Optimizing Authorship Profiling of Online Messages

Adeola O. Opesade
Africa Regional Centre for Information Science,
University of Ibadan, Nigeria
morecrown@gmail.com

ABSTRACT

Authorship profiling is of growing importance in the current information age, partly due to its application in digital forensics. Methodologies of profiling like any other authorship analysis consist majorly of feature extraction and application of analytical techniques. Choice of feature sets and analytical techniques may significantly affect the performance of authorship analysis. Hence, a need for methods that can help improve on the success of authorship profiling undertakings. The present study sought through experiments, the writing features, analytical technique and number of class labels that can help improve the effectiveness of profiling the country of affiliation of authors of online messages. The experiment showed that the most effective model was achieved when all feature set types in our study were used within a two-class dataset that was analysed with the Neural Network (Multilayer Perceptron) machine learning scheme. The study recommends a need for further studies in finding models that can maximize both effectiveness and efficiency in profiling the authorship of online messages.

CCS Concepts

• General and reference → Cross-computing tools and techniques →Experimentation

Keywords

Authorship profiling, Machine learning, Computational linguistics, Natural Language Processing, Nigerian English

1. INTRODUCTION

Electronic messages are extensively used to distribute information over such channels as e-mail, Internet newsgroups, Internet chat rooms, Internet forums and other user generated contents on the Web. These messages are quite different from other forms of writings particularly, because of their brevity. Unfortunately, unethical hands and criminals exploit the convenience of these media to carry out their obnoxious goals. Digital forensics require the use of scientifically derived and proven methods towards the preservation, collection, validation, identification, analysis, interpretation, documentation and presentation of digital evidence derived from digital sources for litigation purposes.

Authorship profiling is one of the major classes of authorship attribution problems. It seeks the demographic or psychological group of the author of an anonymous text. Its application in forensics and digital security has made it to be of growing importance in the present information age. Methodologies of profiling like any other authorship analysis consist majorly of feature extraction and application of analytical techniques. Choice of feature sets and analytical techniques may significantly affect the performance of authorship analysis [1]; thus, studies into optimization of authorship profiling of online messages can assist in improving the success of identifying sources of security threats perpetrated through web-based channels.

CoRI'16, Sept 7–9, 2016, Ibadan, Nigeria.

A number of previous studies ([1]; [22]; [3]) have investigated some parameters that could affect the effectiveness of authorship attribution undertakings. These studies, however, focused on authorship identification problem and not on authorship profiling. Considering the potential of authorship profiling in investigating transnational digital breaches, the present study seeks to find through experiments the writing-style features, classification techniques as well as possible number of class options that can maximize the effectiveness of profiling the authorship of electronic messages. The following research questions were pursued in order to achieve the purpose the study:

Research Question 1: Which feature type set maximizes the effectiveness of profiling the country of affiliation of writers of online messages?

Research Question 2: Which classification scheme maximizes the effectiveness of profiling the country of affiliation of writers of online messages?

Research Question 3: Which class labelling option maximizes the effectiveness of profiling the country of affiliation of writers of online messages?

Research Question 4: What is the performance of the resultant model in classifying electronic messages to writers' countries of affiliation?

2. LITERATURE REVIEW

2.1 Authorship Attribution Problems

Authorship attribution is a process of examining the characteristics of a piece of writing in order to draw conclusions about its author. Authorship attribution problems vary in complexity. They have been categorized into three major classes, namely, authorship identification, authorship profiling and authorship verification. The most straightforward version of these three is the identification problem which involves the determination of the actual author of a given text among a small set of candidate authors. Given a set of writings of a number of authors, the task in authorship identification is to assign a new piece of writing to one of them [4]. In authorship verification, there is no closed candidate set but there is one suspect and the challenge is to determine if the suspect is or is not the author. In this case, examples of the writing of a single author are given and the task is to verify that a given target text was or was not written by this author. Hence, verification can be thought of as a one-class classification problem and it is significantly more difficult than basic authorship identification problem [5].

In authorship profiling (also known as authorship characterization problem) there is no candidate set at all; the challenge is to provide as much demographic or psychological information as possible about the author. Unlike the identification problem, authorship profiling does not begin with a set of writing samples from known candidate authors. Instead, it exploits the sociolinguistic observation that different groups of people speaking or writing in a particular genre and in a particular language, use that language differently; that is, they vary in how often they use certain words or syntactic constructions in addition to variation in pronunciation or intonation [6]. Profiling problem is concerned with determining such characteristics as gender, educational and cultural backgrounds, language familiarity and so on of the author that produced a piece of work. This is a harder problem than the identification problem since it characterizes the writing style of a set of writers rather than the unique style of a single person [7].

