RESEARCH ARTICLE OPEN ACCESS

Text Mining Approach in Curtailing Cyber-Crimes in Nigeria

Kasali Funmilayo, Kuyoro Afolashade, Awodele Oludele
Department of Computer Science
Babcock University
Ilishan Remo Ogun State Nigeria
Ilishan Remo - Nigeria

ABSTRACT

The issue of cyber-crimes has continually pose a major threat both locally and globally since the growth in the use of computers and the internet. Various measures are continually been used by all bodies concerned to curb this trend that has gained prominence among youths although only minor success has been recorded. Hence, there is need for a more flexible, robust and adaptable approach in curbing this anomaly. The emergence of text mining as a technique in deriving high quality information from unstructured textual data that is available in enormous quantity on the web is recently gaining attention from researchers as it has been seen and verified to be effective in curtailing the activities of cyber-criminals. This study used the method of Review in research to explore how Social media monitoring and the application of Text Mining techniques can be used in curtailing crimes in Nigeria and relevant information was extracted using the Inductive approach. The work explored the novel world of text mining, its basic concepts, different application areas and an approach in using it to curb cyber-crimes.

Keywords:- Text Mining, Cyber-crime, Social Network Analysis.

I. INTRODUCTION

As technological advancements continue to open a world of opportunities in different areas of life ranging from communication, education, and industrial sciences, it also simultaneously creates a new era of criminality: cybercrime. From professional hacking, cyber bullying and virus writing to identity theft and fraud, cyber criminals are always finding new ways of committing crimes at a very fast pace and it is becoming more difficult for them to be caught. The rate at which digital crime occurs leaves little hope for human intervention to prevail without human error. Artificial Intelligence has a much better chance at detecting and analyzing appropriate defenses against cybercrime because it is also a product of technology and it may be the only chance left for mankind now to curb this anomaly [1]. Highlighted below are some cyber-crime facts as stated by Norton [2];

- Cybercrime has now surpassed illegal drug trafficking as a criminal moneymaker
- Somebody's identity is stolen every 3 seconds as a result of cybercrime
- Without a sophisticated security package, unprotected Personal Computers can become infected within four minutes of connecting to the Internet.

Criminals committing cybercrime use a number of methods, depending on their skill-set and their goal. Here are some of the different ways cybercrime can take shape: Theft of personal data, Copyright infringement, Fraud, Child pornography, Cyberstalking, Bullying and so on.

Intel Security Group estimates that cybercrime costs the global economy more than a whopping $400 billion annually, possibly even maxing out at $575 billion which is more than the national income of most countries and governments [1]. In Nigeria, the National Security Adviser, Sambo Dasuki, admitted that every nine seconds, a Nigerian commits crime on the internet with a sharp rise from 0.9% in the 90s to 9.8% in 2014 [3]. A computer crime and cyber survey conducted recently also indicated that Nigeria is the most internet fraudulent country in Africa [4].

The focus of this work is to explore in detail how text mining techniques can be used to curtail cyber-crimes in Nigeria. The remaining part of this work is arranged as follows: Section 2 gives the literature review presenting the novel concept of text mining, steps to text mining and some areas in which it has been applied; Section 3 gives a detailed exploration on how text mining techniques can be applied in the area of curbing crimes and finally, Section 4 gives the conclusion and recommendation for further studies.

II. LITERATURE REVIEW

2.1 *TEXT MINING*

The World Wide Web (www) contains huge and constantly increasing source of information which security experts may consult for information about cyber criminals by tracing their activities online, but the vast amount of data especially unstructured data is often tremendous. It is thus important for science and new technologies to help discover new relationships and increase the efficiency of cyber security experts. The field of text mining is concerned with algorithms and techniques by which valuable information can be gotten from textual data. This field has been found to be useful when the amount of text is extensive [5]. Text Mining is the discovery of new, previously unknown information, by automatically extracting information from different written resources. A key element is the linking together of the extracted information together to form new facts or new hypotheses to be explored further by more conventional means of experimentation. Text mining is different from what is mostly common in web search. The user is typically looking for something that is already known and has been written by someone else when searching online. The problem is pushing aside all the material that are not in any way relevant so as to concentrate on more relevant ones but in text mining, the aim is to discover unknown information i.e. an information that is not known by anyone and so could not have yet been written down [6]. Text mining also known as text analytics or knowledge discovery from textual databases is an interdisciplinary field which draws on information retrieval, information extraction, Natural Language Processing (NLP), data mining, machine learning, statistics and computational linguistics. Text mining normally involves the process of structuring input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data and finally, evaluation and interpretation of the output. Text mining also involves the process of extracting semi-structured information components from text, usually from multiple documents, and then reason with these semi-structured components to derive patterns. The aim of text mining is to provide the following facilities as outlined by Limika et al., [7]:

- Distill the meaning of a text in a concise form
- View accurate summaries before plunging into full documents
- Navigate efficiently through large textual databases
- Perform natural language information retrieval

It is a known fact that an estimated 80% of data which includes emails, newspaper or web articles, internal reports, transcripts of phone calls, research papers, blog entries, and patent applications, to name a few are unstructured. Moreover, 7 million web pages of text are being added to our collective repository daily hence, the need for an application that can read between enormous amounts of data within seconds. The main goal of text mining is to turn text into data for analysis via application of NLP and analytical methods. Notable areas in which text mining applications have excelled include but not limited to the following as stated by Stephanie [8]:

- Enterprise Business Intelligence/Data Mining, Competitive Intelligence
- E-Discovery, Records Management, Publishing and media
- National Security/Intelligence
- Scientific discovery, especially Life Sciences
- Search/Information Access
- Social media monitoring, Political institutions, Bioinformatics, Business Intelligence and National Security

- Banks, insurance and financial markets

Historically, labour intensive text mining approaches first surfaced in the mid-1980s but the field has expanded over the past decade due to improvement in technology [9]. The emergence of text analytics stems from a refocusing of research in the late 90s from algorithms due to application as described by Prof. Marti Haerst in the paper untangling text mining [10].

Social media monitoring is an active monitoring of social media channels for information, usually tracking of various social media contents such as blogs, wikis, news sites, microblogs such as twitter, social networking sites, photo sharing websites, forums and others as a way to determine the volume and sentiment of online conversation about a brand or topic. More recently, the law enforcement community is increasingly turning to social media monitoring to prevent and investigate crimes. Law Enforcement Agencies (LEA) who recently participated in a LexisNexis survey said social media monitoring has helped them find evidence, identify and locate suspects solicit tips, and alert the public about crimes [11]. In an online study that was conducted among the PoliceOne.com community by LexisNexis Group in 2014, a corporation that provides computer-assisted Legal research, business research and risk management services, of the Federal, State and Local enforcement officials surveyed, 81% said they use social media for investigative purposes, 67% believe that social media is valuable for anticipating crimes, 73% said it can help solve crime faster and 78% said they expect to use it even more this year [12].

2.1.1 *Steps Involved In Text Mining:*

The basic processes involved in text mining is Information Retrieval, Information Extraction and Natural Language Processing [13] and steps are text processing, text transformation, feature selection, text mining methods and interpretation/evaluation.

- **Text Preprocessing:** This step involves tokenization, stop word removal and stemming. Tokenization stage segments unstructured texts into words by removing blank spaces, commas etc. Stop word removal clears texts of HTML, XML tags from web pages. Then the process of removal of stop words such as 'a', 'is', 'of' etc., is performed. In stemming, Stemming refers to the process of identifying the root of a certain word. We have 2 types of stemming which are inflectional and derivation. The most common type of algorithm being used is the Porter's algorithm [14]. These steps are shown in Figure 1

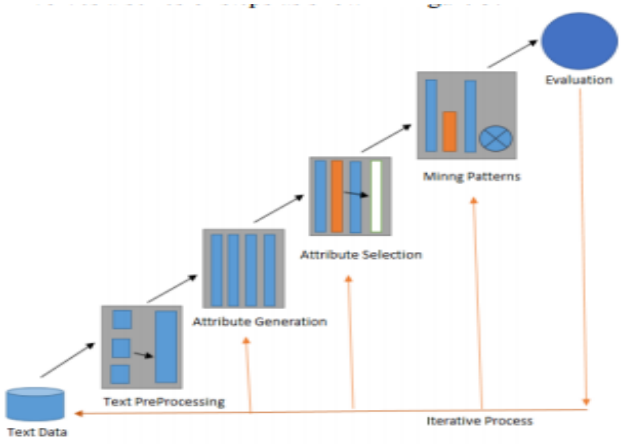

Fig. 1: Process of Text Mining [15]

