International Journal of
Science and Engineering Investigations vol. 6, issue 62, March 2017

ISSN: 2251-8843

A Framework for Knowledge Discovery in the News Media Using Text Mining Technique

Oladejo Bolanle[1], Babajide Olanrewaju[2]
[1,2]University of Ibadan, Computer Science Department
([1]oladejobola2002@gmail.com, [2]juneeagle@yahoo.com)

Abstract- The increase in the media information available to the news audience is growing at an alarming rate and this calls for an efficient means of analyzing and mining information that can be used in making vital decisions. This is most especially true in Nigeria where there is an alarming increase in the crime rate that needs to be analyzed. The work aimed at the use of text mining techniques for deducing the rate of occurrence of crime in Nigeria cities. A review of various text mining techniques with their comparative advantages is presented. The work modeled a text mining framework for the analysis, extraction and organization of news media objects. Mathematical and algorithmic text mining techniques were applied in the development of the crime analysis system. Information was captured from online news material and on ontology was built for representation of specific domain of crime using frames knowledge representation language and Protégé ontology builder. The k-means algorithm is used for knowledge discovery to extract relevant information from the news corpus. The applied clustering technique showcased resulted patterns with graphical representation that aids deductions of the rate of occurrence of crimes. In conclusion, the approach is found useful for sociological analysis of occurrence of crime which invariably aids decision makers on national security.

Keywords- *Text Mining, Crime, News Corpus, Ontology*

I. Introduction

The massive heap of information generated on a daily basis triggers a lot of concern as to the management of such rich and voluminous resources, in order to harness a good and optimal information gathering and retrieval system that will assist in making a veritable decision in our day to day business operation. There is increase in the availability of information in different sectors of life: in the academic, journalistic and even in the social world, information seems to be flowing everyday more than ever before. In fact, the fact that we are living in the information age is a forgone conclusion. For example, on the internet, there is more information available than ever before and also most big multinational companies like MTN, GLO, Cadbury Nigeria Ltd, etc. generate massive amounts of numerical data, which if analyzed will be of benefit to the companies themselves and the general public. Although much research has gone into data mining in recent years, there is the increasing awareness that most of the information that people access is textual and therefore the need for textual knowledge discovery techniques.

In fact, 80% of the information on the internet is textual. This realization has given birth to the field of study called Text Mining. Text Mining (TM) is an emerging field of study that involves the mathematical and scientific analysis textual data for discovering potential hidden information. Text Mining can help in organizing the large swarm of data that is available everywhere. TM is defined as knowledge discovery in textual databases which allows us to create a technology that combines a human's linguistic capabilities with the speed and accuracy of a computer [1]. Since news (as in print and online news) are the primary way for getting information nowadays and this source of information is increasing at an alarming rate, it will not be possible for a human user to analyze all the news in the world. This constraint invariably results to difficulty in making inference from news for the purpose of crime analysis, commercial analysis or whatever purpose he or she wants to achieve,. Hence, calling for an automated system of analyzing this textual information, as demonstrated here in this research. So it is important to analyze this information in order to gain knowledge or even more information. The technology of text mining makes the information age more effective and powerful.

Sequel to the significance of news towards national development, this paper aimed at review of text mining techniques with the intent to apply the appropriate and computationally effective techniques to analysis of news media. The rest of this paper is divided into five sections. Section 2 considers the relevant theoretical background, review of text mining frameworks with related works and comparison of text mining techniques are presented in Section 2.1 and 2.2 respectively. Section 3 focuses on modeling of the text mining framework presented in this work while discussion of research results is considered in section 4. The paper is concluded in section 5 with further work recommendation.

II. Theoretical background

A. *Text Mining*

Text mining is about discovering knowledge in text and it uses many methods from statistics, linear algebra, and Artificial Intelligence. In text mining, textual data is represented with a mathematical model such that mathematical and algorithmic operations can be performed on the data. The first state in any text mining operation is that the document is preprocessed and refined. Then the document is represented in a vector space model in which the words are in the rows and the documents are in columns, but this is not fixed as sometimes it is the other way round. After this stage, various data mining tasks can then be carried out on the text in order to analyze the textual information. Clustering, classification and other types of learning methods can be applied to the structured text in order to deduce patterns from the data. Text Mining is not distinct from data mining as it uses many techniques from data mining. For instance, both types of systems rely on preprocessing routines, pattern-discovery algorithms, and presentation-layer elements such as visualization tools to enhance the browsing of answer sets [2].

The most basic thing in any text mining operation is a collection of documents or a corpus. There are two types of document collections. The first is static and the second is dynamic. In the static collection, the initial document collection remains unchanged while in the dynamic collection, the document collection is updated and changed from time to time. Text Mining systems rely on algorithmic and heuristic approaches to consider distributions, frequent sets, and the various ways in which concepts are associated at an inter-document level. The news collection we intend to analyze is basically a dynamic collection because new reports keep coming out on daily basis. There are different algorithms and technologies applied in the process of text mining as discussed subsequently.

