compared with TC. In text classification the texts are assigned to one or more predefined classes based on the categories where as in AA the texts are assigned to one or more predefined classes based on the author set [3].

Authorship Attribution can be defined in three ways. Firstly, for a given test document, find the author of the text from the defined set of authors. Secondly, for a given test document, believed to be written by one author from a set of authors then find which one, if any. Thirdly, for a given test document, who is the author. There are two flavours of AA tasks: closed-class and open- class. The first definition is a closed class problem whereas second and third definitions are open classes problems. In closed class problem the author to be identified is one from the given set of authors where as in open set problems the author to be identified may or may not in the defined author set.

2. LITERATURE SURVEY

Authorship Attribution can be viewed as one of the oldest problem and one of the newest research problem in the field of Information Retrieval. Stylometry is the statistical analysis of literary style. The main assumption behind stylometry is that the authors make certain subconscious and conscious choices in their writing. Some of the features that were used in stylometry include average sentence length, average syllables per word, average word length, distribution of parts of speech, function word usage, the Type-Token ratio, Simpson's Index, Yule's Characteristic K, entropy, word frequencies, and vocabulary distributions [4]. Some models that were used in stylometry include n-grams [8], feature counts, inductive rule learning, Bayesian networks, radial basis function networks, decision trees, nearest neighbour classification, and support vector machines [5]. Mosteller & Wallace [6] propose to select

semi-automatically the most common terms composed mainly by various function words for AA. The earliest studies of AA were reported by [9] and Yule [10], in which statistical methods were used limit data, not only the size of the experimental corpus but also the size of feature set. Yang [9] graphically represented the word-length as characteristic curves, and he also in [10] used sentence length to differentiate between authors text.

Grammatical-based or syntax-based features in AA, which were applied by several researchers [7, 13]. Chi-square ($\chi 2$) measure is often used to determine relevant features in authorship attribution [14, 12]. The cumulative sum technique [15] looks at the frequencies of a range of possible habits in use of language. Principal Component Analysis (PCA) [9, 8, 10], Markov chains [7], and compression based techniques [16] are typical of computational approaches that were proposed for authorship attribution (AA). N-grams are widely used in authorship attribution [17,11]. Juola in [18] proposed a similar approach that were applied to AA, in which the unigram model on the character level was used . Benedettoin [19] used compression approach to different applications including AA . Machine Learning approaches were applied to AA in recent years, including Neural Networks [26], Bayesian classifiers , SVMs , and decision trees.

In general, applications of AA include resolving historical questions of unclear or disputed authorship. In recent years, practical applications for author identification have grown in areas such as intelligence, criminal law, civil law, and computer security. AA has a long history with multiple application areas that include spam filtering [1], cyber bullying, plagiarism detection [2], author recognition of a given program [3], and web information management [2]. In forensic investigations where verifying the authorship of e- mails and newsgroup messages, or identifying the source of a piece of intelligence also considered as an AA applications.

3. NEED OF STUDY

India is the home of different languages, due to its cultural and geographical diversity. The official and regional languages of India play an important role in communication among the people living in the country. In the Constitution of India, a provision is made for each of the Indian states to choose their own official language for communicating at the state level for official purpose. In the eighth schedule as of May 2008, there are 22 official languages in India.

The availability of constantly increasing amount of textual data of various Indian regional languages in electronic form has accelerated. So the Authorship identification of text documents based on languages is essential. The objective of the work is the representation and identification of Indian language text documents using text mining techniques.

South Indian language corpus such as Kannada, Tamil and Telugu language corpus, has been created. Several text mining techniques such as naive Bayes classifier, k-Nearest-Neighbor classifier and decision tree for Authorship Attribution for various languages have been used. There is no work done in Authorship attribution in Indian languages. Authorship Attribution in Indian languages is challenging as Indian languages are very rich in morphology, giving rise to a very large number of word forms and hence very large feature spaces.

4. PHASES IN AUTHORSHIP ATTRIBUTION MODEL

For the development of a model for authorship analysis for Indian languages text, the following methods need to be followed in the below order.

1. **Data collection:** There is no standard data corpus is available for local languages such as Telugu, Tamil, Kannda, Hindi for the task of Authorship Analysis. The required data sets need to collect from various sources like news papers, articles.

2. **Pre processing stage:** In this stage the raw data collected need to be pre-processed using different phases like normalization, segmentation, stemming, tagging and stop word identification and stop word elimination.

Tokenization is the process of chopping a document into small units called tokens which usually results in a set of atomic words having a useful semantic meaning. This phase outputs the article as a set of words by removing the unnecessary symbols like semicolons, colons, exclamation marks, hyphens, bullets, parenthesis, numbers etc.

As in [5] a stop list is a list of commonly repeated features which appear in every text document. The common features such as pronouns, conjunctions and prepositions need to be removed because they do not have effect on the classification process. For the same reason, if the feature is a special character or a number then that feature should be removed.

Stemming is the process of removing affixes (prefixes and suffixes) from features as in [6]. This process is used to reduce the number of features in the feature space and improve the performance of the classifier when the different forms of features are stemmed into a single feature.

3. **Feature Extraction:** The best suitable subset of the features for authorship analysis, various features such as lexical features, syntactic features and structural features and their combinations need to test.

4. **Feature Selection:** The aim of feature selection methods is to reduce the dimensionality of dataset by removing irrelevant features for the classification task. As in [7], some types of features, such as character and lexical features can considerably increase the dimensionality of the features' set. In such case, feature selection methods can be used to reduce such dimensionality of the representation. Features which are not positively influencing the TC process is removed without affecting the classifier performance, known as Dimensionality reduction (DR).

