# Classification of Social Blogs Comments Using Text Mining

**Adesesan B. Adeyemo[1] and Adebola K. Ojo[2*],**

[1,2]Computer Science Department, University Of Ibadan
Ibadan, Oyo State, Nigeria

*Corresponding Author

## Abstract

In this work, a novel e-governance framework that captured the societal impact of public sector regulations in an attempt to decipher the public's stance towards governmental decisions was proposed. This was done by exploring text mining techniques towards firstly capturing the public's opinions (communicated online) about governmental decision and secondly analyzing the polarity of the mined opinions so that they are considered in subsequent governmental decisions. Citizens' opinions and comments that were posted on online blogs and Facebook were decomposed in order to evaluate how government decisions were perceived by the public and hence, how the public's implicit feedback should be interpreted by government bodies in their subsequent actions. The motivation for our study is that up-to-date government social web sites are not consistently evaluated in the government decision-making process and those citizens' voices are most of the times heard in a limited audience.

***Keywords:*** *e-governance, text mining, blogs, government, decision-making*

## 1. Introduction

Text Mining [1] is a process of extracting new, valid, and actionable knowledge dispersed throughout text documents and utilizing this knowledge to better organize information for future reference. Mining implies extracting precious nuggets of ore from otherwise worthless rock [2] and Text Mining is the gold hidden in mountains of textual data [3]. This study proposes a framework for discovering previously unknown and hidden information from public opinions and views.

With the increasing awareness among citizens about their rights and the resultant increase in expectations from the government to perform and deliver, the whole paradigm of governance has changed. Government, today, is expected to be transparent in its dealings, accountable for its activities and faster in its responses. This has made the use of Text Mining imperative in any agenda drawn towards achieving good governance. It has also led to the realization that such technologies could be used to achieve a wide range of objectives and lead to faster and more equitable development with a wider reach. [4]

Text mining techniques are used to draw out the occurrences and instances of key terms in large blocks of text, such as articles, Web pages, complaint forums, or Internet chat rooms and identify relationships among the attributes [5]. Often used as a preparatory step for data mining, text mining often translates unstructured text into a useable database-like format suitable for data mining for further and deeper analysis [6]. [7] also described text mining as an emerging technology that can be used to augment existing data in corporate databases by making unstructured text data available for analysis.

To assist policy makers in devising their own plans and initiatives, the processes of e-governance implementation is divided into three phases which are not dependent on each other or mutually exclusive, but conceptually they offer three ways to think about the goals of digital governance. A modified 3-phase e-Governance [8] structure showing Publish, Interact and Transact phases, and their interactions with Text Mining is shown in Figure 1. The various components of Figure 1, showing their importance and interdependency towards efficient e-governance are presented:

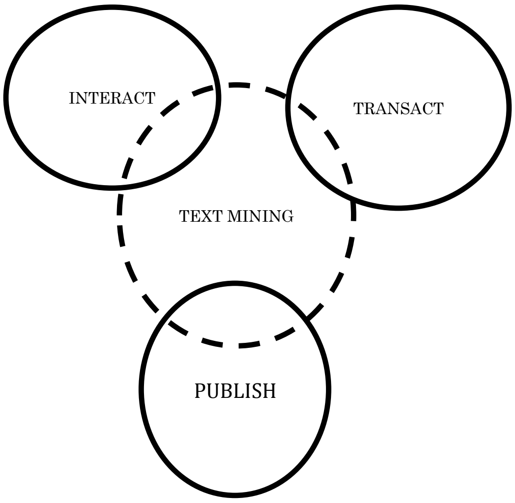


Figure 1: Phases of E-Governance Interaction with Text Mining

PHASE 1: PUBLISH- using ICT to expand access to government information. Publish sites seek to disseminate information about government and information compiled by government to as wide an audience as possible. In doing so, publish sites serve as the leading edge of e-Government.

PHASE 2: INTERACT- broadening civic participation in government. Interactive e-Government involves two-way communications, starting with basic functions like email contact information for government officials of feedback forms that allow users to submit comments on legislative or policy proposals.

PHASE 3: TRANSACT- Allowing citizens to obtain government services or transact business with the government online. A transact website offers a direct link to government services, available at any time. Transact sites can enhance productivity in both the public and private sector by making processes that require government assistance or approval simpler, faster and cheaper. [9]

We were able to achieve only Phase 1 (Publish-Text Mining Interaction) in this work. This is reflected in the following section. The other phases will be addressed in future work.

