Style based Authorship Attribution on English Editorial Documents

N. V. Ganapathi Raju
Associate Professor,
Dept. of CSE,
GRIET, Hyderabad,
India

Ch. Sadhvi
Department of
I.T.,GRIET
GRIET, Hyderabad,
India

P. Tejaswini
Department of
I.T.,GRIET
GRIET, Hyderabad,
India

Y. Mounica
Department of
I.T.,GRIET
GRIET, Hyderabad,
India

ABSTRACT
The aim of the authorship attribution is identification of the author/s of unknown document(s). Every author has a unique style of writing pattern. The present paper identifies the unique style of an author(s) using lexical stylometric features. The lexical feature vectors of various authors are used in the supervised machine learning algorithms for predicting the unknown document. The highest average accuracy achieved is 97.22 using SVM algorithm.

General Terms
Style based Classification; Information Retrieval; Natural Language Processing; Authorship Attribution; Machine Learning

Keywords
Style based Classification; Lexical features; Function words/Stop words; Authorship Attribution/Profiling;

1. INTRODUCTION
Authorship attribution is the process of drawing conclusions on its authorship by examining piece of writing with characteristics. Its roots are from stylometry which is linguistic area, which refers to statistical analysis of literary style. TheVariable ways that the language is used is in certain genres, periods, situations and individuals that refers to the style in written language. 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.

Various fields such as computational linguistics, natural language processing, information retrieval and machine learning has significant impact of authorship attribution. Authorship attribution has diverse applications including intelligence, criminal law and civil law, computer forensics and literary research. Since last decade the vast amount of electronic texts are available through internetmedia in the form of e-mails, blogs, online forum messages, news groups, source code, etc. This has also given rise to various kinds of misuses of content and it is becoming highly difficult to identify the original author of the documents.

The authorship attribution methods identifies the authors based on their attribute or style of writers. Every human has his own writing style. They consciously orunconsciously use certain terminology in his writing style. According to Van Halteren the term "human stymie," represents a specific set of measurable traits that can be used to uniquely identify a given author. Resemblance, consistency, and population models are the most fundamental models of authorship analysis according to McMenamin [8]. The resemblance model employs nonlinguistic evidence in order to narrow down the group of suspect authors to one or a limited number of authors and identify just the author; the consistency model employs sample of writings in order to determine whether two or more writings have been produced by the same author or multiple authors; the population model employs external (nonlinguistic) evidence to opt for the suspect author or authors out of a large number of candidate authors.

2. LITERATURE SURVEY
Many earlier researchers studied authorship attribution by quantifying the writing style of the authors in terms of lexical/ syntactic/ semantic/ application specific levels. The present paper uses style based authorship attribution using character, word based features. Zhenget.al. [1] used 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 of online messages. Cheng et.al.[2] investigated authorship identification for short length, multi-genre, context-free text found in the internet by considering 545 psych-linguistic features. Stamatatos [3,4] 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. Grieve [9] 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. Zhao et.al [12]. On collection of 634 texts by 55 authors on English poems authorship identification is explored. Elder [13] attempted authorship on literary texts using frequencies of the most frequent words.

3. METHODOLOGY
The present paper performs authorship identification by analyzing stylistic features on English editorial documents. The paper considers three types of textual features that are identified in authorship identification research are extracted from editorial columns, and several machine learning techniques are used to build feature-based classification models to perform authorship identification.

3.1 Style based Features

Three kinds of style based text features has been considered for the experimentation, includes character-based, word-based and function words. Character-based features include 53 stylometric features adopted in earlier authorship attribution studies, such as number of letters (a-z), number of uppercase characters (A-Z), digits (0-9), number of white spaces, a number of special characters (e.g., %, &,), etc. Word-based features include 19 statistical metrics such as hapaxlegomena (words that occur only once), hapaxdislegomena(words that occur only twice), average word length, vocabulary richness, average sentence length, type token ratio, number of bi-gram, tri-gram, quad-gram characters, and Vocabulary rich number measure(total number of different words/total number of words) such as Yule's K, Simpson's D, Sichel's S, Honore's R, Entropy measures are considered for the attribution. Total 227 stylistic features are considered for the experimentation. The most common words (articles, prepositions, pronouns, etc.) called function words that have little lexical meanings or are found to be among the best features to discriminate between authors. Totally 150 function words were considered as features for identifying the author's task. The present research was conducted on English editorial documents with style based character, word, function word based features and feature value extraction was implemented in our Java program.

3.2 Performance measures

Standard information retrieval metrics of precision, recall, and F1has been used for evaluating authorship identification.

