Forensic Investigation of Linguistic Sources of Electronic Scam Mail: A Statistical Language Modelling Approach

Adeola O Opesade[1], **Mutawakilu A Tiamiyu**[1], **Tunde Adegbola**[2]
[1] Africa Regional Centre for Information Science, University of Ibadan, Nigeria
[2] African Languages Technology-Initiative (ALT-I), Ibadan, Nigeria
E-mail: morecrown@gmail.com, mutatiamiyu@yahoo.com, taintransit@hotmail.com

Abstract

Electronic handling of information is one of the defining technologies of the digital age. These same technologies have been exploited by unethical hands in what is now known as cybercrime. Cybercrime is of different types but of importance to the present study is the 419 Scam because it is generally (yet controversially) linked with a particular country - Nigeria. Previous research that attempted to unravel the controversy applied the Internet Protocol address tracing technique. The present study applied the statistical language modelling technique to investigate the propensity of Nigeria's involvement in authoring these fraudulent mails. Using a hierarchical modelling approach proposed in the study, 28.85% of anonymous electronic scam mails were classified as being from Nigeria among four other countries. The study concluded that linguistic cues have potentials of being used for investigating transnational digital breaches and that electronic scam mail problem cannot be pinned down to Nigeria as believed generally, though Nigeria could be one of the countries that are prominent in authoring such mails.

Keywords: digital forensics, 419 scam, statistical language modelling

1. Introduction

1.1 Background to the Study

Electronic handling of information is one of the defining technologies of the digital age. The age is characterized by the application of computer technology as a transformative tool to enhance effectiveness and efficiency in personal, commercial, educational, governmental, and other facets of modern life (Reith, Carr, and Gunsch, 2002). However, alongside the benefits derivable from the ability to automatically manage so much information are threats to communications and transactions conducted electronically. Just as ICT provides new opportunities to operate and expand one's presence and reach, it also presents opportunities for those with criminal intentions, leaving its numerous users highly exposed to the threat of cyber attack and cybercrime (Choo, 2011).

The term cybercrime encompasses traditional crimes such as fraud, scam, theft, forgery, harassment, blackmail, embezzlement, in as much as the computer or its network is employed for perpetration; it also involves a host of new criminal activities which cannot be perpetrated but with computer or its network, such as spam, denial of service attacks and distribution of viruses (Torosyan, 2003; Alshalan, 2006).

The advance fee fraud (also known as Internet Scam, Nigerian 419 Scam or *Yahoo-Yahoo*) is a type of internet fraud that is generally believed to have originated from Nigeria and perpetrated mostly by Nigerians. This stance has however, been faced with much controversy. Perceptions on the extent of Nigeria's involvement in the current status of advance fee fraud can be categorised into three; while the first two views are distinctly opposite, the third stands somewhere in between them. Holding the first view are sources that believe that Nigeria is the main, if not the sole source, of this fraud (State Department Publication, 1997). Holding the second view are sources that refute the fact that Nigeria is as prominent in cybercrime as portrayed by those other reports. They argued that the increased consciousness about crime and the 419 scams in particular is only a motive to criminalize the Nigerian state (Bayart, Ellis and Hibou, 1999). Holding the third view are those of the opinion that the term "Nigerian Advance Fee Fraud" is only partially accurate and the problem is truly one of international dimension with victims and offenders being located across the globe (Smith, Holmes and Kaufmann, 2001).

Previous studies that focused on unravelling the controversy used e-mail header (Edelson, 2003) and Internet Protocol (IP) address tracing technique (Longe and Osofisan, 2011). Effective as their methods are however, they are constrained by geographical boundaries and cannot detect the nationality of a scammer who sends an electronic scam mail from outside the shores of his own country. The methods are also not effective against some computer security frauds such as phishing, spoofing and masquerading. There is therefore, a need for new methodologies for investigating sources of such fraudulent e-mail.