## 2. Methodology

This study proposes a framework which integrates text and data mining methods for modelling the public's opinions, feedbacks and evaluations of the government decisions. The citizens' feedbacks are then mined and analysed in order to derive the sentiment orientation of the public opinions and the underlying correlation between mined opinions and the formulation of new governmental decisions on related issues as shown in Figure 2.

### 2.1 Document Collection

The first phase in Figure 2 is the document collection. In this phase, the citizens' opinions, complaints and comments were extracted as documents collected through the online web application. These text documents were stored in different formats (such as pdf, doc, txt, html and xls) depending on the nature and type of the data as shown in Figure 3.

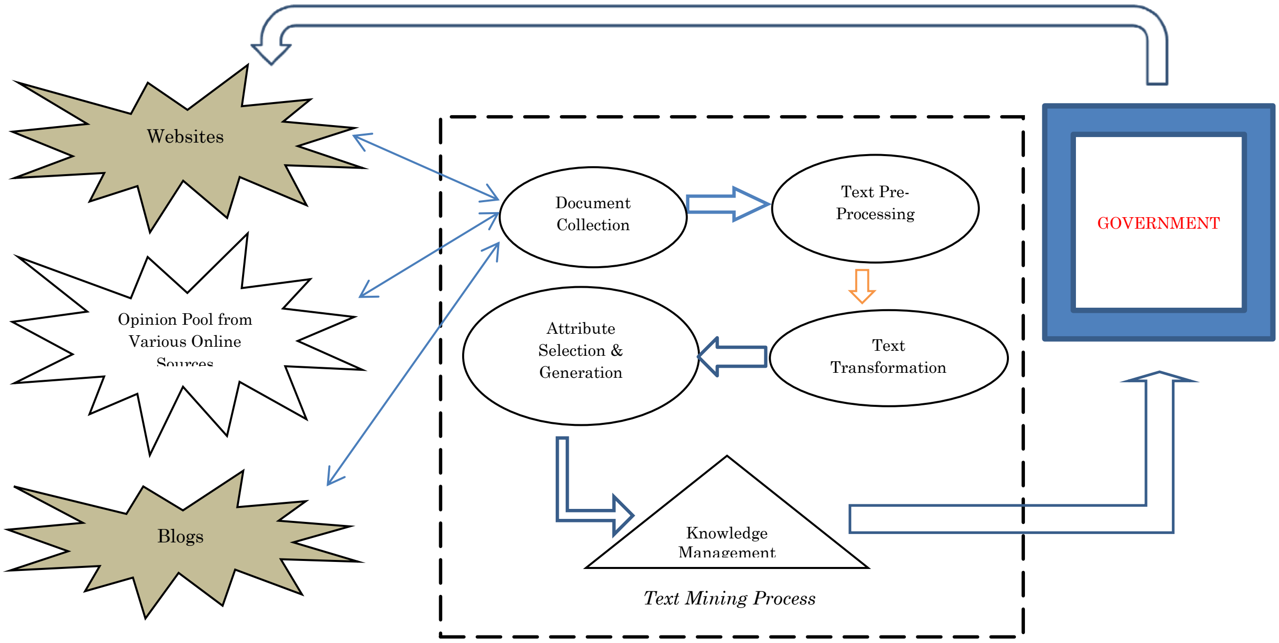


Figure 2: Framework for Text Mining Aided E-Government Content Management

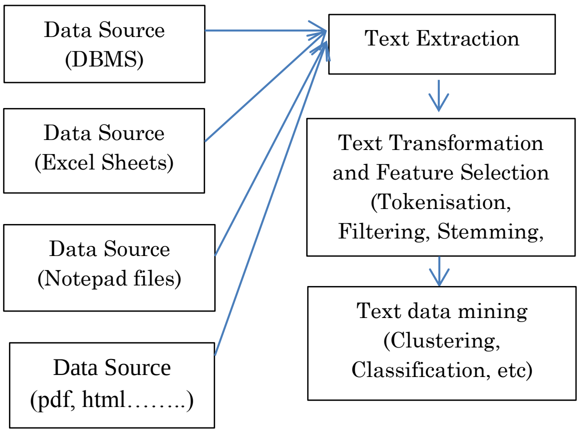

Data Source (DBMS)
Data Source (Excel Sheets)
Data Source (Notepad files)
Data Source (pdf, html……..)
Text Extraction
Text Transformation and Feature Selection (Tokenisation, Filtering, Stemming,
Text data mining (Clustering, Classification, etc)