Precision, for a particular author A, is defined as the fraction of attributions that a system makes to A that are correct:

$P_A = \mathrm{Correct(A)/Attributions(A)}$

Recall, for a particular author A, is defined as the fraction of test documents written by A that are (correctly) attributed to A:

$R_A = \mathrm{Correct(A)/documents - by(A)}$

F1 is defined as the harmonic mean of recall and precision:

$F1 = 2\ P_A\ R_A / P_A + R_A$

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. 50 documents of each author has been considered for both training and testing purpose.The editorials are collected from the leading newspapers of India namely The Hindu, Times of India and Indian express. On the training document the same is evaluated and given to Naive Bayes, Support Vector Machines, Multilayer Perceptron classifiers using Weka (Waikato Environment for Knowledge Analysis) software package Version 3.7 for an effective author attribution.

Table 1 shows that accuracy of authorship attribution on various classifiers along with precision, recall and F measure. From the table it is observed that support vector machines and multilayer perceptron algorithms are performing well in identification of author of an unknown document.

Table 1: Results of style based classification on various Machine learning classifiers

Style based Lexical Features												
	NB classifier				SVM SMO classifier				MLP classifier			
Author Name	% classified	PA	RA	F1	% classified	PA	RA	F1	% classified	PA	RA	F1
Akbar	79.45	0.819	0.795	0.79	96.63	0.967	0.966	0.966	95.26	0.955	0.953	0.952
Chetan Bhagat	76.87	0.812	0.769	0.767	97.27	0.975	0.973	0.973	96.63	0.967	0.966	0.966
Panneerselvan	76.87	0.784	0.769	0.76	97.95	0.981	0.98	0.979	95.27	0.955	0.953	0.953
Raja Mohan	75.51	0.775	0.755	0.749	97.63	0.978	0.976	0.977	96.63	0.967	0.966	0.966
Tavleen Singh	77.55	0.795	0.776	0.773	96.63	0.968	0.966	0.966	94.63	0.947	0.946	0.946

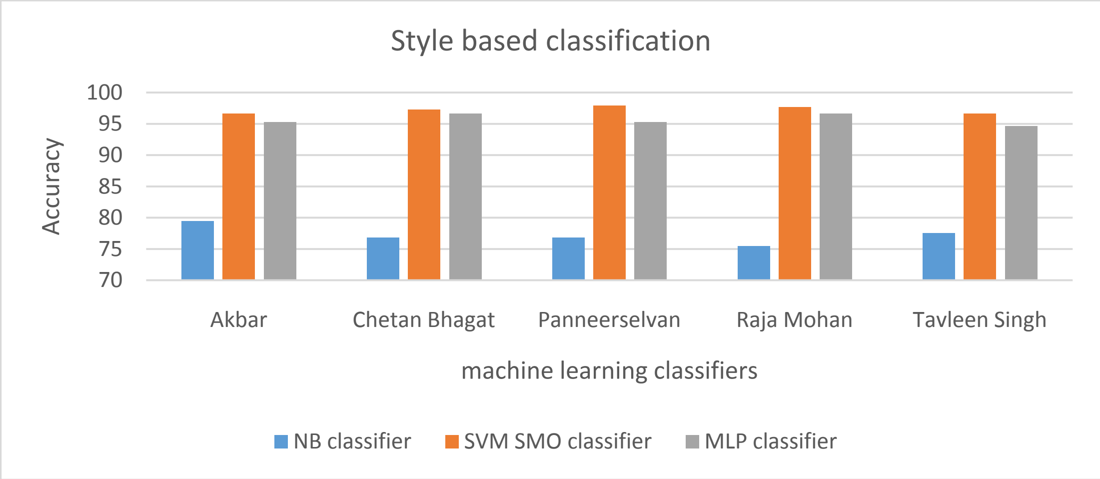

Fig 1: Accuracy for authorship identification on classifiers

Table 2: Average accuracy of style based classification

	NB classifier				SVM SMO classifier				MLP classifier			
Author Name	% of classified	PA	RA	F1	% of classified	PA	RA	F1	% of classified	PA	RA	F1
Style based Features	77.25	0.8	0.77	0.77	97.22	0.97	0.97	0.97	95.68	0.96	0.96	0.96

From fig.1 shows the pictorial representation of above table1. From the table 2, it is observed that average accuracy of SVM classifier outperforms other classifiers with an average accuracy of 97.22.