Despite variations in the complexities of authorship problems, choices of appropriate linguistic features and analytical techniques are paramount.

2.2 Authorship Attribution Methods

One of the main components of authorship attribution methods is the extraction of linguistic features that represent the writing style of an author or author group. Language, like genetics, can be characterized by a very large set of potential features that may or may not show up in any specific sample, and that may or may not have obvious large-scale impact. By identifying the features characteristic of a group or individual of interest, and then finding those features in an anonymous document, one can support a finding that the document was written by that person or a member of that group [8]. The various feature sets, otherwise known as feature metrics in computational linguistics can be classified into four main classes, which are the lexical, syntactical, content-specific and structural features [9]. Researchers vary in their choices of linguistic features; while some used feature(s) that belong to a single class (for example, [10]; [11]; [12]; and [9], others (such as [6]; [2]; [4]; [3]; [7]; [1]; [13]; [14]) used features across multiple feature classes.

The second component is the application of analytical techniques to feature sets for supervised or unsupervised learning. Different analytical techniques have been used in previous authorship attribution studies. These techniques can be classified into three, namely, the unitary invariant, multivariate and machine learning approaches [8]. Machine learning examines previous examples and their outcomes and learns how to reproduce these and make generalisations about new cases. Machine learning algorithms differ in terms of level of data and abilities to resolve data ambiguities such as noise or missing data. Machine learning techniques include rule based algorithms such as OneR, neural networks such as Multilayer Perceptron, statistical modelling algorithm such as Naive Bayes, decision trees such as J48, linear models such as linear regression and Support Vector Machine and instance-based learning algorithm such as Nearest Neighbour.

Unlike in the choice of feature sets, researchers are less varied in their choices of analytical techniques. While older studies tend to favour the use of Principal Component Analysis, the more recent ones tend towards the use of Support Vector Machine. Most previous studies reported the use of only a single analytical technique. Considering such statement as made by [15].

> Experience shows that no single machine learning scheme is appropriate to all data mining problems. The universal learner is an idealistic fantasy. Real datasets vary and to obtain accurate models, the bias of the learning algorithm must match the structure of the domain. Data mining is an experimental science (pg 365).

Choice of machine learning scheme should be based on the result of a prior experiment that validates its suitability to the dataset.

2.3 Related Authorship Studies

A number of previous studies have shown relative performances of a number of feature types and analytical techniques in authorship analyses. [3] studied the results of authorship identification using many authors and limited data on learning. Their result showed that systematically increasing the amount of authors under investigation led to a significant decrease in performance. Their study also revealed that providing a more heterogeneous set of features improves the system significantly. [1] investigated the types of writing-style features and classification techniques that were effective for identifying the authorship of online messages. They reported that the accuracy kept increasing as more types of features were used and that Support Vector Machine (SVM) outperformed Neural Networks (NN), which in turn outperformed the C4.5 classifier. The best accuracy was achieved when SVM and all feature types were used but classifier performance reduced as the number of authors increased. [2] through experiment demonstrated that inclusion of stylistic idiosyncrasy features to letter n-grams, function words and to a combination of n-grams and function words consistently led to improved accuracy for identifying the native language of the author of a given English language text.

The studies of [3] and [1] are situated within the identification domain of authorship attribution problems because they started with a close number of candidate authors, while that of [2] was a profiling problem. However, their focus was majorly to show the ability of idiosyncrasies in detecting writer's native language. It therefore, did not address some of the salient issues covered by [1] which are relative performances of analytical techniques and effect of increasing the number of candidate authors. Also, the corpus used by [2] was the International Corpus of Learner English (ICLE) which had between 579 and 846 words. These numbers were quite high for an online message, which are usually very short. The present study focuses on shorter texts which characterise online messages. Therefore, the present study seeks to find the writing-style (linguistic) features, classification techniques as well as possible number of class options that can maximize the effectiveness of profiling the native language of the author of an online message.

3. EXPERIMENTATION FOR OPTIMIZING AUTHORSHIP PROFILING OF ONLINE MESSAGES

3.1 Problem formulation

Given a number of online messages written in English language by nationals of selected African countries, namely, Cameroon, Ghana, Liberia, Nigeria and Sierra-Leone. The goal is to find the types of writing-style features, the classification technique as well as possible number of class options that can maximize the effectiveness of profiling the linguistic origin of anonymous electronic texts written by the nationals of any of the selected countries.

