Authorship Attribution on Imbalanced English Editorial Corpora

O. Srinivasa Rao, PhD
Associate Professor,
COEJNTUK. Kakinada,
Andhra Pradesh, India

N. V. Ganapathi Raju, PhD
Professor of CSE, Gokaraju
Rangaraju Inst. of Eng. and
Tech., Telangana, India.

V. Vijaya Kumar, PhD
Professor & Dean of CSE,
Anurag Group of Institutions,
Telangana, India

ABSTRACT
Authorship attribution is one of the important problem, with many applications of practical use in the real-world. Authorship identification determines the likelihood of a piece of writing produced by a particular author by examining the other writings of that author. Every author has a unique style of writing pattern. This paper identifies the unique style of an author(s) using lexical stylometric features including function words using balanced training corpus. The present paper calculates the frequencies of the lexical based stylometric features by balancing training and test corpus on English editorial documents. The present paper compares various machine learning algorithms for the authorship attribution and achieved highest average accuracy 95.58 using Random Forest classifier and 92.59 using Multilayer Perceptron algorithms.

General Terms
Imbalanced corpora; Editorial Corpus; Function words; Lexical Features;

Keywords
Authorship Clustering; Stylometry; Supervised Classification

1. INTRODUCTION
Authorship attribution is a problem since last decade. The rapid growth of the electronic documents in internet in the form of emails, blogs, social networking, news groups, twitter, Face book, etc. has created multitude ways to share information across the World Wide Web. The surfeit of available electronic texts revealed the potential of authorship analysis in various applications areas including intelligence, criminal law and civil law, computer forensics, literary research etc. [6]. Hence authorship identification has become an emerging research area in information retrieval research. Authorship attribution is a process of examining the characteristics of a piece of writing to draw conclusions on its authorship. Its roots are from a linguistic research area called stylometry, which refers to statistical analysis of literary style. Style in written language refers to the variable ways that language is used in certain genres, periods, situations and individuals. The purpose of evaluating stylistics is to identify writer's subconscious habits of writing style. The present research measures textual features in term of quantitative for various authors then compares known writings of authors with unknown (anonymous) text and assigns the unknown text to the correct author.

The authorship attribution methods identifies the authors based on their attribute or style of writers. Every human has his own writing style. He consciously or unconsciously use certain terminology in his writing style. According to Van Halteren [10] the term "human stylome," represents a specific set of measurable traits that can be used to uniquely identify a given author. The factors that influence authorship attribution are training and test corpus size, number of candidate authors, distribution of training corpus over the authors (balanced or imbalanced corpus). A common problem in authorship identification is the lack of text samples of undisputed authorship to be used for training. In some cases extremely limited text samples to be available for some authors. On the other hand, a big amount of text samples may be available for other candidate authors. Note that text samples should be of comparable length. Another realistic scenario is to have (more or less) equal amount of texts of undisputed authorship for all the candidate authors, however short texts are available for some of them and long texts for others. Hence, in the procedure of normalizing the length of training text samples, few text samples will be produced for some authors and many text samples for others. Many studies have shown that balanced dataset provides improved overall classification performance compared to an imbalanced data set [8, 9, 14, 15].

The present research identifies writing style of the author by analyzing stylistic features of on English editorial documents. The editorial documents written by various authors may not have same length, hence leads to class imbalance problem. To overcome the class imbalance problem, the normalized method is used to calculate various lexical based stylistic features for the authorship attribution. The present paper considers frequencies of average length per word, frequencies of 150 function words, frequency occurrences of punctuations and frequency occurrences of bigram, trigram, quadgram words are calculated in authorship identification on editorial columns, and several machine learning techniques are considered to build feature-based classification models to perform authorship identification.