- **Text Transformation**: Text document is represented by the words it contains and their occurrences. Two approaches used for document representation are bag of words and vector spaces.
- **Feature Selection:** This is also known as variable selection. It is the process of selecting a subset of important features for use in model creation. This phase mainly performs removing features which are redundant it is the subset of feature extraction.
- **Text Mining Methods:** At this point, data mining methods are now used for text mining. Some of the techniques include clustering, classification, information retrieval etc.
- **Interpretation/Evaluation:** Analyzing results gotten after mining.

2.2 *Review of Text Mining Applications*

In a world of big data boom, filtering through enormous amount of semi structured and unstructured information in the form e-mails, spam, Wikis, microblogs amongst others go beyond the reach of ordinary data mining tools, text mining goes further by locating the source of information, connect disjointed concepts in widely separated documents, maps relationships between activities and helps discover new answers to questions [16].

Ronen et al., [5] presented and used an end-to-end text mining methodology for relation extraction of adverse drug reactions (ADRs) from medical forums initially not covered in clinical trials for evaluating new drugs that come to market but were only reported later on by the FDA as a label change. In this work, more online message boards or social media sources should have been mined to ascertain high level of efficiency and certainty in the methodology used. The pre-processing steps should also have been extended by taking care of irregularities such as corrupt grammar, typographical errors, extensive use of acronyms and informal syntax.

Antonette, Elizabeth & Helen [17] proposed a text mining approach to automate team work assessment in chat data using open source Ajax chats environment and 272 randomly selected students. The unstructured data gotten was pre-processed by removing stop words and they also used indicative term dictionary for generating rules for categorization analysis of chat data. Naïve-Bayes, k-NN and Support Vector Machine (SVM) were applied for topic classification using training and test sets from their data which was checked for reliability by comparison with manual coding. Other datasets should still be used to test reliability of the classifiers used as they claimed that their findings will be useful in identifying appropriate classifiers and visualization techniques for chat data analysis.

Volkan & Turgay [18] applied Text Mining and Social Network Analysis (SNA) on Computer Science and Engineering theses in Turkey by examining 6,834 Masters and PhD theses conducted between 1994 and 2013. The data was gotten from the YOK theses web portal and they developed a web crawling and parsing tool in C# programming language for downloading and processing these records using PRETO, a text mining tool they developed. After downloading the HTML files, they cleaned, parsed and now extracted useful information from these files. They obtained bigram occurrence frequencies within abstracts after the application of text mining tools, they now did Social Analysis by creating a network representation from the bigram occurrence data. They used Gephi 0.8.2, a SNA tool for network visualization and analysis. The results of the analysis were presented visually and they were able to know that in the last 5 years, Machine Learning were most popular and they form a separate community rather than being part of Neural Network related community. Also, Genetic Algorithms had also been popular in the past decade, they were also able to see that Computer Network related concepts had shifted towards Sensor Networks and Wireless Networks. They witnessed a paradigm shift from database (1999 – 2003) to Machine Learning. The work could go further by also applying dynamic network visualization to explore and understand structure transition and content propagation over time.

Sanghee & Min [19] used text mining as a method of analyzing health questions in Q &A that people post in Social Q & A where they obtain and share information, advice and experiences. 69,363 of health questions about Sexually Transmitted Diseases (STDs) posted from 2009 to 2012 were randomly collected from Yahoo! Answers. In other to critically review these questions, IBM SPSS Modeler Premium was used because of the predictive models it contains. The tools extracted major concepts from texts of STD questions, counted the frequency of the concepts shown in one question and listed in other to identify the popular concepts discussed. The tools also generated concept maps to identify relationship among concepts easily. Interpretation of text mining results is mostly based on terms without considering the contexts in which the questions are shared. The scope of the study should be extended to other health related areas as well and the problem of ambiguity is inherent in this study. The data cleaning method should also be upgraded to ensure data quality before analysis takes place.