Feature selection deals with several measures such as document frequency, DIA association factor, chi-square, information gain, mutual information, odds ratio, relevancy score, GSS coefficient. These methods are applied to reduce the size of the full feature set. DR by feature extraction is to create a small set of artificial features from original feature set, which can be

done using Term clustering and Latent semantic indexing. In Indian languages, the number of features are be even higher compared with English text because of richness in morphology.

5. **Classification:** The goal of Machine Learning (ML) is to construct programs that automatically learn from the training dataset. ML algorithms are able to discover rules from training examples [12]. There are two types of machine learning algorithms named as eager learning and lazy learning algorithms. The k-Nearest Neighbour algorithm is an example of a Lazy Learning algorithm. All other learning algorithms which are considered as eager learning algorithms [6], to identify the most suitable machine learning approaches for local data sets for authorship analysis.

6. **Author identification:** For a given test document the name of the author will be returned. For this purpose four steps need to perform. These steps are same steps that are performed on the training data set. Data preprocessing is performed which involves tokenizing, stopword removal and stemming of the input test document, feature extraction is performed after that reduce the dimensionality of the feature set then input the classifier with reduced feature set of the test document.

7. **Data Compression:** An alternative procedure for the task of Authorship Analysis is usage of data compression techniques. Various data compression techniques and various compression distance measures and their performance on Authorship analysis for the given text set need to be measured using various measures such as F1 measure, accuracy, macro F1 measure and micro F1 measure.

5. CONCLUSIONS

Still today authorship attribution to Indian language-based text is not attempted by the researchers. The influence of attribution methods and features which are applied to text in various languages may not be suitable for Indian language- based text for adaptation. Hence, in the proposed work, the existing statistical approaches, machine learning techniques, data compression techniques and various features such as lexical, syntactic and structural features which are thoroughly tested on text of various languages need to be tested for its most likely adaptability for text in the Indian languages.

Developing such a system for the text developed in Indian languages will be useful for identification of the source of the text, the writer of a piece of the text, age, area, gender and to which era that particular author belongs, detecting the plagiarized content from the text, identification of the owner ship of the legal documents.

REFERENCES

[1] Sebastiani, F. (2002) Machine Learning in Automated Text Categorization. ACM Computing Surveys, 34(1), pp. 1-47.

[2] Zheng, R., Li, J., Chen, H. & Huang, Z. (2006). A Framework for Authorship Identification

of Online Messages: Writing-Style Features and Classification Techniques. Journal of the American Society for Information Science and Technology, 57(3): 378-393.

[3] Chaski, C. E. (2007). The Keyboard Dilemma and Authorship Identification. In P. Craiger & S. Shenoi (Eds.), Advances in Digital Forensics III (pp. 133-146). Ney York, NY: Springer.

[4] Juola, P. (2008). Authoship Attribution. Hanover, MA: Now Publishers.

[5] Koppel , M., Schler, J., & Argamon, S. (2009). Computational Methods in Authorship Attribution. Journal of the American Society for Information Science and Technology, 60(1), 9-26.

[6] Koppel, M., Schler, J., & Messeri, E. (2008). Authorship Attribution in Law Enforcement

Scenarios. In C.S. Gal, P. Kantor, & B. Saphira (Eds.), Security Informatics and Terrorism: Patrolling the Web (pp.111-119). Amsterdam: IOS.

[7]H. Baayen, H. V. Halteren, A. Neijt, and F. Tweedie. An experiment in authorship attribution. In Proceedings 6th International Conference on the Statistical Analysis of Textual Data, pages 29–37, 2002.

[8]D. I. Holmes, M. Robertson, and R. Paez. Stephen Crane and the New York Tribune: A case study in traditional and non-traditional authorship attribution. Computers and the Humanities, 35(3):315–331, 2001.

[9]P. Juola and H. Baayen. A controlled-corpus experiment in authorship identification by cross- entropy. Literary and Linguistic Computing, 20:59– 67, 2003.

[10]J. Burrows. Delta: A measure of stylistic difference and a guide to likely authorship. Literary and Linguistic Computing, 17:267–287, 2002.

[11]E. Stamatatos, N. Fakotakis, and G. Kokkinakis. Automatic authorship attribution. In Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics, pages 158–164, Bergen, Norway, 1999. Association for Computational Linguistics.

[12]E. Stamatatos, N. Fakotakis, and G. Kokkinakis. Computer-based authorship attribution without lexical measures. Computers and the Humanities, 35(2):193–214, 2001.

[13] O. V. Kukushkina, A. A. Polikarpov, and D. V. Khmelev. Using literal and grammatical statistics for authorship attribution. Problem of Information Transmission, 37(2):172–184, 2001.

[14]Y. M. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In Proceedings of the 14th ICML International Conference on Machine Learning, pages 412–420, Tennessee, USA, 1997. Morgan Kaufmann Publishers.

[15] J. M. Farringdon. Analysing for Authorship: A Guide to the Cusum Technique. University of Wales Press, 1996.

[16] D. Benedetto, E. Caglioti, and V. Loreto. Language trees and zipping. The American Physical Society, 88(4):048702, 2002.

[17] D. Jurafsky and J. H. Martin. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Prentice Hall, 2000.

[18]P. Juola. What can we do with small corpora? Document categorization via cross-entropy. In Proceedings of the Interdisciplinary Workshop on Similarity and Categorization, Edinburgh, UK, 1997.

[19] D. Benedetto, E. Caglioti, and V. Loreto. Language trees and zipping. The American Physical Society, 88(4):048702, 2002