Name: SURESH KUMAR.M
Reg.No: 2101305001
Title: AUTHORSHIP ANALYSIS

INTRODUCTION

The main goal of most Natural Language Processing (NLP) applications is to automatically extract meaning from text. Text Mining is an essential branch of this process. It denotes the tasks that try to extract useful information by doing a linguistic analysis of large quantities of text and by detecting usage patterns. An important step in the Text Mining domain is Text Categorization also called Document Classification [1]. Applications of Text Categorization are numerous. The most important ones are document indexing and filtering, word sense disambiguation, the hierarchical categorization of web pages and web search engines [1]. Authorship Analysis (AA), like Language Identification, is a classification problem.

Authorship analysis will be used as a broad term that covers a number of related forensic linguistics tasks aimed at making inferences about the author of an anonymous piece of text. Authorship Analysis tasks can be broadly categorized into four ways. They are authorship attribution (AA), author profiling, author identification and clustering. AA can be termed as

author identification as in [2, 3], closed-class task as in [4], categorization task as in [5], needle-in-a-haystack problem [6] and vanilla authorship attribution [5].

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 flavors 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.

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 neighbor 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, and 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. Benedetto in [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 [20], SVMs [21], and decision trees [22].

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 application.

NEED FOR THE 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.

Objectives of the Study

There are few web sites storing texts in Indian languages, such as Indian language newspapers and magazines. Indian languages like many other languages in the world have little content on the Internet when compared with English. To make the matters worse, these languages are usually linguistically more sophisticated and are high entropy languages when compared to languages like English, contributing to the information access problem. The problem facing the research community in the western countries working on English and other languages were trying to deal with the information overload. Whereas, in the case of Indian language queries, the problem is of retrieving relevant articles are relatively fewer when compared with articles in English language.

Language based corpus and statistical approaches were well established elsewhere in the world, India is still lagging far behind. Inadequate plain text corpora is available and annotated corpora is hardly found in Indian languages. Around three million Small plain text corpora was developed in late Eighties on major Indian Languages with the initiatives of the TDIL (Technology Development in Indian Languages) group of Department of Electronics, Government of India carried out basic statistical analysis of these corpora.

One of the major issues related to Indic scripts is linked with the heuristic grammar rules on individual words in the representative form of pluralities, present and past tenses etc. The resultant is found to be multiplied in vocabulary. The basic philosophy of language representation reflecting the phonetic sequences is the basis for this complexity. More amount of complexity is involved when a large set of grammar rules are applied to combine two or three

words formulating into a single word. These formulations can neither be treated as phrases nor individual words.

Authorship attribution depends on three basic factors. Good knowledge in linguistics gives better choice of selecting style markers for author stylistic analysis. Statistics are the good measures to quantify these style markers. Classification techniques are used to discriminate the obtained style markers correctly among the authors in the author set.

The research on authorship attribution is evaluated on multiple data sets with various corpus sizes makes it difficult to compare the methods used in different approaches. The attribution methods which were successful for one type of data set may not be successful for other data sets. Hence it is not clear whether these methods are scalable, reliably effective, or robust. The problem of authorship attribution is not yet attempted on the text of Indian languages. There are many research papers published on authorship attribution for various languages like English, Arabic, Dutch, Chinese and Greek with various corpus sizes. Many features such lexical, syntactic and structural features and their combinations are also experimented for feature extraction for various languages text. The features which are more suitable for one language text may not be suitable for other languages text. Feature selection measures are also explored for dimensionality reduction of the training and test set. Various classifiers and their combinations are also tested for different texts developed on various languages.

Data compression techniques are also thoroughly tested for authorship attribution on various languages text. It is required to identify the influence of machine learning techniques on Authorship attribution for Indian languages text. It is required to evaluate the influence of various features such as lexical, syntactic and structural features on the text developed for Indian languages. There is a need to study the impact of data compression techniques and data compression distance measures.

Based on the above discussions, it is required to address the various problems in terms of feature extraction, feature selection, classification and also the influence language characteristics on the various concepts. The problems need to addressed are mentioned below.

1. Identify the influence of various classifiers on Authorship Analysis for Indian languages text
2. Identify the linguistic information specific to each language to obtain the best results on authorship analysis task for the given text
3. Identify the influence of number of subjects, number of documents in each subject and the number of potential authors on Authorship Analysis.
4. Identify the influence of lexical, syntactic and structural features and their combination.
5. Identify the influence of Data Compression Techniques for Authorship Analysis on Indian languages text.