Figure 3: Text Data Extraction and Processing

## 2.2 Text Pre-processing

This is otherwise known as tokenization or text normalization. Documents are transformed into a suitable representation for the clustering and classification tasks. During term extraction, character text is first parsed into words. This process also strips away words that convey no meaning. Adjectives, adverbs, nouns and multi-word are extracted from the document. Noisy data, such as, tags, punctuation marks, white spaces, special characters and digits are extracted as well. Also, certain words occur very frequently in text data. Examples include "the" and "a". These words are removed from the term collection because they have no meaningful content. By creating a list of such stop words and eliminating them, the number of indicator variables created is reduced. Many of these stop words do not appear in the claim description data, but appear frequently in text data. After removal of stop words, stemming is performed.

## 2.3 Text Transformation

Word frequency and inverse document frequency are two parameters used in filtering terms. Low term frequency (TF) and document frequency (DF) terms are often removed from the indexing of those documents. In "Bags of words" representation each word is represented as a separate variable having numeric weight. The most popular weighting schema is normalized word frequency *tfidf*:

$$tfidf\,(\mathrm{w}) = t\,f.\,log\left(\frac{N}{df(w)}\right) \qquad (1)$$

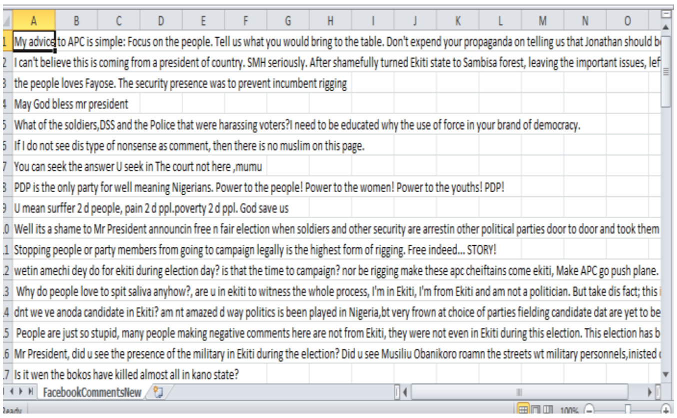
*tf(w)* is the term frequency (number of word occurrences in a document); *df(w)* represents document frequency (number of documents containing the word); *N* gives the number of all documents; *tfidf(w)* is the relative importance of the word in the document. The Transform Cases Operator transforms cases of tokens in a document. This operator transforms all characters in the document to lower case.

## 2.4 Feature Selection and Attribute Generation

In this stage, a subset of the features was selected to represent a document. This created an improved text representation since many features have little information content. Stop Words were removed, and words stemmed down to their roots. Stemming identifies a word by its root and reduces dimensionality (number of features). Features were selected based on classification and some irrelevant attributes were removed.

# 3. Case Study

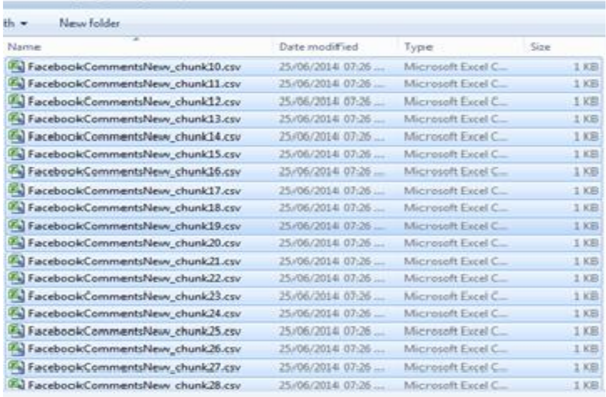
The framework for text mining aided e-government content management was tested using a Nigerian blog site (Nairaland) and the President's Facebook page (www.facebook.com/jonathangoodluck).The aim was to generate a consistent line of thought from the hitherto unstructured/uncoordinated comments by diverse individuals. K-means clustering using Euclidian distance was applied to the matrix of extracted terms from comments on Nigerian President's Facebook page. Each cluster that is created from a k-means clustering procedure has a centre referred to as the centroid. The centroid is the vector of average values for the cluster for each variable entering the clustering procedure. This corpus consisted of 76 comment lines from the citizens on the Governorship Election in Ekiti State in Nigeria, held in June 2014.