2. LITERATURE SURVEY
Many researchers are worked on authorship attribution to quantify the writing styles of the authors by considering various stylometric features. The features can be classified as lexical, character, syntactic, semantic and function words. Lexical features include word length, sentence length, word frequencies, vocabulary richness functions, word n-grams etc. Character features include frequency of character types, frequency of character n-grams. The majority of authorship attribution studies is based on lexical features to represent the style and out of these N-gram model is a relatively simple idea, and it has been found to be effective in authorship attribution. To extract the style markers of an author N-gram based features are considered, where "N-gram" is the term for any sequence of n words/n characters. In natural language processing the presence of one-, two-, and three-word sequences is known as unigrams, bigrams, and trigrams, respectively. In the literature, it has been demonstrated that character or word n-grams and function words are among the

most effective stylometric features that improve the performance of an attribution model. For authorship attribution, the most frequent words have contributed as the most utilitarian feature. The most common words (articles, prepositions, pronouns, etc.) are the best features to distinguish between authors. They carry no semantic information and they are usually called 'function' words. The selection of the function words is based on arbitrary criteria which is generally language-dependent. Various authors worked on functional and significant words of English language for author attribution. Due to their high frequency in the language and highly grammaticalized roles, function words are questionable to be subject to conscious control by the author. Also to be considered is that the frequencies of different function words vary extensively across different authors and genres of text – hence the hope that modeling the interdependence of different function word frequencies with style will result in effective attribution.

Most lexical features are highly author and language dependent. The common approach to determining authorship is to use stylistic analysis that proceeds in two steps: first, specific style markers are extracted, and second, a classification procedure is applied to the resulting description. Authorship attribution methods are divided in to two categories i.e. instance-based and profile-based [6]. Most of the earlier researchers in this field carried out author attribution based on these categories only. The instance based methods consider each training document individually whereas profile based methods consider all the documents of an author as a single document (cumulative) [6]. Many researchers attempted authorship attribution problem by extracting style markers on both balanced and imbalanced training corpus using several lexical, character, syntactic and semantic features on imbalance training and test data sets. The performance of the authorship attribution degrades using imbalance training data sets compared with balanced training corpus. [1,7] tried authorship on four types of writing-style features (lexical, syntactic, structural, and content-specific features) are extracted and inductive learning algorithms are used to build feature-based classification models to identify authorship on English and Chinese online-newsgroup messages. [2] attempted authorship on few training texts at least for some of the candidate authors or there is a significant variation in the text-length among the available training texts of the candidate authors. [3] Assumption of quantitative authorship attribution is that the author of a text can be selected from a set of possible authors by comparing the values of textual measurements in the text to their corresponding values in each author's writing sample on English poems. [4] et. al. found that Document Author Representation(DAR) can be very useful in AA tasks, because it provides good performance on imbalanced data, getting comparable or better accuracy results. [5, 6] investigated the class imbalance problem and tests several methods for compensation of imbalanced data sets. He concludes that the best method uses many short text samples for minority classes and less but longer ones for the majority classes. [18] observed that the amount of training material has more influence on performance than the amount of test material. In order to obtain reliable performance, they find that 5,000 words in training can be considered a minimum requirement. But none of the earlier researchers attempted normalized imbalanced training and testing data sets for the authorship attribution. In the present paper the authorship attribution on imbalanced editorial documents is presented using average word length, frequencies of function words, frequency of bigram/trigram/quadgram words, and frequency occurrence of punctuations.

3. METHODOLOGY

The present paper focused on the closed-class authorship attribution problem in which the real author is one of several possible candidates and the attribution model used is an instance based approach, i.e., each training text is individually represented as a separate instance of author style.

3.1 Algorithm

Step 1:- The present paper initially performs a preprocessing method on all documents. The pre-processing method makes the corpus text as case insensitive, and performs other operations like data cleaning, tokenizing, normalization, removal of stop words for effective feature extraction. By this step, spaces, numbers, special characters from the corpus articles are eliminated.

Step 2:- The present paper considers four kinds of lexical based stylometric features are extracted for authorship identification.

2.1:- A significant advantage of average length per word is that it can be applied to any language and any corpus with no additional requirements except the availability of a tokenizer. Average Length per Word (AW) can be calculated as

$$AW=\frac{total\ number\ of\ characters}{total\ number\ of\ words} \quad \text{……. (1)}$$

2.2:- The present paper focused on the use of lexical-based features of an author's style, and evaluated most frequently used N-gram (unigram, bi-gram and tri-gram with and without overlapping) features after performing the pre-processing step. Frequency occurrences of bigrams, trigrams and quadgrams are T. By normalizing it becomes

$$\frac{100}{TC} * T \quad \text{……..…. (2)}$$

Where TC total number of characters.