RESEARCH METHODOLOGY

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, Kannada, Hindi for the task of Authorship Analysis. The required data sets need to collect from various sources like newspapers, articles.

2. **Preprocessing stage:** In this stage the raw data collected need to be preprocessed 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 [23, 24, and 25] 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 [26]. 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 [27, 28], 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, and 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 [26]. There are two types of machine learning algorithms named as eager learning and lazy learning algorithms. The k-Nearest Neighbor algorithm is an example of a Lazy Learning algorithm. All other learning algorithms which are considered as eager learning algorithms [26], 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, stop word 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.

PERIOD OF STUDY

Collection of related literature and corpus creation takes around one year. Collection related tools for preprocessing such morphological analyzers, POS taggers and preprocessing takes for a period of six months. The publication of related papers takes around two years. The thesis write-up takes six months.
Around four years are required to complete the task.

IMPLICATIONS OF THE STUDY

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.

CHAPTER SCHEMES

The proposed work can be organized in eight chapters. Chapter 1 deals with the problem of authorship analysis and its significance in Indian context. The concept of text mining with respect to authorship analysis. Various features of text classification and data compression techniques towards authorship analysis need to be discussed briefly. The applications, motivation to attempt the problem and organization of thesis need to present.

In chapter 2, the detailed discuss is presented on text categorization techniques, authorship analysis techniques, features, classification techniques such as machine learning techniques, statistical techniques and data compression techniques. In chapter 3, the characteristics of Indian languages, the process of data collection, flow of preprocessing of a text such as normalization, stemming, tagging and stop word elimination need to be described with a suitable example.

Chapter 4 describes about influence of different machine learning techniques such as decision trees, K-nearest neighbors, and Naive Bayes and Support vector machines on Authorship attribution. The different evaluation measures are also required to address. Influence of number of authors in the data set, influence of number of documents in the data set and feature extraction and feature selection measure such chi-square also need to address in chapter 5.

In chapter 6 Influence of lexical, syntactic and structural features and their combination on authorship attribution for Indian languages text need to empirically evaluate. The evaluation measures such precision, recall, F1 measure and accuracy to measure the impact of features on

authorship attribution. Chapter 7 explains about the influence of Data Compression Techniques on authorship attribution the text, various compression techniques, various compression distance measures and micro and macro F1 measures need to evaluate and present. Finally, Chapter 8 gives salient features of the work, the conclusions of the proposed work from the results obtained from the previous chapters and also possible further extensions.

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). Authorship 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, Nashville, 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.
[20] P. Domingos and M. J. Pazzani. On the optimality of the simple bayesian classifier under zerone loss. Machine Learning, 29(2/3):103–130, 1997.
[21] E. Gabrilovich and S. Markovitch. Text categorization with many redundant features: Using aggressive feature selection to make SVMs competitive with C4.5. In Proceedings of the 21st ICML International Conference on Machine Learning, pages 321–328, Banff, Alberta, Canada, 2004. ACM Press.
[22] 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.
[23] Stamatatos, E.: Author Identification: Using Text Sampling to Handle the Class Imbalance Problem. Information Processing and Management, 44(2), pp. 790--799 (2008).
[24] B.Vishnu Vardhan,P.Vijaypal Reddy,A.Govardhan"Analysis of BMW model for title word selection on Indic scripts", International Journal of Computer Application (IJCA) Vol 18 Number 8 March 2011 pp 21-25
[25] B.Vishnu Vardhan,P.Vijaypal Reddy, A.Govardhan"Corpus based Extractive summarization for
Indic script", International Conference on Asian Language Processing (IALP) IEEE Computer Society (IALP 2011) pp 154-157
[26] P. Vijay pal Reddy, Vishnu Murthy.G, Dr. B. Vishnu Vardhan, K. Sarangam "A comparative study on term weighting methods for automated Indian language-based text categorization with effective classifiers" International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.6, November 2013
[27] B. Vishnu Vardhan ,B. Padmaja Rani, A. Kanaka Durga, A.Govardhan, L. Pratap Reddy, A. Vinay
Babu, Impact of dimensionality reduction on the categorization of phonetic based language documents- A case study on Telugu" Geetham journal of Information and Communication Volume1 , Issue 1, July-December 2008, pp 93-98
[28] B. Vishnu vardhan,P.Vijayapal reddy, B Sasidhar, B Harinatha reddy,L. Pratap reddy , A. Govardhan, " Approaches of Dimensionality Reduction for Telugu Document Classification" International Conference on Asian Language Processing (IALP) IEEE Computer Society 2009, pp 259-264.