5. CONCLUSIONS

The present paper implements authorship attribution using lexical based stylometric features including 150 function words and predicts the author of an unknown document using supervise machine learning classifiers on English editorial documents. The highest average accuracy achieved is 97.22 using SVM algorithm. In future authorship identification using syntactic and semantic features need to be explored to increase the accuracy and also to implement authorship profiling features for the identification of gender identification.

Authorship profiling has many applications in forensics, security, and marketing.

6. REFERENCES

[1] Zheng, Rong, et al. "A framework for authorship identification of online messages:

[2] Writing- style features and classification techniques." *Journal of the American Society for Information Science and Technology* 57.3 (2006): 378-393.

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

[4] Juola, Patrick. "Authorship attribution." *Foundations and Trends in information Retrieval* 1.3 (2006): 233-334.

[5] Stamatatos, Efstathios, Nikos Fakotakis, and George Kokkinakis. "Automatic authorship attribution." *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 1999.

[6] Gamon, Michael. "Linguistic correlates of style: authorship classification with deep linguistic analysis features." *Proceedings of the 20th international Conference on Computational Linguistics*. Association for Computational Linguistics, 2004.

[7] Stamatatos, Efstathios, Nikos Fakotakis, and Georgios Kokkinakis. "Computer-based authorship attribution without lexical measures." *Computers and the Humanities* 35.2 (2001): 193-214.

[8] Khachatryan, R. Kh. "Desideratum of New reality: forensic authorship attribution." *Вестник Московского государственного лингвистического университета* 24 (2012).

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

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

[11] Chen, Zhili, et al. "More than Word Frequencies: Authorship Attribution via Natural Frequency Zoned Word Distribution Analysis." *arXiv preprint arXiv:1208.3001* (2012).

[12] Dinu, Liviu P., and Marius Popescu. "Ordinal measures in authorship identification." *Proc. SEPLN*. 2009.

[13] 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.

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

[15] 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.

7. AUTHOR PROFILE

Ch Sadhvi 13241A1260 Department of IT

P Tejaswini 13241A1279 Department of IT

Y Mounica 13241A12B Department of IT

8. APPENDIX

Style based text features

F1	Total number of characters(C)
F2	Total number of letters (a-z)/C
F3	Total number of upper characters/C
F4	Total number of digital characters/C
F5	Total number of white-space characters/C
F6	Total number of Special characters/C
F7-F32	Frequency of Upper case characters A, to Z (26)
F33-F53	Frequency of special characters (~,@,#,$,%,^,&,*,-,_,=,+,>,<,[,],{,},/,\,|)
F54	Total No. of Words (N)
F55	Average Length per word
F56	total no of short words
F57	Total No.of Different words/no of words
F58	Hapax Legomena
F59	Hapax Dis legomena
F60	Average sentence length in terms of characters
F61	Average sentence length in terms of words
F70-F219	Frequency of Function words
F220	Number of Bi-Gram Characters
F221	Number of Tri-Gram Characters
F222	Number of Quad-Gram Characters
F223	Simpsons D measure $\sum_{i=1}^{v} Vi\,\frac{i}{N}\,\frac{i-1}{N-1}$
F224	Sichel's S measure $count\ of\ Hapax\ Dislegomena/V$
F225	.Honores R measure $\frac{100\log_{10}N}{1-\frac{\text{count of Hapax Legomena}}{V}}$
F226	$\sum_{i=1}^{N} V_i\left(-\log_{10}\frac{i}{N}\right)\frac{i}{N}$ .Entropy
F227	$10^4\left(-\frac{1}{N}+\sum_{i=1}^{V} V_i\left(\frac{i}{N}\right)^2\right)$ .YulesK

V: number of different words
Vi: number of different words that occur i times
N : total number of words.

Function Words (150)

a, between, in, nor, some, upon, about, both, including, nothing, somebody, us, above, but, inside, of, someone, used, after, by, into, off, something, via, all, can, is, on, such, we, although, coos, it, once, than, what, am, do, its, one, that, whatever, among, down, latter, onto, the, when, an, each, less, opposite, their, where, and, either, like, or, them, whether, another, enough, little our these which any every lots outside they while anybody everybody, many, over, this, who, anyone, everyone, me, own, those, whoever, anything, everything, more, past, though, whom, are, few, most, per, though, whose, around, following, much, plenty, till, will, as, for, must, plus, to, with, at, from, my, regarding, toward, within, be, have, near, same, towards, without, because, he, need, several, under, worth, before, her, neither, she, unless, would, behind, him, no, should, unlike, yes, below, I, nobody, since, until, you, beside, if, none.