3.2 Research Method

A multistage sampling technique was used to select a representative sample of electronic texts from the population of texts contained in the relevant country pages of the website www.topix.com. To get the texts that could be useful for a supervised learning approach of the study, each text was opened, read and assessed based on the number of words contained and a sense of affiliation to the respective country as depicted in the content. A comment was considered to be affiliated to (and labelled to be from) a particular country if it was found in that country's forum and if it contained such phrases as 'our country', 'our beloved country' and other related ones in its discourse. Initially the researcher targeted selecting texts with a hundred or more words; however, this was reduced to texts with twenty (20) or more words because of the scarcity of large texts on the discussion forums. The numbers of texts selected for the study in November 2011 and based on the assessment criteria are as shown in Table 1.

Table 1: Training Data Set

Country's forum website	No. of pages	Pages selected	No. of selected texts
www.topix.com/forum/world/nigeria	31	2,8,13,25	425
www.topix.com/forum/world/ghana	9	2,3,6.9	317
www.topix.com/forum/world/liberia	4	1-4	130
www.topix.com/forum/world/cameroon	4	1-4	241
www.topix.com/forum/world/sierra-leone	4	1-4	357
Total no. of Texts			1,470

3.2.1 Text Pre-processing and Processing

The corpora were subjected to pre-processing in order to put them in the format expected by the relevant software for text processing. The pre-processing tasks included deletion of e-mail headers, removal of control codes, text aggregation, and removal of non-ASCII characters. Text processing was achieved by extracting linguistic features from the sampled texts using computer codes written by the researcher in Python 2.6.4 programming language, based on the natural language toolkit (NLTK) version 2.0. Some of the specific issues handled in the course of text processing were tokenization, part of speech tagging and linguistic feature extraction.

Although there is no agreement on a best set of features for a wide range of application domains, selected feature metrics must be reliable characteristic of attribution domain [21]. Certain features were extracted in the present study, based on their relevance as determined from relevant literature on authorship attibution and Nigerian Englishes ([16]; [17]). Extracted features were syntactic features comprising the twenty (20) most frequent function words in the *topix.com* corpus, Idiosyncratic features comprising frequency of occurrence of spelling errors, adverb-verb part of speech (POS) bigram distribution and article omission/inclusion distribution. Structural features comprising lexical diversity; and content specific features consisting of twenty (20) most frequent noun, adjective, verb and adverb unigrams in the *topix.com* corpus. The features extracted and their denotations are as shown in Table 2.

Table 2: Extracted Linguistic Features

Feature type	Feature metric	Denotation
Lexical	Vocabulary richness	F1
Syntactic	Probabilities of occurrence of most occurring function words	F2
Idiosyncrasies	Probabilities of occurrence of article deletion, verb -adverb sequence and spelling errors.	F3
Content specific	Noun unigrams, adjective unigrams, verb unigrams, adverb unigrams.	F4

The decision to extract twenty most frequent features (function word, noun, adjective, verb and adverb unigrams) was as a result of a prior experiment which showed that the summation of the frequencies of occurrence of the twenty most frequent features accounted for at least 60% of the cumulative frequency of all features extracted in each case.

3.3 Experimental Setup

i. Class Labelling: According to the study of [3] learner's performance changes with number of candidate authors. To find out the effect of varying the number of classes on the classification performance in the present study, the dataset was copied into three different files having all parameters being the same except the class labels. The class labels were controlled as presented in Table 3.

Table 3: Dataset Class Labelling Options

File Name	No of Class Labels	Class Labels	Remark
Dataset1	5	Nigeria, Ghana, Cameroon, Liberia, Sierra-Leone	Labelling according to texts' original classes.
Dataset 2	3	Nigeria, Ghana, Non-Ghana-Nigeria	Labelling informed by language similarities between the selected countries as found in a previous study [21].
Dataset 3	2	Nigeria, Non-Nigeria	Testing a 2-class labelling scheme which can enable the identification of online texts from a country from those of other countries put together.

The texts in Dataset 1 bear their original class labels, that is, the actual countries of affiliation of the writers as determined from the forums and the texts. There are therefore five different class labels, representing the five country sources of the texts. Dataset 2 has three class labels; texts from Nigeria and Ghana bear their original country source labels while those from the other three countries were combined and labelled 'Non-Ghana-Nigeria'. This was informed by a previous study that showed varying degrees of similarity in the English language usage among the selected countries. Dataset 3 labelled texts from Nigeria as Nigeria while texts from the other four countries were combined under the label ' Non-Nigeria'. This was done to achieve a two-class dataset option.