| | A |
|---|---|
| 1 | My advice to APC is simple: Focus on the people. Tell us what you would bring to the table. Don't expend your propaganda on telling us that Jonathan should b |
| 2 | I can't believe this is coming from a president of country. SMH seriously. After shamefully turned Ekiti state to Sambisa forest, leaving the important issues, lef |
| 3 | the people loves Fayose. The security presence was to prevent incumbent rigging |
| 4 | May God bless mr president |
| 5 | What of the soldiers,DSS and the Police that were harassing voters?I need to be educated why the use of force in your brand of democracy. |
| 6 | If I do not see dis type of nonsense as comment, then there is no muslim on this page. |
| 7 | You can seek the answer U seek in The court not here ,mumu |
| 8 | PDP is the only party for well meaning Nigerians. Power to the people! Power to the women! Power to the youths! PDP! |
| 9 | U mean surffer 2 d people, pain 2 d ppl.poverty 2 d ppl. God save us |
| 10 | Well its a shame to Mr President announcin free n fair election when soldiers and other security are arrestin other political parties door to door and took them |
| 11 | Stopping people or party members from going to campaign legally is the highest form of rigging. Free indeed... STORY! |
| 12 | wetin amechi dey do for ekiti during election day? is that the time to campaign? nor be rigging make these apc cheiftains come ekiti, Make APC go push plane. |
| 13 | Why do people love to spit saliva anyhow?, are u in ekiti to witness the whole process, I'm in Ekiti, I'm from Ekiti and am not a politician. But take dis fact; this |
| 14 | dnt we ve anoda candidate in Ekiti? am nt amazed d way politics is been played in Nigeria,bt very frown at choice of parties fielding candidate dat are yet to be |
| 15 | People are just so stupid, many people making negative comments here are not from Ekiti, they were not even in Ekiti during this election. This election has b |
| 16 | Mr President, did u see the presence of the military in Ekiti during the election? Did u see Musiliu Obanikoro roamn the streets wt military personnels,inisted |
| 17 | Is it wen the bokos have killed almost all in kano state? |

Figure 4: Extracted Raw Comments Saved in an Excel CSV File Format

All of these comments were extracted into a CSV file. Each line of comment was then converted into a single file using File-Splitter. The Rapid Miner text processing add-in was used for pre-processing and text mining processing.



| Name | Date modified | Type | Size |
|---|---|---|---|
| FacebookCommentsNew_chunk10.csv | 25/06/2014 07:26 ... | Microsoft Excel C... | 1 KB |
| FacebookCommentsNew_chunk11.csv | 25/06/2014 07:26 ... | Microsoft Excel C... | 1 KB |
| FacebookCommentsNew_chunk12.csv | 25/06/2014 07:26 ... | Microsoft Excel C... | 1 KB |
| FacebookCommentsNew_chunk13.csv | 25/06/2014 07:26 ... | Microsoft Excel C... | 1 KB |
| FacebookCommentsNew_chunk14.csv | 25/06/2014 07:26 ... | Microsoft Excel C... | 1 KB |
| FacebookCommentsNew_chunk15.csv | 25/06/2014 07:26 ... | Microsoft Excel C... | 1 KB |
| FacebookCommentsNew_chunk16.csv | 25/06/2014 07:26 ... | Microsoft Excel C... | 1 KB |
| FacebookCommentsNew_chunk17.csv | 25/06/2014 07:26 ... | Microsoft Excel C... | 1 KB |
| FacebookCommentsNew_chunk18.csv | 25/06/2014 07:26 ... | Microsoft Excel C... | 1 KB |
| FacebookCommentsNew_chunk19.csv | 25/06/2014 07:26 ... | Microsoft Excel C... | 1 KB |
| FacebookCommentsNew_chunk20.csv | 25/06/2014 07:26 ... | Microsoft Excel C... | 1 KB |
| FacebookCommentsNew_chunk21.csv | 25/06/2014 07:26 ... | Microsoft Excel C... | 1 KB |
| FacebookCommentsNew_chunk22.csv | 25/06/2014 07:26 ... | Microsoft Excel C... | 1 KB |
| FacebookCommentsNew_chunk23.csv | 25/06/2014 07:26 ... | Microsoft Excel C... | 1 KB |
| FacebookCommentsNew_chunk24.csv | 25/06/2014 07:26 ... | Microsoft Excel C... | 1 KB |
| FacebookCommentsNew_chunk25.csv | 25/06/2014 07:26 ... | Microsoft Excel C... | 1 KB |
| FacebookCommentsNew_chunk26.csv | 25/06/2014 07:26 ... | Microsoft Excel C... | 1 KB |
| FacebookCommentsNew_chunk27.csv | 25/06/2014 07:26 ... | Microsoft Excel C... | 1 KB |
| FacebookCommentsNew_chunk28.csv | 25/06/2014 07:26 ... | Microsoft Excel C... | 1 KB |

