Author Identification using word N Grams – A case study on Telugu Text

P.Vijayapal Reddy
Department of Computer Science and Engineering
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
vijayapalreddy76@gmail.com

Authorship Identification (AI) or Authorship Attribution (AA) deals with identify the author of an anonymous text from known author set. The AA problem can be viewed as a classification problem. The different steps involved in AA are data preprocessing for vector representation of the text, feature extraction for quantitative representation of the text, feature selection is to reduce the dimensionality feature space, classification algorithms for pattern generation and finally author identification for the given unknown document. There are four categories of features such as lexical, character, syntactic, and semantic features. In this paper word N gram features are considered for feature extraction in combination with various classifiers to learn the training document set and to identify the author of a unknown text. The performance of the method is evaluated with precision, recall and F1 measures on Telugu Texts.

Keywords : Authorship Identification, word N grams, Naive Bayes classifier, Support Vector machine classifier.

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 behave differently in AA [1]. 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 [2]. Thus the texts in AA are categorized into different classes based on the given set of authors.

From the last decade, research on authorship identification is extensively explored. In the beginning authorship attribution was manually conducted by observing the linguistic information embedded in a text corpus as there is no sophisticated natural language processing tools available. These techniques were based on linguistic markers like term frequency and word similarity [3]. Most of the proposed methods are centered on the detection of authorship for literary texts. Well known study in the field of authorship attribution is the identification of an author of Federalist papers where there was a dispute about twelve of the authors [4]. Another study to identify the authorship of Shakespeare‟s plays in question [5].

In this paper the main focus is on identifying the author of a given text using various steps. The various steps are data preprocessing, feature extraction, feature selection, classification and author identification. Data preprocessing contains text tokenizing, stopword removal and text stemming. Feature extraction involves the process of extracting various features such as lexical, syntactic, structural and character level features. In this paper word level features are considered for feature extraction. Classification algorithms are used to make generalizations and discover rules from feature set. Classifiers such as Naive Bayes, Support vector machine, K nearest neighbour and decision tree are used for pattern generation from the training author sets. Performance of different classifiers in combination with different features are evaluated on Telugu data set.

2. Related Work

In [6] and [7], researchers used a variety of statistical methods to identify characteristics which are not variant for a given author but which varies from author to author. In [8] the researcher working on Federalist Papers identified a set of functional words which are less frequent could serve as best features to identify authorial style. Yule in [7], [9] and [10] identified that complexity-based features such as sentence length, word length, type and token ratio, automated parsing, POS tagging and POS n-grams are useful features in authorship attribution. Peng in [11] modeled each author by a vector of most frequent n-grams in the text. Fung in [12], used SVM to determine the authors of Federalist papers. Stamatatos in [13] used Multiple Regression classifier in combination with syntactic style features. A character n-gram based method of author attribution has been proposed by Keselj [11]. N-gram models have been successfully applied in speech recognition [14], natural language processing [15] and spell correction[16]. It is also successfully applied in author attribution [17].

In latest research of AA, there are number of evaluations are performed on short texts in various languages. Short text authorship attribution is challenging compared with data set having long texts. Short texts requires reliable representation of such texts and Machine Learning (ML) algorithm that can handle with limited data. In [18], it is reported that the samples of texts should be long enough therefore the text representation features can sufficiently represent their style. In [19] showed that reducing the length of the training samples has a direct impact on performance. Some studies have shown promising results with short texts of 500 characters [20] or 500 words [21]. In [22] stated that the longer is the text; the better is the identification.

In this paper the data set contains 300 different texts written by 12 authors. This paper focuses on employment various classifiers such as Naive Bayes classifier, KNN, SVM and DT in the classification task. In contrast to NB, this study also implements SVM which is more suited to extremely big datasets [23]. In this paper, we focus an author attribution of Telugu texts by extracting various features and applying different classifiers. The rest of the paper is organized as section 3 discusses about the AA model which contains various steps such as data preprocessing, feature extraction, feature selection, classification and author identification. The characteristics of Telugu language is explained in the section 4. The experimental evaluations and discussions on the results are presented in section 5. Section 6 deals with the conclusions drawn from the results and also it contains the possible extensions to the proposed work.

3. Proposed Model

In this model various steps are data preprocessing, feature extraction, feature selection, classification and author identification. The data set is separated into training and testing set. In the first phase, features are extracted from the data, after that training and test instances are created, on the basis of these features. In the second phase, an classification model is built from training data, so as to be tested on unknown test data. The training and test instances are numerical feature vectors which represent term frequency of every selected feature, followed by the author number. Labeled training data are used to train a machine learner, as it allows the evaluation of classification. The task of AA is conducted as multi class AA.