Dyrud (2005) analysed ninety-three (93) electronic scam letters (which he called Nigerian scam letters) which he received over a period of 10 months. While analysing one of the letters he commented:

> *Some are ludicrously transparent, such as John Kredoski's November, November 13, 2004 e-mail. He lives in Reno, has suffered a heart attack, and has chosen to write to me because I am "a fellow American." "Do not," he cautions, "associates this letter as numerous letters we receive from Africa. If you need my passport to prove my identity, I will send it to you." For a native-born American, his syntax and vocabulary are curiously similar to his Nigerian counterparts.* (Dyrud, 2005: p. 7).

Although Dyrud (2005) did not present any objective means of deducing that the received mails were Nigerian, he was able to differentiate the language usage of the writer from his own American style, despite the writer's claim of being an American. Exploitation of linguistic cues can therefore, aid in further detection of sources of anonymous

electronic mails; this is because a writer's natural language is less susceptible to deception.

Language is an intricate system of structural components for encoding and decoding information (Cruz-Ferreira and Abraham, 2006). All human languages exhibit a great deal of internal variations in terms of a specific set of linguistic items or human speech patterns such as words, idiosyncratic, structural or grammatical features which can uniquely associate with some external factors such as geographical area or social group, time, situation, genre, subject, author amongst others (Adetugbo, 1979; Banjo, 1979; Wardhaugh, 1994). The phenomenon of Standard Nigerian Spoken English was proposed by Banjo in 1971 (Jowiit, 1991). According to Adetugbo (1979) the fact that English language in Nigeria is regarded as a dialect of English means that it preserves the same common core items of vocabulary and structure found in other varieties of English everywhere else, however, the implications of spatial dialect differentiation, the mode of acquisition of the English language in Nigeria, interference of native languages on English in its second language situation in Nigeria and the non-equivalence of the social and cultural environment in Nigeria with that of any other English language speaking community, demanded that communicative competence in English in Nigeria cannot be the same as for any other variety of English.

The features of the Nigerian variant of the language have been well researched in the humanities, for example, on the level of morphology, due to the influence of the dominant local language or mother tongue the spoken English shows variations in accent. On the lexical level, the problem with the use of the article (definite and indefinite) is endemic (Egbe, 1979). Kujore (1985) and Jowiit (1991) composed specific variations alongside the standard British English in order to show the growing divergence between the Nigerian usage and British usage of the language. Opesade, Adegbola and Tiamiyu (2013) validated quantitatively the presence of English language variants of five African countries, Cameroon, Ghana, Liberia, Nigeria and Sierra-Leone. If Nigerian and other countries' variants of the English language exist, then this could be a means of identifying Nigerian texts written in the English language from non-Nigerian text written in the other variants of same language.

1.2 Objective of the Study

The objective of the study was to establish and apply statistical language models for the detection of electronic scam mail likely to be of Nigerian origin, based on linguistic cues. To achieve the objective of the study, the following sub-objectives were pursued:

1. To determine a precise model for separating electronic texts of Nigerian linguistic origin from those of other countries (Cameroon, Ghana, Liberia, Sierra-Leone).
2. 2. To apply the model to classify electronic scam mail into countries of origin based on linguistic cues.

1.3 Research Questions

1. What is the performance of the (NigGh_Other) model that classifies Nigerian and Ghanaian texts from those from the other three countries (Liberia, Cameroon and Sierra Leone)?
2. What is the performance of the (Nig_Gh) model that classifies Nigerian and Ghanaian texts only?
3. What percentage of electronic scam mail are classified to be Nigerian based on linguistic cues.

3. Research Methodology

3.1 Research Design

The research was carried out using modelling and simulation methodologies. In this study, online English texts of five African countries were modelled statistically. Out of a population of all English speaking countries in West Africa including Cameroon, the domains of estimation selected for this study were Nigeria, Ghana, Liberia, Cameroon and Sierra-Leone. Nigeria is the country of focus in this investigation, while Ghana, Liberia, Cameroon and Sierra-Leone are selected for the purpose of comparing their English language usages to that of Nigeria as a means of providing alternative candidate countries in the attribution of the electronic scam mails.