2.3:- Punctuation marks count, is easily available for any natural language and corpus and it has been proven to be quite useful to quantify the writing style for authorship identification. Making strange use of punctuation, as it usually happens in e-mails or online forum messages. It is well known that punctuation has the potential of being a successful attributor of authorship, it has only really been successful when combined on its own with an understanding of its syntactic role in a text. Frequency occurrences of Punctuations: Total 8 features are considered for punctuations. They are ",", ".", "?", "!", ":", ";", "''", "" " are calculated as

$$\frac{100}{total\ number\ of\ sentences} * frequency\ occurance\ of\ punctuations \quad \text{…… (3)}$$

2.4:- Frequency occurrences of Function Words (FW). Total 150 function words are considered as features.

$$FW = \frac{100}{total\ number\ of\ words} * frequency\ occurance\ of\ function\ words \quad \text{… (4)}$$

Step 3:- Once the feature vectors are calculated on every document based on the author using Vector Space Model then assign a class label to all documents of the author.

Step 4:-The feature vectors are given as input to supervised machine learning classifiers for the prediction of author of an unknown document.

4. RESULTS AND DISCUSSION

The style based features are implemented on a collection of 250 editorial documents from the seven leading columnists of India i.e...(1) M.J.Akbar, (2) Chetan Bhagat, (3) A.S.Panneerselvan, (4) C.Raja Mohan and (5) Tavleen Singh. The editorials are collected from the leading newspapers of India namely The Hindu, Times of India and Sunday Guardian. 50 documents of each author has been considered for both training and testing purpose. On the training document the same is evaluated and given to Support Vector Machines, Multilayer Perceptron, Bagging, Logic Boost, Random Forest classifiers using Weka (Waikato Environment for Knowledge Analysis) software package Version 3.7 for an effective author attribution.

Table 1: Accuracy of various classifiers on English editorial documents

	Support Vector Machines	Multi-Layer Perceptron	Bagging	Logic Boost	Random Forest
PS	90.3	91.81	85.54	93.03	95.15
RM	91.51	93.63	85.15	90	96.06
CB	90.6	92.42	82.42	90.3	95.45
AKB	90.3	92.72	85.75	91.81	95.45
TS	90.6	91.81	82.12	91.81	96.66

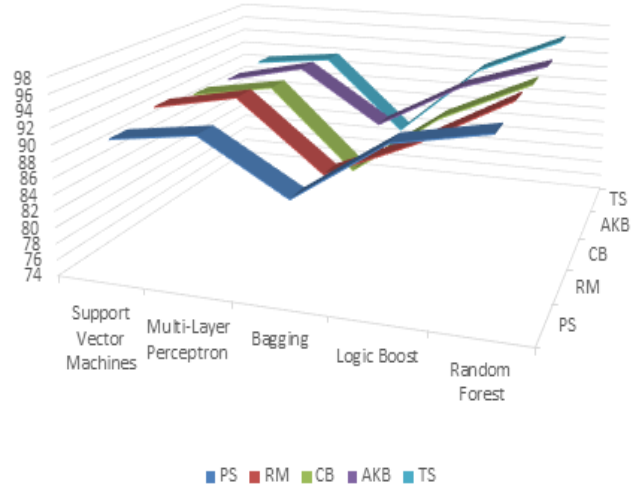

Fig 1: Accuracy of various classifiers on English editorial documents

Based on the above Tables 1, 2 and Figure 1, 2, it is observed that Random Forest classifier outperforms all other algorithms with an average accuracy of 95.74 then Multilayer Perceptron classifier with 92.47 in identifying the author of an unknown document.

Table 2: Average accuracy of various classifiers on English editorial documents

Classifier	Support Vector Machines	Multi-Layer Perceptron	Bagging	Logic Boost	Random Forest
average	90.662	92.478	84.196	91.39	95.754