Manabu et al., [20] used text mining techniques to speed up manual curation of phosphorylation

information which is a global regulator of cellular activity. They enhanced the Rule based Literature Mining System for protein Phosphorylation (RLIMS-P), a rule based Information Extraction (IE) system to identify this information by integrating new Natural Language Processing (NLP) techniques to reduce efforts used for system development and maintenance. The newly enhanced system was applied to the publicly available PubMed in Central (PMC) open access subset and they obtained promising results in mining the full text article, also in RLIMS-P focuses only on protein phosphorylation information but the new system can generally be used for other Post-translational modification (PTM) types. The newly enhanced system should be further refined and thoroughly evaluated in different settings and corpora. The system should also be ported to the extraction of other IE tasks in the domain such as various PMT types other than phosphorylation.

Andrew et al., [21] applied text mining techniques in bone biology to enable bone scientists have easy access to enormous literature available on the internet in biological literature databases so as to stay abreast of new information using a thesaurus based method to obtain a significant improvement over random guessing to discover the existence of relationships. More work should be done on extending the set of vocabulary terms and in developing visualizations which are more meaningful to a bone biologist information experts, to further improve the accuracy and efficiency of the approach used, it should be applied to other similar biological domains.

Rathi, Shirgaonkar & Dhote [22] applied text mining techniques to detect bomb blast using the EART system, a user friendly application developed in other to simplify rule mining in textual documents collection. The main aim of the system is to discover new previously unknown information by automatically extracting information from different written resources. The extracted information are then linked together to form new facts/hypotheses which are now further explored through more conventional means of experimentation. The system could only work for small unstructured texts. More so, for a system that was claimed to detect bomb blast, a real life investigation of a bomb detection done by it should have been ascertained to really show its effectiveness and reliability.

Judith & Micheal [23] applied Text Mining techniques in Indexing by using Image Analysis based on letters sharing a common baseline and having characteristics aspect ratio and size to find text in maps, document structure to find captions and titles and then text mining to assign each map to a subject category, a geographical place and a time period. They measured how accurately the system classifies maps by giving the system unseen maps to compare the categories it assigned to that of the categories assigned by human indexers with a test set of 50 and the result showed automatic classification to be above 60%.

Ritu [24] applied text mining techniques to identify articles related to a particular domain by collecting unstructured texts from various sites online and applying classification techniques to remove articles which do not belong to the domain of interest and clustering techniques to form subgroups between the classified articles. In this work, 43 files were analyzed for classification, each document was tokenized and reduced to a "a bag of words" by the program the resulting words were now stemmed automatically using intelligently created Match files, the articles were now compared for similarity using Jaccard's coefficient (a statistic used for comparing the similarity and diversity of sample sets and it is defined as the size of the intersection divided by the size of the union of the sample sets) before clustering. For clustering, they used a Neo-hybrid algorithm, a modified and enhanced version of k-NN and the single link clustering algorithm. The main problem of NLP which is ambiguity still persists in this work although they claimed that the approach has a wide scope of application in Grid Computing Environment.

III. USING TEXT MINING TECHNIQUES IN CURBING CRIMES

In curtailing crimes, Social Network Analysis (SNA) which is a new type of Intelligence is required, what it does is to get individual nodes which could be people, events, places and so on depending on the type of network, which are connected by complicated

yet logical associations that form the networks [25]. These are pervasive networks with simple laws and instructions, they form the fundamental core of many organizations, events and social processes. Technological advancement has made social networking so easy as a result of the proliferation of many social networking sites and increasing internet availability and lower cost in accessing it. According to the Nigeria Communications Commission (NCC), they recently pronounced that the number of internet users on the country's telecom networks increased to 83,362,814 as at February 2015 [26] which is almost 40% of the present total population. Most cyber-criminals have a network in which they exchange new hacking software, ideas and information. The most common cyber-crime committed in Nigeria is cyber-financial fraud as Mr. Adebayo Adelabu, Deputy Governor, Financial Systems Stability, Central Bank of Nigeria (CBN) said about N159 billion was lost to electronic fraud between 2000 and the first quarter of 2013 [27]. This openly indicates that if unchecked, could wreck disastrous havoc on the country's economy particularly with the new Cashless policy in place. It is a gigantic task battling cyber criminals but it is a war that needs to be won as a nation. Developed countries all over the world have various agencies to tackle this problem but Nigeria has done very little in this war as there is currently no effective policy against cyber-crime in the country. It is not only financial fraud that cyber-criminals are targeting, other cyber-crimes too take place like cyber-bullying, terrorism, online identity theft and so many others which usually go unreported. The better approach in curbing this trend in criminality is to put in place adequate Social Network Analysis by monitoring suspects online social activities. Most cyber-criminals operate as gangs and they have forums online either through the use of Facebook, twitter, e-mails amongst others in which they communicate and in doing this, they generate lots of unstructured texts which could contain hidden knowledge that can be analyzed using text mining approach to get required and adequate information to trace, monitor and curtail whatever activities they intend to do or to even catch up with them. There presently exist different Social media tools that can be used to monitor people's social media interaction online although choice of tools should depend on goals, sources monitored, ability to respond (in case of listening tools), historical data and other stuffs depending on the miner. When this information has been monitored and analyzed using SNA, Text mining tools can now be used to mine the extracted information. Table 1 shows some text mining technologies as offered by some commercial vendors. Some open source text mining tools include Carrot 2, GATE, Natural Language Toolkit, Text Mechanic, R programming language amongst others.