3.2 Target Populations

To achieve the objective of this study, two populations were examined. They were:

i. All readers' comments posted on the fifty–two (52) pages contained (as of November 2011) in *www.topix.com* websites of the five countries selected for the study and the additional four pages posted (as of February 2012) in the same websites.
ii. All eight hundred and seventy-three (873) electronic scam mails available in scam archives such as scam bating sites as of 14th March 2011. Scam baiting, also known as counter scamming, is the practice of feigning interest in a fraudulent scheme in order to manipulate a scammer.

3.3 Sampling Techniques for Composition of Study Corpora

A multistage sampling technique was employed for textual data collection. The sampling procedure for each of the two populations is as presented below:

3.3.1 Sampling Technique for the www.topix.com Texts

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 researchers 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.

Another set of one hundred and fifteen (115) texts were collected from the country forums of *www.topix.com* in February 2012. A complete enumeration of all texts on the

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

Table 1: Training Data Set

Country's forum website	No. of pages	Pages selected	No. of selected texts
www.topix.com/forum/world/nigeria	35	1	30
www.topix.com/forum/world/ghana	10	1	45
www.topix.com/forum/world/liberia	4	1	10
www.topix.com/forum/world/cameroon	4	1	15
www.topix.com/forum/world/sierra-leone	4	1	15
Total no. of Texts			115

Table 2: Validation Data Set

first page of each country's website was carried out; all texts which met the conditions of at least twenty words with evidence of affiliation to the relevant country, but which had not been selected previously were selected. The one hundred and fifteen texts selected (as the validation set) are as shown in Table 2.

3.3.2 Sampling Technique for the Electronic Scam Data

A multistage sampling technique was also used for this population. The scam baiting sites selected for the study, the total number of electronic scam mails baited in each and the sample size are shown in Table 3.

Scam Baiter's website	No. of scam mails	No. selected
www.419eater.com	133	53
www.419hell.com	30	12
www.419baiter.com	19	8
www.scamorama.com	598	239
www.ebolamonkeyman.com	24	10
www.monkeyspit.net/inbox/	7	3
www.sweetchillisauce.com/nigeria.html	41	16
www.whatsthebloodypoint.com	11	4
http://419.bittenus.com	10	4
Total	873	349

Table 3: Scam Baiters' Websites and Sample Sizes (Date: 14/3/2011)

3.4 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 predetermined linguistic features from the sampled texts using computer codes written by the researchers in the Python2 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.

Extracted features were syntactic features comprising the twenty (20) most frequent function words in the *topix.com* corpus, twenty (20) most frequent function words in the scam mail corpus (out of which seventeen most frequent function words that are common to *topix.com* corpus and scam mail corpus were used for analysis). Idiosyncratic features comprising adverb-verb part of speech (POS) bigram distribution and article omission/inclusion distribution. Structural features comprising lexical diversity; and content specific features twenty (20) most frequent noun, adjective, verb and adverb unigrams in the *topix.com* corpus, twenty (20) most frequent noun, adjective, verb and adverb unigrams in the scam mail corpus (out of which thirteen most frequent content words that are common to *topix.com* corpus and scam mail corpus were used for analysis).

The decision to extract twenty most frequent features (function word, noun, adjective, verb and adverb unigrams) was a result of a prior experiment by the researchers 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.

3.5 Data Analysis

Although academic researchers have tended to favour the use of Support Vector Machines (SVMs), there is still a division on the choice of the best machine learning method, particularly in the anti-spam community (Sculley and Wachman, 2007). Also in the words of Witten and Frank (2005):

> *Experience shows that no single machine learning scheme is appropriate to all machine learning 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*

	TP Rate	FP Rate	Precision	Recall	F-score	ROC Area
NotNGGH	0.630	0.279	0.689	0.630	0.659	0.68
NGGH	0.721	0.370	0.665	0.721	0.692	0.68
Weighted Average	0.676	0.325	0.677	0.676	0.675	0.68

Table 4: Detailed Performance of the NigGh_Others Model

> *domain; machine learning like data mining is an experimental science.* (Witten and Frank: p. 365).