Experiments were carried out using the Experimenter interface of the open source Waikato Environment for Knowledge Analysis (WEKA) machine learning tool. In this study, four machine learning algorithm implementations in WEKA namely naïve Bayes, SMO (SVM implementation), J48 and Multilayer perceptron (Neural network implementation) were used. The experiment was carried out to compare the performances classifier models in the phase of:

a. Changing the number of classes.
b. Changing the linguistic feature sets.
c. Changing classifier algorithms.

Each of the three datasets (Dataset 1, Dataset 2 and Dataset 3) with each of the feature set types (F1, F2, F3, F4) and all their possible combinations (F1+F2, F1+F2+F3, F1+F2+F3+F4, F1+F2+F4, F1+F3, F1+F4, F2+F3, F2+F3+F4, F2+F4, F3+F4, F3+F4+F1) were analysed using the four machine learning algorithms.

Ten fold cross validation was used to evaluate the models' performances based on percent correct (percentage of all datasets that are classified correctly) and Kappa statistic (measure of the agreement between predicted and observed categorization, while correcting for agreement that happens by chance.

3.4 Evaluation of the Experiments

Tables in Appendix 1 show the percent correct and kappa statistic values derived for each of the datasets in our experiment. The results are presented successively for Naive Bayes, SMO, J48 and multilayer perceptron. It could be observed from the tables that the percent correct values appear to be highest for Dataset 3 while Kappa statistics appear to be highest for Dataset 2. This observation cuts across virtually all features sets and classifiers. This implies that classifiers were better able to classify Dataset 3 correctly compared to other datasets while classifications achieved in Dataset 2 gave better agreement between predicted and observed categorization having corrected for agreement that happened by chance. Worthy to be noted is the result of SMO in Dataset 3, although the percent correct values were relatively high, Kappa statistics were all zero. Lack of coherence in the directions of the two performance measures led us to using the product of the two measures (percent correct and kappa statistic) as a basis for comparing models' performances.

This decision to use the product was informed by the theory of Dimensional Analysis which is a problem-solving method that uses the fact that any number or expression can be multiplied by one without changing its value. One can only meaningfully add or subtract quantities of the same type but can multiply or divide quantities of different types. When two measurements are multiplied together the product is of a type depending on the types of the measurements. This analysis is routinely applied in physics and it is an engineering tool that is widely applied to numerous engineering problems for designing and testing all types of engineering and physical systems ([18]; [19]). The result of the dot products of the two measures is as presented in Appendix 2. The table in Appendix 2 presents the performances of our models taking into consideration the two performance measures. We consider this table more representative of the models' performances because it combines the strengths and weaknesses of the two performance measures. Answers to research questions will, therefore, be based on the content of this table.

4. RESULTS AND DISCUSSION

Research Question 1: Which feature set type maximizes the effectiveness of profiling the country of affiliation of writers of online messages?

Figure 1 is a derivative of the table in Appendix 2, it shows the product of percent correct and kappa statistic values derived for the feature set types in our experiment. The results are presented successively for Naive Bayes, SMO, J48 and Neural Network.

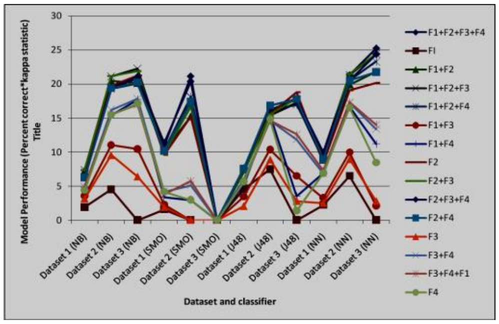

Figure 1: Comparison of feature sets performances

Across all the three datasets, the feature set that combined all feature types (F1+F2+F3+F4) performed best. This is followed by (F2+F4), (F2+F3+F4) and (F1+F2+F3), while the performance of F1 was the least. Our result shows that inclusion of all features from all the four types (lexical, syntactic, idiosyncrasies and content specific) produced the most effective model. Again the result was consistent with those of [20] and[2] and [1] pg 365 who reported that combining feature types in their studies gave a better result. Using vocabulary richness only produced the poorest result probably because of the short length of online messages in the study.

Research Question 2: Which classification scheme maximizes the effectiveness of profiling the country of affiliation of writers of online messages?

Figure 2 shows the relative performances of the four classifiers across all feature types (F1+F2+F3+F4) and datasets.

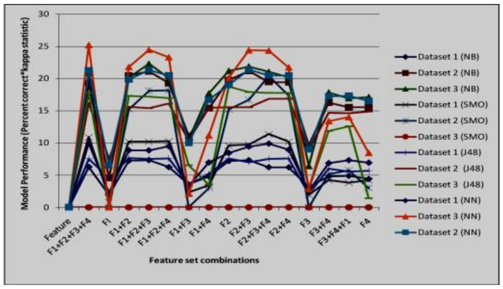

Figure 2: Relative performances of the four classifiers across all feature and data sets.