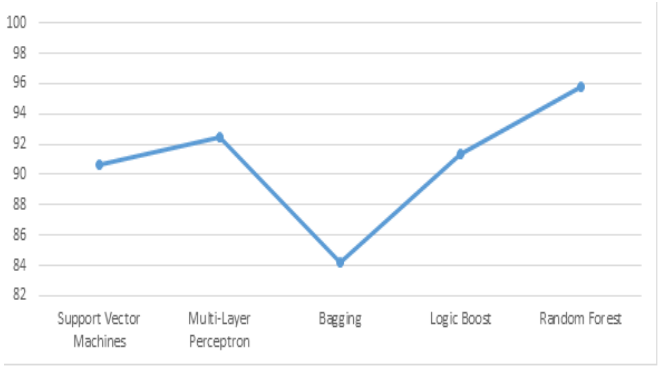

Fig 2: Average accuracy of various classifiers on English editorial documents

5. CONCLUSIONS

This paper predicts author of an unknown/disputed document using lexical based stylometiec features with various supervised machine learning classifiers. The paper uses normalized stylometric features on imbalanced training data sets. The method used achieved average accuracy of 95.74% using Random Forest classifier. The method is useful in forensic applications. The future scope of the paper is to present syntactic and semantic stylistic features on imbalance corpora.

6. REFERENCES

[1] Zheng, Rong, et al. "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 (2006): 378-393.

[2] Stamatatos, Efstathios. "Author identification: Using text sampling to handle the class imbalance problem." Information Processing & Management 44.2 (2008): 790-799.

[3] Grieve, Jack. "Quantitative authorship attribution: An evaluation of techniques." Literary and linguistic computing 22.3 (2007): 251-270.

[4] López-Monroy, Adrián Pastor, et al. "A new document author representation for authorship attribution." Mexican Conference on Pattern Recognition. Springer Berlin Heidelberg, 2012.

[5] Luyckx, Kim, and Walter Daelemans. "The effect of author set size and data size in authorship attribution." Literary and linguistic Computing 26.1 (2011): 35-55.

[6] Stamatatos, Efstathios. "A survey of modern authorship attribution methods." Journal of the American Society for information Science and Technology 60.3 (2009): 538-556.

[7] Cheng, Na, Rajarathnam Chandramouli, and K. P. Subbalakshmi. "Author gender identification from text." Digital Investigation 8.1 (2011): 78-88.

[8] Layton, Robert. "A Simple Local n-gram Ensemble for Authorship Verification." CLEF. 2014.

[9] Wei, Qiong, and Roland L. Dunbrack Jr. "The role of balanced training and testing data sets for binary classifiers in bioinformatics." PloS one 8.7 (2013): e67863.

[10] Van Halteren, Hans, et al. "New machine learning methods demonstrate the existence of a human stylome." Journal of Quantitative Linguistics 12.1 (2005): 65-77.

[11] V. Vijaya Kumar, N V Ganapathi Raju, O Srinivasa Rao, "Histograms of Term Weight Feature (HTWF) model for Authorship attribution",International Journal of Applied Engineering Research (IJAER), vol10, number 16 , pp 36622-36628, ISSN 0973-4562, 2015

[12] N V Ganapathi Raju, V. Vijaya Kumar,O Srinivasa Rao, "Authorship attribution of Telugu texts based on Syntactic features and Machine learning techniques", Journal of Theoretical and Applied Information Technology (JATIT), volume 85, No.1, ISSN: 1992-8645, march 2016

[13] N V Ganapathi Raju, V. Vijaya Kumar, OSrinivasa Rao, "Author based Rank Vector Coordinates (ARVC) model for Authorship attribution", International Journal of Image, Graphics and Image Processing (IJIGSP), Vol. 8, No. 5, May 2016.

[14] McMenamin, Gerald R. "Style markers in authorship studies." International Journal of Speech Language and the Law 8.2 (2007): 93-97.

[15] Stamatatos, Efstathios. "Text Sampling and Re-Sampling for Imbalanced Authorship Identification Cases." Frontiers in Artificial Intelligence and Applications 141 (2006): 813.

[16] Zhao, Ying, and Justin Zobel. "Searching with style: Authorship attribution in classic literature." Proceedings of the thirtieth Australasian conference on Computer science-Volume 62. Australian Computer Society, Inc., 2007.

[17] Eder, Maciej. "Style-markers in authorship attribution a cross-language study of the authorial fingerprint." Studies in Polish Linguistics 6.1 (2011): 99-114.

[18] Sanderson, C., & Guenter, S. "Short text authorship attribution via sequence kernels, Markov chains and author unmasking: An investigation", In Proceedings of the International Conference on Empirical Methods in Natural Language Engineering, Pages 482-491,2006.