Data preprocessing is a very important step in authorship attribution. Text documents in their original form are not suitable for meaning patterns generation. They must be converted into a suitable input format. It can be converted into a vector space since most of the learning algorithms use the attribute, value representation. This step is important for the next stages. Data preprocessing involves tokenization, stopword removal and stemming. 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.

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. Stop word list is identified using Parts of speech (POS) tagging done Telugu morphological Analyser (TMA). Stemming is the process of removing affixes (prefixes and suffixes) from features. 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. By using the tool Telugu morphological analyser (TMA) stem forms of the inflected words are identified.

There are four main kinds of features that contains authorial impressions for authorship. The are lexical, character, syntactic, and semantic features. In this paper, empirical evaluations are carried using lexical and character features because they are more reliable than semantic features. The different character level features considered in this paper are character unigram, bigram, trigram and tetragrams. Character unigram takes individual characters as tokens where as character bigram considers two consecutive characters, trigram consider three consecutive and tetragram considers four consecutive characters as features. Character ngrams demonstrated that they are able to handle limited data [19] effectively. Lexical ngrams features are the most widely used kind of features. Whereas researches done in [25] and [19] stated that lexical features are good for small datasets. Word unigram, word bigram, word trigram and word tetragram features are used as lexical features in this experiments. Word unigram takes single word as a feature, word bigram takes two consecutive words, word trigram takes three consecutive words and word tetragram considers four consecutive words as a feature.

4. Author identification

In this step, for a given test document the name of the author will be returned. For this purpose four steps are performed. 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.

5. Results and Discussion

For authorship identification system, the dataset is collected from Telugu news papers. The topics are ranges from editorials, business and sports. The dataset cotains 300 news articles written by 12 authors. The average number of words is 547 per document. In our experiments, we have separated our dataset into two groups: testing and training data. The training set contains 20 different texts for each one of 12 different authors. On the other hand, for the test set there are 5 different texts for each one of 12 different authors. The training data set is used to create a data patterns of each author and was treated as data with a known author. We extracted the same profile from the testing data and this data was treated as data with unknown author. Taking the profiles of each testing document one by one, we compared them with each of the training profiles belonging to each author to identify author of unknown test document.

5.1. Evaluation measures

In order to compare the results of all possible features with classifiers, we computed the precision, recall and F1 measure. Precision is the proportion of examples labeled positive by the system that were truly positive, and recall is the proportion of truly positive examples that were labeled positive by the system. where F1 is computed based on the following equation:

$$F_1 = \frac{2 * Recall * Precision}{Recall + Precision}$$

where,

$$Precision = \frac{X}{X+Y}$$

$$Recall = \frac{X}{X+Z}$$

where X is documents assigned and correct, Y is documents assigned but not current and Z is documents not assigned but correct. The precision, recall and F1 values obtained for different classifiers in combination with different word level features are presented in the below tables.

S.No	**Feature**	**F_1 value**			
		NB	KNN	SVM	DT
1	Word uni-gram	0.76	0.81	**0.83**	0.70
2	Word Bi-gram	0.68	0.72	0.75	0.62
3	Word Tri-gram	0.58	0.64	0.67	0.54

4	Word Tetra-gram	0.54	0.61	0.63	0.49

Table 1:The F_1 measure for various word N gram features

Precision, recall and F1 values are calculated for various character level features and lexical features on different classifiers. From the values obtained it can be concluded that character trigram feature is performing well out of all character level features. From the view of classifiers, SVM performance is good compared with all other classifiers. Word unigram is identified as a best feature for authorship attribution when compared with all other character level features and word level features. Compared with word level features, on an average character level features can be considered as good indicators for authorship identification.

6. Conclusions & Future Scope

In this work, an AA task has been experimented on an Telugu dataset In this work different texts of various authors are selected. These texts are preprocessed. Several features have been tested for Telugu dataset. In our experiments tried to find the best authorship attribution conditions using character level features and lexical features. Empirical evaluations are carried on the text set using different machine learning classifiers such as naive bayes, k-nearest neighbour, support vector machine and decision tree classifiers in combination with different character and word level features. From the results it can be concluded that the character level features are better than the word based features. The word unigram feature gave the best F1 score obtained as 0.83 for authorship classification. The SVM classifier shows good performance in this experiment of AA compared with all other classifiers. This work of AA, which is first work done on Telugu text.

As a part of future work we may experiment with other machine learning algorithms with more number of authors and with small dataset of texts. It is also possible to experiment with other types of features such as syntactic and semantic and also with the combination of different types of features.