Neural Network (multilayer perceptron) performed best when compared to the other three classifiers. Its performance was particularly the highest on the feature set (F1+F2+F3+F4) contained in our two-class option dataset (Dataset 3). Most previous studies considered SVM most appropriate in authorship attribution (though most times without carrying out a prior experiment). [1] however, reported that there were no significant performance differences between SVM and neural networks. It could be observed that SVM implementation (SMO) outperformed the other three classifiers when the texts contained their natural class labels (Dataset 1) and performed most terribly on Dataset 3. This corroborates the submission of [15] that no single machine learning scheme is appropriate to all data mining problems because real datasets vary and to obtain accurate models, the bias of the learning algorithm must match the structure of the domain. Meaning that the structure of our Dataset 3 is most amenable to neural network than any of the other machine learning schemes (Naive Bayes, SMO, J48) in our study. Worthy of note also is the usefulness of our application of the dimensional analysis principle which informed the multiplication of the two performance measures in our study. For example, if our comparison had been based on percent correct (in Appendix 1) only, we might have erroneously rated the performance of SMO relatively high on Dataset 3.

Research Question 3: Which class labelling option maximizes the effectiveness of profiling the country of affiliation of writers of online messages?

Fig. 3 shows the percent correct values derived for each of the datasets in our experiment using the most precise classification scheme (Neural Network) and all feature sets (F1+F2+F3+F4) only. The results are presented successively for Naive Bayes, SMO, J48 and Neural Network.

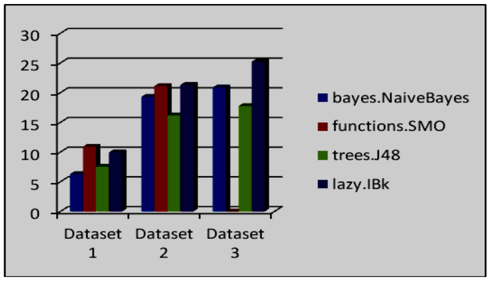

Figure 3: Column Chart of Classifier Performances with Varied Class Labelling Options

The figure shows that the dataset having two class options (Dataset 3) performed best followed by the one having three class options (Dataset 2) and lastly the one having the instances labelled naturally, having five classes (Dataset 1). The result is consistent with those of [3] and [1] that reported that authorship attribution success improves with reduction in the number of authors or author classes. In the specific however, the present result shows that if we can reduce an authorship profiling problem to a two-class one, we can get an appreciable improvement in the effectiveness of authorship profiling task.

Research Question 4: What is the performance of the resultant model in classifying electronic messages to writers' countries of affiliation?

Using the *TrainTestSplitMaker* component of WEKA's knowledge flow interface to evaluate the performance of our model in classifying electronic messages to writers' countries of affiliation. Separate two-class label file was created for each country, resulting in a dataset for each country, where all attributes except the class attribute were the same. The class attribute for a particular country had instances labelled either as 'the country name' such as (Nigeria, Ghana, Cameroon) or as 'non country name' such as (Non-Nigeria, Non-Ghana, Non-Cameroon). Tables 4 shows the effectiveness of profiling authors' countries of affiliation by the resultant model.

Table 4: Effectiveness of Profiling Authors' Countries of Affiliation

Country	Percent Correct	Kappa Statistics	PC*KS
Nigeria	75.80	0.34	25.95
Cameroon	73.80	0.10	7.68
Ghana	78.40	0.27	21.54
Liberia	88.20	0.04	3.23
Sierra Leone	70.80	0.28	19.59

PC*KS denotes Percent correct* Kappa statistics

Application of our optimization method resulted in a remarkable improvement in the profiling of each country from the others. The study showed that we could achieve a percent correct ranging between 70.8% and 88.2% at Kappa statistics ranging between 0.04 and 0.34 compared to the highest possible percent correct value of 43.8% at kappa statistics of 0.26% if our method was not applied. This however is a trade-off on the efficiency of the profiling process because we needed to create separate labels for the class attribute. The extent of improvement in model performance however can be said to outweigh the additional effort. The detailed performance of the model is as shown in Table 5.