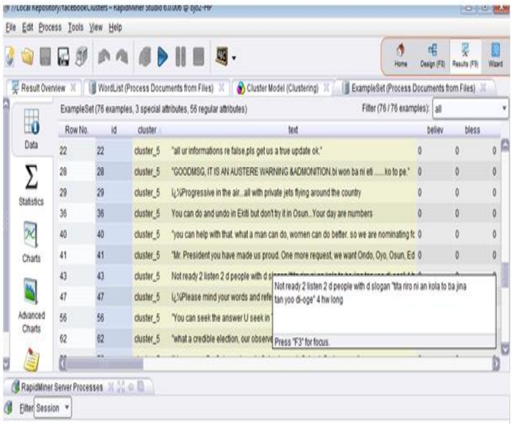
Figure 5: Files Created from the Extracted Comments Using File-Splitter Software Tool.

These files were then text-processed by first transforming them to lower cases and then tokenised. The tokens were filtered by length using minimum of four characters and maximum of twenty-five characters to reduce and eliminate irrelevances. Stop words, such as ‘a’, ‘an, ‘and’, ‘the’, and so on, were also removed/filtered from the entire document. The remaining tokens were finally stemmed down using Snowball stemmer. These were then clustered together using K-means clustering algorithm. Each cluster that was created from a k-means clustering procedure has a centre referred to as the centroid. The centroid is the vector of average values for the cluster for each variable entering the clustering procedure. The clustering procedure was used to create six clusters.

Table 1: The Cluster Model

| Cluster Model |
|---|
| Cluster 1: 52 items |
| Cluster 2: 6 items |
| Cluster 3: 5 items |
| Cluster 4: 3 items |
| Cluster 5: 5 items |
| Cluster 6: 5 items |
| Total number of items: 76 |

The Pruning method used was absolute: words that appeared in less than 5% of all documents as well as those words that appeared in more than 10% of all documents were ignored (pruned). Words that have high representation within a cluster were highlighted. The most commonly occurring words were identified and used to label the cluster as shown in Figure 6.



//Local Repository/FacebookClusters – RapidMiner Studio 6.0.006 @ ojo2-HP
File Edit Process Tools View Help
Home
Design (F8)
Results (F9)
Wizard
Result Overview
WordList (Process Documents from Files)
Cluster Model (Clustering)
ExampleSet (Process Documents from Files)
Data
Statistics
Charts
Advanced Charts
RapidMiner Server Processes
Filter
Session

ExampleSet (76 examples, 3 special attributes, 56 regular attributes)

Filter (76 / 76 examples): all

| Row No. | id | cluster | text | believ | bless | |
|---|---|---|---|---|---|---|
| 22 | 22 | cluster_5 | "all ur informations re false,pls get us a true update ok." | 0 | 0 | 0 |
| 28 | 28 | cluster_5 | "GOODMSG, IT IS AN AUSTERE WARNING &ADMONITION.bi won ba ni eti.......ko to pe." | 0 | 0 | 0 |
| 29 | 29 | cluster_5 | ï¿½Progressive in the air...all with private jets flying around the country | 0 | 0 | 0 |
| 36 | 36 | cluster_5 | You can do and undo in Ekiti but don't try it in Osun...Your day are numbers | 0 | 0 | 0 |
| 40 | 40 | cluster_5 | "you can help with that. what a man can do, women can do better. so we are nominating fc | 0 | 0 | 0 |
| 41 | 41 | cluster_5 | "Mr. President you have made us proud. One more request, we want Ondo, Oyo, Osun, Ed | 0 | 0 | 0 |
| 43 | 43 | cluster_5 | Not ready 2 listen 2 d people with d s | | | |
| 47 | 47 | cluster_5 | ï¿½Please mind your words and refe | | | |
| 56 | 56 | cluster_5 | "You can seek the answer U seek in | | | |
| 62 | 62 | cluster_5 | "what a credible election, our observ | | | |

Not ready 2 listen 2 d people with d slogan "tita riro ni an kola to ba jina tan yoo di-oge" 4 hw long

Press "F3" for focus.

Figure 6: The Clusters Generated by RapidMiner Software

Tables 2 shows the six clusters with their attributes.