TABLE 1: Text Mining technology and Commercial Vendors

	Vendors							
Features	Inxight	Autonomy	Clearforest	SAS	Convera	Megaputer	SPSS	IBM
Information Extraction	X	X	X	X	X	X	X	X
Topic Tracking		X						
Summarization	X	X			X	X		X
Categorization	X	X	X	X	X	X	X	X
Concept Linkage		X	X	X				
Clustering		X			X	X		X
Information Visualization	X						X	
Quest		X				X		

ion Answering								

Source: Weiguo et. al. [16]

IV. CONCLUSIONS AND RECOMMENDATIONS FOR FURTHER STUDIES

Text Mining is still a novel area in information analysis hence all the three basic approaches which are essential in successfully mining unstructured texts should be explored by interested researchers which are Information Retrieval, Information Extraction and NLP. Moreover, unstructured texts vendors like Facebook, Twitter, microblogs owners amongst others could start hoarding this enormous information available or start charging text miners before they can have access to petabytes of information in different domains available online especially vendors of research/journals databases. Information security is another roadblock to text miners, ethical issues, privacy violation and law protecting the use of plethora of unstructured texts available on the internet are also challenges that need to be surmounted. This study has explored the trendy world of text mining, areas of application, benefits and the process involved in text mining. The work gave details on cyber-crimes, the need for social media monitoring as a way of curbing cyber and related crimes. Moreover, Nigeria as a country should rise fully to this fight and ensure they put adequate measures in place to win it and embracing Social media monitoring, using Social Network Analytical tools and Text Mining Technology will at least go a long way in curtailing cyber-criminal acts. The identified gaps in Text Mining is still the main issues involved in NLP, how to deal with ambiguous words and statements present in unstructured texts, more work still need to be done in developing algorithms that can analyze semantic meaning of word and in the context which they are used since people use language freely.

ACKNOWLEDGEMENT

We thank the entire Babcock Postgraduate Students especially the present 2015/2016 PhD class for their objective criticisms and contributions towards the success of this work.

REFERENCES

[1] Kelly, T. (2015). Artificial intelligence may be the only solution to keep-up with cyber criminals. Retrieved from http://thescienceexplorer.com/technology/using-artificial-intelligence-take-down-cyber-criminals

[2] Norton by Semantec, (2015). Prevention tips. Retrieved from http://us.norton.com/cybercrime-preventionhttp://us.norton.com/cybercrime-prevention

[3] Richard, U. (2014). Rate of cyber-crime in Nigeria is alarming, retrieved from vanguardngr.com/2014/06

[4] This day, (2014). Nigeria ranked third in the world for cyber-crime says survey, retrieved from balancing act-africa.com/news/en/issue-no-302/computing

[5] Ronen, F., Oded, N., Aviv, P., Binyanmin, R., (2015). Utilizing Text Mining on Online Medical Forums to Predict Label Change due to Adverse Drug Reactions ACM. ISBN 978-1-4503-3664-2/15/08, DOI: http://dx.doi.org/10.1145/2783258.2788608. Pg. 1779

[6] Steven, O. K. (2006). Text Mining for Business Intelligence, INSEAD – UNILEVER workshop "Scenario Planning and Scanning on Biofuels and Related Commodities" 19-20 Oct 2006.