Table 5: Detailed Prediction Performance of the Resultant Model

	TP Rate	FP Rate	Preci-sion	Re-call	F-score	ROC Area
Nigerian	0.380	0.080	0.671	0.380	0.485	0.721
Non-Nigerian	0.920	0.620	0.776	0.920	0.842	0.721
Weighted Average	0.758	0.458	0.744	0.758	0.735	0.721
Cameroon	0.299	0.182	0.230	0.299	0.260	0.652
Non-Cameroon	0.818	0.701	0.865	0.818	0.841	0.652
Weighted Average	0.738	0.621	0.767	0.738	0.751	0.652
Ghanaian	0.333	0.092	0.5	0.333	0.400	0.671
Non-Ghanaian	0.908	0.667	0.832	0.908	0.868	0.671
Weighted Average	0.784	0.543	0.760	0.784	0.767	0.671
Liberian	0.036	0.013	0.250	0.036	0.063	0.671
Non-Liberian	0.987	0.964	0.892	0.987	0.937	0.671
Weighted Average	0.882	0.859	0.822	0.822	0.841	0.671
Sierra-Leonean (SL)	0.582	0.256	0.39	0.582	0.467	0.748
Non SL	0.744	0.418	0.863	0.744	0.799	0.748
Weighted Average	0.708	0.383	0.759	0.708	0.726	0.748

The resultant model performed well when we consider the weighted averages of the performance measures of each dataset. It could however, be observed that the model was better at identifying texts that were not from the country as against those that were from the country in each case. It could also be observed that the performance of the model in predicting each country's texts vary directly with the number of each country's texts in the study corpus. The best performance was achieved in profiling Nigerian electronic texts from Non Nigeria texts, followed by that of Sierra Leone and then Ghana. Thus, it could be deduced that performance of our model could be much improved with bigger sub-corpora sizes.

5. CONCLUSION

The study through experiments sought the number of class options, feature set types and machine learning scheme that maximize the effectiveness of identifying the countries of affiliation of authors of online messages composed in English language. The online messages in our corpus were collected from online forums of five African countries with average length of 52 to 102 words. Using a product of percent correct and kappa statistics as our bases for model justification, the experiment showed that we achieved the most effective model when all feature set types, contained in a two-class dataset was analysed with the neural network (multilayer perceptron) machine learning scheme. Application of the parameters of the most effective model (derived from the experiment) to profiling the countries of affiliation of authors of the online messages resulted in about a hundred percent improvement in effectiveness.

The study achieved greater effectiveness but with a trade-off on efficiency. We look forward to having a model that **can** maximize both effectiveness and efficiency in profiling the authorship of online messages, and this constitutes a need for further studies. This approach in its present state can be very appropriate if a group is suspected and the purpose of authorship attribution is to affirm one's thought about the suspect's group of affiliation.

6. REFERENCES

[1] Zheng, R., Li, J., Chen, H. and 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.

[2] Koppel, M., Schler, J. and Zigdon, K. 2005. Automatically determining an anonymous author’s native language. *Lecture Notes in Computer Science (LNCS) 3495*. Eds. Kantor, P.B., Muresan, G., Roberts, F., Zeng, D.D. and Wang, F : ISI 2005, Berlin: Springer-Verlag. 209 – 217.

[3] Luyckx, K. and Daelemans, W. 2008. Authorship attribution and verification with many authors and limited data. In: *Proceedings of the 22nd International Conference on Computational Linguistics* held in Manchester from 18-22 August 2008. 513–520.

[4] Koppel , M., Schler , J., Argamon, S. and Messeri, E. 2006. Authorship attribution with thousands of candidate authors. In: *Proceedings of the 29th annual international ACM SIGIR (Special Interest Group on Information Retrieval) conference on research and development in information retrieval.* Aug. 6-11 2006, Seattle, Washington, USA.

[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] Argamon, S., Koppel,M., Pennebaker, J.W. and Schler, J. 2009. Automatically Profiling the Author of an Anonymous Text. *Communications of the ACM* 52(2). 119-123.

[7] De Vel, O. , Anderson, A., Corney, M. and Mohay, G. 2001. Mining E-mail Content for Author Identification Forensics. *Special Interest Group on Management of Data (ACM SIGMOD) Record* 30(4). 55-64.

[8] Juola, P. 2007. Future trends in authorship attribution. *International Federation for Information Processing* 24(2). 119-132.

[9] Iqbal, F., Hadjidj, R., Fung, B.C.M and Debbabi, M. 2008. A novel approach of mining write-prints for authorship attribution in e-mail forensics. *2008 Digital Forensic Research Workshop.* Elsevier Ltd. Retrieved Nov. 16, 2009, from www.elsevier.com/locate/diin. 2008.05.001

[10] Holmes, D.I. 2003. Stylometry and the civil war: the case of the Pickett letters. *CHANCE* 16(2) 18-25.

[11] Binongo, J.N.G. 2003. Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution. *CHANCE* 16(2) . 9-17.