Table 2: The Six Clusters with their Order of Frequencies

| **Cluster 1** | **Cluster2** | **Cluster 3** | **Cluster4** | **Cluster5** | **Cluster 6** |
|---|---|---|---|---|---|
| Comment | campaign | educate | kill | govern | congratulate |
| Jonathan | politician | democracy | jonathan | politician | politics |
| Politics | Rig | comment | rig | believe | educate |
| Sambisa | governor | believe | | sambisa | govern |
| Secure | secure | respect | | rig | |
| Respect | jonathan | kill | | secure | |
| governor | govern | governor | | | |
| rig | respect | secure | | | |
| democracy | politics | congratulate | | | |
| politician | | | | | |
| govern | | | | | |
| believe | | | | | |

The clusters above revealed positive comments on the outgone election in Ekiti State, Nigeria. Most of the comments were based on congratulating the Governor-elect of Ekiti State and the President on the success of the election. The masses liked the way the election was conducted: it was very free and fair. Respect was given to the President (Jonathan) by the way the political election was conducted. There was security and no rigging and killing. From people’s point of view, it was a true democracy (People were educated about what democracy in Nigeria is all about). However, Government was intimated about high rate of killings happening in the country and an advice was given to the President to deploy soldiers to Sambisa forest and rescue the abducted Chibok girls as this election was a huge success.

## 4. Conclusion and Future Work

In this work an e-governance framework was proposed and implemented, which captured people's opinions and feedback (opinion pool from various sources). The process involved converting these text data to structured data. Text mining techniques were used on the structured database that resulted from the feature extraction of those text data. The extracted features were then clustered for knowledge discovery using K-Means clustering algorithm. This knowledge obtained from the clustered text data can be used for strategic decision support for e-governance. We were able to achieve only Phase 1 (Publish-Text Mining Interaction) in this work. Interaction of text mining with other phases (Interact and Transact Phases) will be addressed in future work.

## References

[1] M. W. Berry. 2004, "Automatic Discovery of Similar Words, in Survey of Text Mining: Clustering, Classification and Retrieval", Springer Verlag, New York. LLC, 24-43.

[2] M. Hearst, 1999, "Untangling Text Data Mining," in the Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, 1999.

[3] J. Dorre, P. Gerstl and R. Seiffert R, 1999, "Text Mining: Finding Nuggets in Mountains of Textual data", In Proc. 5th ACM Int. Conf. on Knowledge Discovery and Data Mining (KDD-99), pages 398-401, San Diego, US, 1999. ACM Press, New York, US

[4] A. Padmapriya, 2013, "E-Governance: A move towards paperless Administration in India", International Journal of Computer Trends and Technology. 4(3) 2013 http://www.internationaljournalssrg.org

[5] D. Robb, "Taming Text", 2005, Retrieved from http://vnweb.hwwilsonweb.com/hww/jumpstart.jhtml?recid=0bc05f7a67b1790e8bd354a88a41ad89a928d23360302a4959035699f17e2ba8a63e2dd032c73f8a7fmt

[6] P. Cerrito, 2005, "Inside Text Mining", Retrieved from http://wilsontxt.hwwilson.com/pdffull/06619/275n6/gs9.pdf

[7] L. Francis and M. Flynn, 2010, "Text Mining Handbook", Casualty Actuarial Society E-Forum

[8] Center for Democracy and Technology, 2002, E-Government Handbook. Retrieved on February 28, 2014 from http://www.cdt.org/egov/handbook/

[9] A. Al-Hashmi and A. B. Darem, 2008,. "Understanding Phases of E-Government Project", Emerging Technologies in E-Government, 2008. Pp. 152-157

**Dr. Adesesan Barnabas ADEYEMO** is a Senior Lecturer at the Computer Science Department of the University of Ibadan. He obtained his PhD, M. Tech., and PGD Computer Science degrees at the Federal University of Technology, Akure. His research interests are in Data Mining, Data Warehousing & Computer Networking. He is a member of the Nigerian Computer Society and the Computer Professionals Registration Council of Nigeria. Dr Adeyemo is a Computer Systems and Network Administration Specialist with expertise in Data Analysis and Data Management.

**Adebola K. OJO** is a PhD student in the Department of Computer Science, University of Ibadan, Nigeria. She is a registered member of the Computer Professional of Nigeria (CPN). She had her Masters of Science Degree in Computer Science from University of Ibadan, Nigeria. Her research interests are in Digital Computer Networks, Data Mining, Text Mining and Computer Simulation. She is also into data warehouse architecture, design and data quality via data mining approach.