[7] Limika, D., Muhammed, A., Jahiruddin, Guarav, S., (2007). Text Mining through Entity-Relationship Based Information Extraction, IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology – Workshops 0-7695-3028-1/07 $25.00 © 2007 IEEE, DOI 10.1109/WI-IATW.2007.75

[8] Stephanie, P. (2013). What is text mining, Retrieved from syr.edu/2013/04/23/what-is-text-mining

[9] Wikipedia, (2015). Text Mining, retrieved from https://en.wikipedia.org/wiki/Text_mining

[10] Hearst, Marti A. (1999). "Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics". pp. 3–10.doi:10.3115/1034678.1034679. ISBN 1-55860-609-2

[11] William, C. (2015). Law Enforcement Agencies (and Corporate Security) benefit for Social Media Monitoring, retrieved from cyberalert.com/blog/index.php

[12] LexisNexis Risk Solutions, (2014). Survey of Law Enforcement Personnel and their use of Social Media, www.lexisnexis.com/investigations

[13] Sayantani, G., Sudipta, R, Samir, K. B., (2012). A tutorial review on Text Mining Algorithms, International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 4, Pg. 223

[14] Sumathy, K. L., Chiddambaram, M. (2013). Text Mining: Concepts, Applications, Tools and Issues: An Overview, International Journal of Computer Applications (0975 – 8887) Vol. 80, No.4, Pg. 29-31

[15] Lokesh, K., Parul, K. B. (2013). Text Mining: Concepts, Process And Applications, Journal of Global Research in Computer Science, ISSN-2229-371X, Vol. 4, No. 3, Pg. 37

[16] Weiguo, F., Linda, W., Stephanie, R. & Zhongju, Z. (2006). Tapping the Power of Text Mining, Communications of the ACM, Vol. 49, No. 9, Pg. 77

[17] Antonette, S., Elizabeth, K. & Helen, H. (2015). Text mining approach to automate teamwork assessment in group chats, ACM 978-1-4503-3417-4/15/03. http://dx.doi.org/10.1145/2723576.2723648, Pg. 434-435

[18] Volkan, T., Turgay, T. B. (2014). Text Mining and Social Network Analysis, International Conference on Computer Systems and Technologies, Publication rights licensed to ACM. ISBN 978-1-4503-2753-4/14/06, http://dx.doi.org/10.1145/2659532.2659639 187.

[19] Sanghee, O., Min, S. P. (2013). Text Mining as a Method of Analyzing Health Questions in Social Q&A, ASIST 2013, November 1-6, 2013, Montreal, Quebec, Canada.

[20] Manabu, T., Cecilia, N. A., Qinghua, W., Cathy, H. W., Vijay-Shanker, K. (2013). Text Mining of Protein Phosphorylation Information Using a Generalizable Rule-Based Approach, ACM 978-1-4503-2434-2/13/09, Pg. 201-209

[21] Andrew, H., Snehasis, M., Qian, Y., Shiaofen, F., Yuni, X. & Joseph, B. (2010). Text Mining for Bone Biology, ACM 978-1-60558-942-8/10/06, Pg. 522- 530

[22] Rathi, S., Shirgaonkar, S., Dhote, C. A. (2010). Text Mining: Applied for Bomb Blast Detection, International Conference and Workshop on Emerging Trends in Technology, Copyright 2010 ACM 978-1-60558-812-4… Pg. 558-559

[23] Judith, G., Micheal, L. (2009). Text Mining for Indexing, ACM 978-1-60558-322-8/09/06. Pg. 467

[24] Ritu, A. (2005). Text Mining: Classification & Clustering of articles related to sports, Conference Paper, Kennesaw, GA, USA. ACM 1-59593-059-0/05/0003, Pg. 153-154

[25] Steve, R. (2006). Social Network Analysis as an Approach to Combat Terrorism: Past, Present, and Future Research, The Journal of the NPS Center for Homeland Defense and Security, Issue 2, Vol. 2, Retrieved from https://www.hsaj.org/articles/171

[26] Vanguard (2015). 83m internet users in Nigeria – NCC, Retrieved from http://www.vanguardngr.com/2015/05/83m-internet-users-in-nigeria-ncc/

[27] PM News, (2015). Cyber Crime in Nigeria, Retrieved from http://www.pmnewsnigeria.com/2015/06/24/cyber-crime-in-nigeria/