Thus, an experiment was carried out to determine the most precise machine learning algorithm for the study data set using the experimenter view of the open source Waikato Environment for Knowledge Analysis (WEKA) data mining software. Based on the result of the experiment, the Instance Based K-nearest (IBK) Neighbour algorithm was found to be the most precise in terms of accuracy and kappa statistic. Hence, the Instance Based K-nearest (IBK) Neighbour algorithm was used for the classification of the study's data sets based on a proposed hierarchical modelling approach.

3.5 Proposed Hierarchical Modelling Approach

This is a recursive approach proposed by the researchers based on the fact that human languages (and variants of a language) are not exclusively independent of one another (Bird, Klein and Loper, 2007), and on our understanding of the linguistic similarities among the English language variants of the five African countries in the study as revealed in Opesade, Adegbola and Tiamiyu (2013). Although each variant of the English language in the five countries differed from one another, there were still similarities among different pairs. For example, the Nigerian variant was found to be closest to the Ghanaian, followed be Sierra Leonean and Liberian and lastly Cameroonian.

In the proposed approach, binary classification tasks were performed such that the Nigerian and Ghanaian sub-corpora were first treated as one and separated from the other three, after which the two sub-corpora were classified into their respective country classes (Nigeria or Ghana). This was done because of the discovered similarity between the Nigerian and Ghanaian variants.

4. Presentation of Results

Research Question 1: What is the performance of the (NigGh_Others) model that classifies Nigerian and Ghanaian texts from those from the other three countries (Liberia, Cameroon and Sierra Leone)?

Based on 10-fold cross validation repeated ten times, the model's accuracy and kappa statistic were 67.619% and 0.3518 respectively. The detailed performance of the Instance based (IBK) model that classified training data set of Nigerian and Ghanaian linguistic origin from those of the three other countries is as shown in Table 4.

The precision values for the Nigeria/Ghana class (NGGH), others (NotNGGH) were 0.665 and 0.689 respectively, with a weighted average of 0.677. The model's recall values for the Nigeria/Ghana (NGGH), others (NotNGGH) were 0.721 and 0.63 respectively and with a weighted average of 0.676. The F-measures for the Nigeria/Ghana class (NGGH) and others (NotNGGH) were 0.692 and 0.659 respectively. The confusion matrix of the classification (Table 5) shows that four hundred and fifty-nine (459) out of seven hundred and twenty-eight (728) others (NotNGGH) texts were classified correctly, giving an accuracy value 63.0%.

NotNGGH	NGGH	Classified as	Accuracy (%)
459	269	NotNGGH	63.0
207	**535**	NGGH	72.1

Table 5: Confusion Matrix of the NigGh_Others Model using Topix_training Set

Five hundred and thirty-five (535) out of seven hundred and forty-two (742) Nigerian/Ghanaian (NGGH) texts were classified correctly, giving an accuracy value 72.1%. The NigGh_Others model was thereafter used to predict the sources of two hundred and fifty-three anonymous electronic scam mails as Nigerian/Ghanaian or other (Liberian, Cameroonian, Sierra-Leonean). The prediction is as shown in Table 6.

Predicted Class	Count	Percentage classified (%)
Nigerian/Ghana	119	47.04
Others (Liberia, Cameroon, Sierra Leone)	134	52.96
Total	**253**	**100**

Table 6: NigGh_Others model's Prediction of Electronic Scam Mails Classes

About forty-seven percent (47.04%), that is, one hundred and nineteen (119) of the scam mails were classified as being Nigeria/Ghana while the remaining 52.96% was classified as being Others (Liberia, Cameroon, Sierra Leone).

Research Question 2: What is the performance of the model (Nig_Gh model) that classifies Nigerian and Ghanaian texts only?

Texts from countries other than Nigeria and Ghana were removed from the training and Topix_test set, also scam mails classified as being from countries other than Nigeria and Ghana were excluded. The remaining texts were from two classes, that is, the Nigerian and Ghanaian class only. Accuracy and kappa statistic of the model were 67.66% and 0.3359 respectively. The detailed performance of the Nig_Gh model based on 10-fold cross validation is as shown in Table 7.