[12] Binongo, J.N.G. and Smith M.W.A. 1999. The application of principal component analysis to stylometry. *Literary and Linguistic Computing* 14(4). 445-466.

[13] Abbasi, A. and Chen, H. 2006. Visualizing authorship for identification. *Lecture Notes in Computer Science (LNCS) 3975*. Eds. Mehrotra, S., Zeng, D.D., Chen, H., Thuraisingham, B., Wang, F. Berlin: Springer-Verlag. 60–71.

[14] Abbasi, A. and Chen, H. 2008. Writeprints: a stylometric approach to identity level identification and similarity detection in cyberspace. ACM Transactions on Information Systems. 26 (2). doi: 10.1145/1344411.1344413.

[15] Witten, I.H. and Frank, E. 2005. Data mining: practical machine learning tools and techniques. 2nd ed. USA: Morgan Kaufmann publishers.

[16] Kujore, O. 1985. English usage: some notable Nigerian variations. 1-112. Nigeria: Evans Brothers Nigeria Publishers Limited.

[17] Jowitt, D. 1991. *Nigerian English usage: An Introduction*. 1-277. Nigeria: Longman.

[18] Balaguer P 2013 Application of Dimensional Analysis in Systems Modeling and Control Design, The Institution of Engineering and Technology;

[19] Szirtes T 2007 Applied Dimensional Analysis and Modeling. Elsevier/Butterworth-Heinemann Amsterdam; New York.

[20] Ma, J., Teng, G., Zhang, Y., Li, y. and Li Y (2009) A Cybercrime Forensic Method for Chinese Web Information Authorship Analysis. In: PAISI 2009, LNCS 5477 pp. 14-24. H. Chen et al. (Eds.). Springer-Verlag Berlin Heidelberg

[21] Opesade, A., Adegbola, T., & Tiamiyu, M. (2013). Comparative Analysis of Idiosyncrasy, Content and Function Word Distributions in the English Language Variants of Selected African Countries. International Journal of Computational Linguistics Research Vol. 4(3) pp.130-143.

Appendix 1: Experiment Result

	Naive Bayes						SMO					
	Dataset 1		Dataset 2		Dataset 3		Dataset 1		Dataset 2		Dataset 3	
Feature	PC	KS	PC	KS	PC	KS	PC	KS	PC	KS	PC	KS
F1+F2+F3 +F4	34.96	0.18	58.49	0.33	67.34	0.31	43.65	0.25	62.12	0.34	71.09	0.00
FI	31.87	0.06	50.14	0.09	71.09	0.00	31.52	0.05	49.52	0.00	71.09	0.00
F1+F2	36.29	0.20	60.11	0.34	69.41	0.29	42.53	0.24	58.54	0.26	71.09	0.00
F1+F2+F3	36.77	0.20	60.31	0.35	69.78	0.32	42.48	0.24	60.33	0.30	71.09	0.00
F1+F2+F4	34.80	0.18	58.39	0.33	66.97	0.30	42.84	0.24	60.54	0.30	71.09	0.00
F1+F3	32.37	0.11	52.63	0.21	69.77	0.15	32.73	0.07	49.48	0.00	71.09	0.00
F1+F4	32.10	0.15	55.01	0.28	65.59	0.27	34.07	0.10	49.86	0.06	71.09	0.00
F2	36.06	0.20	59.77	0.33	70.74	0.30	41.83	0.23	58.48	0.26	71.09	0.00
F2+F3	36.69	0.20	60.51	0.35	70.63	0.31	42.32	0.23	59.57	0.28	71.09	0.00
F2+F3+F4	34.64	0.18	58.86	0.33	67.93	0.31	43.76	0.26	61.72	0.33	71.09	0.00
F2+F4	34.65	0.18	58.73	0.33	67.46	0.30	42.50	0.24	59.89	0.29	71.09	0.00
F3	31.79	0.10	53.24	0.18	71.46	0.09	31.86	0.06	49.50	0.00	71.09	0.00
F3+F4	32.43	0.15	55.97	0.29	66.14	0.27	35.05	0.12	51.58	0.10	71.09	0.00
F3+F4+F1	32.64	0.15	55.44	0.28	65.23	0.26	34.50	0.11	52.17	0.11	71.09	0.00
F4	31.53	0.14	55.37	0.28	65.93	0.26	34.88	0.12	49.75	0.06	71.09	0.00