References

[1] B. Allison and L. Guthrie, " Authorship attribution of e-mail: comparing classifiers over a new corpus of evaluation," in Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), May 28-30, 2008, Marrakech, Morocco.

[2] Klarreich, E. 2003. Bookish math. Science News 164(25)

[3] T. Merriam, "Heterogeneous authorship in early Shakespeare and the problem of Henry V," Literary and Linguistic Computing, vol. 13, no. 1, 1998, pp. 15-28

[4] Bozkurt, D., Baglıoglu, O., & Uyar, E. (2007), “Authorship Attribution: Performance of Various Features and Classification Methods” Computer and Information Sciences.

[5] Zhao, Y. (2007), “Effective authorship attribution in Large Document Collections”, PhD Thesis, School of Computer Science and Information Technology, RMIT University , Melbourne, Victoria,

Australia.

[6] Holmes, D.: The Evolution of Stylometry in Humanities Scholarship Literary and Linguistic Computing, (1998) 13, 3, 111-117

[7] Koppel, M., Schler, J.: Exploiting Stylistic Idiosyncraises for Authorship Attribution, IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis, Acapulco, Mexico (2003)

[8] Mosteller, F., Wallace, D.L.: Inference and Disputed Authorship: The Federalist Reading, MA:Addison-Wesley (1964)

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

[10] Stamatatos, E., Fakotakis, N., Kokkinakis, G.: Computer-Based Authorship Attribution without Lexical Measures, Computers and the Humanities (2001) 193-214

[11] Peng, F., Schuurmans, D., Keselj, V., Wang, S.: Language Independent Authorship Attribution using Character Level Language Models, 10th Conference of the European Chapter of the Association for Computational Linguistics, Budapest (2003) 267-274

[12] Fung, G., Mangasarian, O.:The Disputed Federalist Papers: SVM Feature Selection via Concave Minimization, Proceedings of the 2003 Conference of Diversity in Computing, Atlanta, Georgia, USA (2003) 42-46

[13] Stamatatos, E., Fakotakis, N., Kokkinakis, G.: Automatic Authorship Attribution, Nineth Conf. European Chap. Assoc. Computational Linguistics, Bergen, Norway (1999)

[14] F. Jelinek, Statistical Methods for Speech Recognition, MIT Press, Boston, Massachusetts, USA, ISBN: 0-262-10066-5, 1998

[15] Foundations of Statistical Natural Language Processing Christopher D. Manning and Hinrich Sch¨ utze (Stanford University and Xerox Palo Alto Research Center) Cambridge, MA: The MIT Press, 1999, xxxvii + 680 pp; hardbound, ISBN 0-262-13360-1

[16] Mayes, E., F. Damerau, et al. (1991). “Context Based Spelling Correction.” Information Processing and Management 27(5): 517-522.

[17] Abbasi, A. and Chen, H. 2008. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace.

[18] Stamatatos, E. (2009), “A survey of Modern authorship attribution methods”, Journal of the American Society for Information Science and Technology, 538-556.

[19] Luyckx, K. (2010), “Scalability Issues in Authorship Attribution”, PhD Thesis, Faculty of Arts and Philosophy, Dutch UPA University.

[20] Sanderson,C. & Guenter, S. (2006), “Short text authorship Attribution via Sequence Kernels, Markov Chains and Author Unmasking”, An Investigation. Proceeding of 2006 Conference on

Empirical Methods in Natural Language Processing (EMNLP), 482-491.

[21] Koppel, M., Schler, J., & Bonchek-Dokow, E. (2007), “Measuring differentiability: Unmasking pseudonymous Authors”, Journal of Machine Learning Research, 8, 1261-1276.

[22] Siham, O. & Halim, S. (2012), “Authorship Attribution of Ancient Texts Written by Ten Arabic Travelers Using a SMO-SVM Classifier”, The 2nd International Conference on Communications and Information Technology (ICCIT): Digital Information Management, Hammamey, 44-47.

[23] Elayidom, M. S., Jose, C., Puthussery, A., & Sasi, N. K. (2013), “Text Classification for Authorship Attribution Analysis”, Advanced Computing: An International Journal (ACIJ), Vol.4, No.5, 1-9.

[24] YANG Y, PEDERSEN J Q. A comparative study on feature selection in text categorization. Proceedings of the 14th International Conference on Machine Learning (ICML), 1997: 2-3.

[25] Hong, R., Tan, R., & Tsai, F, S. (2010), “Authorship Identification for Online Text”, International Conference on Cyberworlds, 155-162.