The precision values for the Ghanaian and Nigerian sub-corpora were 0.626 and 0.712 respectively, with a weighted average precision of 0.675. The model's values of recall for the Ghanaian and Nigerian sub-corpora were 0.603 and 0.732 respectively, with a weighted average recall of 0.677.

	TP Rate	FP Rate	Precision	Recall	F-score	ROC Area
NG	0.732	0.397	0.712	0.732	0.722	0.658
GH	0.603	0.268	0.626	0.603	0.614	0.658
Weighted Average	0.677	0.342	0.675	0.677	0.676	0.658

Table 7: Detailed Prediction Performance of the Nig_Gh Model

The F-measures for the Ghanaian (GH) and Nigerian (NG) classes were 0.614 and 0.722 respectively. Table 8 shows the confusion matrix of the classification (Nig_Gh) model.

Nigeria	Ghana	Classified as	Accuracy (%)
311	114	Nigeria	73.2
126	**191**	Ghana	60.3

Table 8: Confusion Matrix of the Nig_Gh Model on Training Set

The confusion matrix shows that three hundred and eleven (311) out of four hundred and twenty-five (425) Nigerian texts were classified correctly, giving an accuracy value of 73.2% for the Nigerian sub-corpora; while one hundred and ninety-one (191) out of three hundred and seventeen (317) Ghanaian texts were classified correctly, giving 60.3% accuracy value. The Nig_Gh model was thereafter used to predict the sources of the one hundred and nineteen (119) anonymous electronic scam mails classified as Nigerian/Ghanaian in Table 5. The prediction is as shown in Table 9.

Predicted Class	Count	Percentage (%)
Nigerian	73	61.3
Ghanaian	46	38.7
Total	**119**	**100**

Table 9: Nig_Gh model's Prediction of Electronic Scam Mails Classes

Out of the one hundred and nine (119) scam mails classified to be from a combination of Nigerian and Ghanaian texts in Table 5, the Nig_Gh model predicted seventy-three (73) to be from Nigeria and forty-six (46) to be from Ghana.

Research Question 3: What percentage of electronic scam mail are classified to be Nigerian relative to the other four countries, based on linguistic cues?

Table 10 shows the percentage of scam mail classified to be from Nigeria relative to the other four countries. Seventy-three (73) out of the 253 scam mail in the study were classified to be from Nigeria. This shows that about 28.85% of the scam mails are classified to be from Nigeria based on linguistic cues.

4.1 Model Validation

The two-level model approach was validated on the one hundred and fifteen (115) *topix* validation data set. The accuracy and kappa statistic were 60.87% and 0.2153 respectively for the NigGh_Others model and 70.67 % and 0.3774 respectively for the Nig_Gh model**.** The weighted F-Measures were 0.618 and 0.716 for the NigGh_Others and Nig_Gh models respectively.

Predicted Class	Count	Percentage (%)
Nigerian	73	28.85
Ghanaian	46	18.18
Others (Cameroon, Liberia, Sierra Leone)	134	52.96
Total	**253**	**100**

Table 10: Nig_Gh Model's Prediction of Electronic Scam Mails' Classes

5. Discussion of Findings on the Research Questions

Using the two-level modelling approach, 28.85% of anonymous electronic scam mails were classified as being from Nigeria. This showed that among five countries, about one-third of the two hundred and fifty-three electronic scam mail in the study were predicted to be from Nigeria. This result is mid-way between the findings of Longe and Osofisan (2011) whose analysis of electronic mails provided results that deviated from the generally held beliefs and cast some shadows on widely held opinions on the origins of 419 mails and that of Edelson (2003) who reported that 67% of the scam e-mails were from Nigeria. The finding of this study therefore, supports the third view of the controversial views on the acclaimed source of electronic scam mail. This third view as submitted by Smith *et al.* (2001) is that the term "Nigerian Advance Fee Fraud" is only partially accurate and the problem is truly one of international dimension with victims and offenders being located across the globe. However, Nigeria may be one of the countries where scam mail authoring is prominent, as informed by a relatively high percentage accrued to the country among four other countries.