PC = Percent Correct KS = Kappa Statistic

Experiment Result Continued

	Tree (J48)						Multilayer Perceptron (Neural Network)					
Feature	Dataset 1		Dataset 2		Dataset 3		Dataset 1		Dataset 2		Dataset 3	
F1+F2+F3+F4	PC	KS	PC	KS	PC	KS	PC	KS	PC	KS	PC	KS
FI	35.11	0.13	49.92	0.15	71.09	0.00	31.76	0.07	50.01	0.13	71.09	0.00
F1+F2	38.32	0.20	55.59	0.28	72.10	0.24	40.28	0.22	60.31	0.33	72.73	0.30
F1+F2+F3	37.66	0.20	55.05	0.28	71.74	0.24	40.16	0.22	61.05	0.35	74.16	0.33
F1+F2+F4	37.88	0.20	55.65	0.29	70.80	0.24	41.22	0.23	60.32	0.34	72.91	0.32
F1+F3	31.58	0.11	51.93	0.20	72.37	0.09	32.73	0.10	52.52	0.19	70.39	0.03
F1+F4	34.61	0.15	55.34	0.28	70.43	0.05	38.69	0.18	57.62	0.29	69.86	0.16
F2	37.87	0.20	55.57	0.28	72.20	0.26	40.18	0.21	59.50	0.32	71.90	0.28
F2+F3	36.97	0.19	55.41	0.28	71.63	0.25	41.02	0.23	61.19	0.35	73.98	0.33
F2+F3+F4	37.76	0.20	56.08	0.30	71.11	0.25	41.22	0.24	60.37	0.34	73.81	0.33
F2+F4	37.84	0.20	56.18	0.30	71.14	0.25	40.60	0.22	60.30	0.34	72.50	0.30
F3	29.48	0.07	52.54	0.17	70.59	0.04	31.64	0.08	52.90	0.17	70.44	0.04
F3+F4	35.49	0.17	54.41	0.27	69.78	0.17	38.06	0.18	57.67	0.30	70.46	0.19
F3+F4+F1	34.71	0.16	54.12	0.27	69.94	0.18	38.48	0.19	57.69	0.30	70.12	0.20
F4	35.41	0.16	55.06	0.27	70.09	0.02	38.46	0.18	57.03	0.29	70.49	0.12

PC = Percent Correct KS = Kappa Statistic

Appendix 2: Products of Percent Correct and Kappa Statistics

	Naive Bayes (PC*KS)			SMO (PC*KS)			J48 (PC*KS)			Multilayer Perceptron (PC*KS)		
Feature	Dataset 1	Dataset 2	Dataset 3	Dataset 1	Dataset 2	Dataset 3	Dataset 1	Dataset 2	Dataset 3	Dataset 1	Dataset 2	Dataset 3
F1+F2+F3+F4	6.29	19.30	20.88	10.91	21.12	0.00	7.55	16.17	17.77	9.95	21.29	25.22
FI	1.91	4.51	0.00	1.58	0.00	0.00	4.56	7.49	0.00	2.22	6.50	0.00
F1+F2	7.26	20.44	20.13	10.20	15.22	0.00	7.66	15.57	17.30	8.86	19.90	21.82
F1+F2+F3	7.35	21.11	22.33	10.20	18.10	0.00	7.53	15.41	17.22	8.84	21.37	24.47
F1+F2+F4	6.26	19.27	20.09	10.28	18.16	0.00	7.58	16.14	16.99	9.48	20.51	23.33
F1+F3	3.56	11.05	10.47	2.29	0.00	0.00	3.47	10.39	6.51	3.27	9.98	2.11
F1+F4	4.82	15.40	17.71	3.41	2.99	0.00	5.19	15.50	3.52	6.96	16.71	11.18
F2	7.21	19.72	21.22	9.62	15.20	0.00	7.57	15.56	18.77	8.44	19.04	20.13
F2+F3	7.34	21.18	21.90	9.73	16.68	0.00	7.02	15.51	17.91	9.43	21.42	24.41
F2+F3+F4	6.24	19.42	21.06	11.38	20.37	0.00	7.55	16.82	17.78	9.89	20.53	24.36
F2+F4	6.24	19.38	20.24	10.20	17.37	0.00	7.57	16.85	17.79	8.93	20.50	21.75
F3	3.18	9.58	6.43	1.91	0.00	0.00	2.06	8.93	2.82	2.53	8.99	2.82
F3+F4	4.86	16.23	17.86	4.21	5.16	0.00	6.03	14.69	11.86	6.85	17.30	13.39
F3+F4+F1	4.90	15.52	16.96	3.80	5.74	0.00	5.55	14.61	12.59	7.31	17.31	14.02
F4	4.41	15.50	17.14	4.19	2.99	0.00	5.67	14.87	1.40	6.92	16.54	8.46

PC*KS denotes Percent correct* Kappa statistic