The percentage of scam mail classified to be Nigerian was higher than average, although not as high as one would expect as it has been believed generally that Nigeria was responsible for the crime. The implication of this is that electronic scam mail problem cannot be pinned down to Nigeria. However, it is possible that Nigeria is one of the countries that are prominent in authoring such mails. The performance of our model despite the few words contained in the comments of the study corpora shows that the use of linguistic cues in authorship profiling of anonymous texts is promising. Finally, the improvement in model performance when the similarity between Nigerian and Ghanaian texts was put into consideration implies that linguistic similarities between variant of language in consideration should be resolved in order to improve classification model performance.

5.1 Limitations of the Study

The availability of a corpus of electronic scam mail with authenticated country sources could have been more appropriate for the study, in the absence of this, the available corpus of *www.topix.com* was used as a surrogate. Other limitations include the relatively low number of words in the comments of the training and validation dataset, natural language processing applications, especially for the purpose of authorship attribution perform better with large text sizes. Also, the study modelled the specific linguistic features of the Nigerian variant because of the motivation for the study; however, modelling the linguistic features of each of the dialects under consideration could have produced a better performance of the model.

5.2 Recommendations

Based on the finding of this study, it is recommended that researchers in technology should exploit the results of studies in the humanities in general and languages of communication specifically to enable the optimization of technology to solve human problems such as the forensic determination of the origin of scam mail through linguistic analyses. There should be more research in the creation of text corpora in Nigeria. This is necessary because the availability of corpora will go a long way in helping researchers to carry out statistical language modelling of Nigerian text more readily. Also law enforcement agencies of each country, in particular Nigeria, should create a database of investigated fraudulent mails, with contents stored and tagged appropriately with the criminals' profiles. Such data can be useful in investigating transnational digital criminal offences by applying the proposed statistical language modelling technique.

Finally, it is recommended that governmental, non-governmental and education sectors in Nigeria should provide seminars and training on more profitable use of the computer technologies including the Internet and also on the evil of electronic scamming to the Nigerian youth. The information age and its numerous enabling technologies have come to stay, deliberate introduction of positive use of these devices will reduce their possible negative uses.

6. Conclusions

Based on the findings of the study, it could be concluded that linguistic cues have potentials of being used for investigating transnational digital breaches if properly exploited. More research effort in this area, particularly in the deliberate storage of annotated electronic texts from different countries of the world could serve as a useful resource if a need arises to profile the country source of a controversial anonymous text. It could also be concluded that electronic scam mail problem cannot be pinned down to Nigeria as believed generally, though Nigeria could be one of the countries that are prominent in authoring such mails as revealed by the study.

7. Bibliographical References

Adetugbo, A. (1979). Appropriateness of Nigeria English. In E. Ubahakwe (Ed.), Varieties and Functions of English in Nigeria: Selections from the Proceedings of the Ninth Annual Conference of the Nigerian English Studies Association. Ibadan, Nigeria: African Universities Press, pp. 167-183.

Alshalan, A. (2006). Cyber-crime fear and victimization: an analysis of a national survey. Doctoral Thesis. Mississippi State University, Department of Sociology, Anthropology, and Social Work.

Banjo, A. (1979). Beyond intelligibility: A presidential address. In E. Ubahakwe (Ed.), Varieties and Functions of English in Nigeria: Selections from the Proceedings of the Ninth Annual Conference of the Nigerian English Studies Association. Ibadan, Nigeria: African Universities Press, pp.7-13.

Bayart, J., Ellis, S. & Hibou, B. (1999). The Criminalization of the State in Africa. Bloomington: Indiana University Press.

Bird, S., Klein, E & Loper, E. (2007). Natural Language Processing in Python. USA: O'Reilly Media.

Choo, K.R. (2011). The cyber threat landscape: challenges and future research directions. Computers and Security, 30, pp. 719 – 731.

Cruz-Ferreira, M. & Abraham, S.A. (2006). The Language of Language-core Concepts in Linguistic Analysis. 2nd ed. Singapore: Prentice Hall Pearson.

De Vel, O. , Anderson, A., Corney, M. & Mohay, G. (2001). Mining e-mail content for author identification forensics. Special Interest Group on Management of Data (ACM SIGMOD) Record, 30(4), pp. 55-64.

Dyrud, M.A. (2005). I Brought You a Good News: An Analysis of Nigerian 419 letters. In Proceedings of the 2005 Association for Business Communication Annual Convention. Retrieved Feb. 12, 2010, from http://www.businesscommunication.org/conventionsNew/proceedingsNew/2005New/PDFs/07ABC05.pdf.

Edelson, E. (2003). The 419 scam: information warfare on the spam front and a proposal for local filtering. Computers and Security, 22 (5), pp. 392-401.

Egbe D.I. (1979). Spoken and written English in Nigeria. In E. Ubahakwe (Ed.), Varieties and Functions of English in Nigeria: Selections from the Proceedings of the Ninth Annual Conference of the Nigerian English Studies Association. Ibadan, Nigeria: African Universities Press, pp. 86-106.

Iqbal, F., Hadjidj, R., Fung, B.C.M. & Debbabi, M. (2008). A Novel Approach of Mining Write-Prints for Authorship Attribution in E-mail Forensics. In 2008 Digital Forensic Research Workshop. Elsevier Ltd. Retrieved Nov. 16, 2009, from www.elsevier.com/locate/diin. 2008.05.001.

Jowitt, D. (1991). Nigerian English Usage: An Introduction. Nigeria: Longman.

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

Koppel, M., Schler, J., Argamon, S. & 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. Seattle, USA: ACM, pp.659-660.

Koppel, M., Schler, J., & Argamon, S. (2009). Computational methods in authorship attribution. Journal of the American Society for Information Science and Technology, 60(1), pp. 9-26.

Kujore, O. (1985). English Usage: Some Notable Nigerian Variations. Nigeria: Evans Brothers Nigeria Publishers Limited.

Longe, O. & Osofisan, A. (2011). On the Origins of Advance Fee Fraud Electronic Mails: A Technical Investigation Using Internet Protocol Address Tracers. The African Journal of Information Systems, 3(1), 17-26.

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, 4(3), pp. 130-143.

Poplack, S. (1993). Variation theory and language contact. In D.R. Preston (Ed.), American Dialect Research. USA: John Benjamins Publishing Co, pp. 251-286.

Reith, M., Carr, C. & Gunsch, G. (2002). An examination of digital forensic models. International Journal of Digital Evidence, 1(3), pp. 1-12.

Rosenfeld, R. (2000). Two Decades of Statistical Language Modelling: Where Do We Go From Here? In: Proceedings of the European Conference on Speech Communication and Technology. Retrieved Nov. 20, 2010, from citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.140.5518.

Sculley, D. & Wachman, G.M. (2007). Relaxed Online SVMs for Spam Filtering. In: SIGIR 2007 Proceedings. USA: ACM, pp. 415-422.

Smith, R.G., Holmes, M. N. & Kaufmann, P. (2001). Nigerian advance fee fraud. International Society for the Reform of Criminal Law. Retrieved Feb. 14, 2010, from http://www.isrcl.org/Papers/Nigeria.pdf.

State Department Publication. (1997). Nigerian advance fee fraud. Retrieved 16 June, 2009, from http://www.state.gov/documents/organization/2189.pdf.

Torosyan, A. (2003). Cyber crime programs by state and local law enforcement: a preliminary analysis of a national survey. Doctoral Thesis. California State University, Department of Criminal Justice.

Wardhaugh, R. (1994). An Introduction to Sociolinguistics. 2nd. ed. U.K: Blackwell Oxford.

Witten, I.H. & Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques. 2nd ed. USA: Morgan Kaufmann publishers.