B. Review of text Mining Frameworks with Related Works

A lot of researches have proposed different methodologies for building text mining frameworks for different purposes especially for text mining. In [3] a text mining application was developed using data structures which is relatively easy to extract from text. To be more precise, it uses the concept hierarchy as a central data structure. A framework that learns the target ontology from text documents and then uses the same target ontology to improve the effectiveness of both supervised and unsupervised text categorization was built by [4]. It is an approach that integrates ontology learning and text mining. It essentially has methods for automatically constructing ontologies and then exploiting the ontologies themselves for mining text. Also, [5] built a framework for the health industry that can support search, discovery and trending of patient characteristic in order to rapidly transform from data collection to an understanding of a patient's health trajectory. Essentially, the framework leverages the use of skip bigrams or S-grams. S-grams are word pairs in their respective sentence order that skip over words. In [6] a framework that can identify key actors in a body of news and the actions they performed was built. The framework extracts Subject-Verb-Object triplets from text by means of a parser and then constructs a semantic graph that captures narrative structures and relations contained in a text. Also, [7] used a graph based approach to mine patterns from text. The graph was represented as a binary Matrix from which a directed graph is created and association rule mining graph based approach is performed on the matrix. It was observed that in [8], a framework for the world of online newspapers was presented. Essentially, the framework extracts individual news items from the web pages and mines them separately. The main pages of the web newspapers are retrieved at regular intervals and various pattern identification algorithms were used to identify the text on the web page. In [1], a framework for text mining from newspapers using the Generating Association Rules Based on Weighting Scheme (GARW) algorithm was designed and some interesting patterns were generated. A framework that uses text mining techniques to identify a strong sequence of events by examination of word frequencies was designed by [9]. In [10] a framework that encompasses all the major data mining, natural language processing operations such as classification, clustering and speech recognition and recognition of language was designed. Also, [11] designed a framework for clustering newspaper articles and it essentially helps to solve the problem of differentiating the different articles in the newspaper.

C. Comparison of Text Mining Techniques

Quite a number of text mining algorithms are available, such as, the K-Means algorithm, hierarchical and spectral clustering amongst others. The conducted review shows that the K-Means algorithm is better because it is not computationally demanding unlike the hierarchical clustering algorithm which requires time and memory space. Also, hierarchical techniques that are agglomerative in nature perform poorly for textual documents because two documents can often be nearest neighbors without belonging to the same class. Also due to the probabilistic nature of how words are distributed, any two documents may share any of the same words. To make matters worse, because of the way hierarchical clustering works, these mistakes cannot be fixed once they happen [12]. We also compared the K-Means algorithm with the spectral clustering algorithm and we realized that the problem with spectral clustering algorithm is that they are computationally demanding and in general, the memory consumption is too high. This is because computing or even approximating eigenvalues and eigenvectors is not fast for all graphs and hence such methods may face scalability issues when applied to data. Besides, it is not possible to introduce fuzzy methods into spectral clustering since most graph clustering algorithms produce clusters that are too exact and allow objects to belong to only one cluster. Within graph clustering, relatively not much work has been done on fuzzy graph clustering. This is unlike the K-means algorithm in which the memory space requirements are modest and the time requirement for K-means is linear. The advantage the K-means algorithm has over some other algorithms is shown in Table 1.

Sequel to the trade-offs between the K-Means algorithm and afore-mentioned algorithms, K-Means algorithm was chosen as the text mining technique for clustering news media in the context of this work.

III. METHODOLOGY OF KNOWLEDGE DISCOVERY OF CRIME PATTERNS FROM NEWS MEDIA

The choice of methodology adequate for this work is Natural Language Processing (NLP). NLP is the branch of linguistics which deals with computational models of language. It involves processes such as Stemming, Lemmatization, Part of Speech Tagging and stop word removal. The K-Means algorithm was chosen for clustering textual documents in the context of this work.

Knowledge Miner Framework (KMF) depicted in Fig. 1 was modeled to represent the set of processes involved in discovery of the pattern of criminal events in a nation from the Nigeria tabloids.

One major advantage of the system based on this framework is that the information being retrieved is more accurate due to the use of Latent Indexing (LSI) technology. LSI minimizes errors such as synonyms and antonyms during information retrieval. As was mentioned before, text mining is not a single procedure or algorithm but is rather a collection of different algorithms, techniques and procedures. Implementation wise, the framework consists of the following stages:

A. The Ontology Building Stage

Here ontology was built for the framework using Protege 3.4.8 software. It is used for the representation of all the events and people in the news corpus.

B. The Preprocessing Stage

Here lemmatization, stop word removal and other operations were performed on textual document in order to structure the document for proper mining.

C. The Representation Stage

Here the normal Vector Space (VSM) was used to represent the documents in the corpus but then there is a choice of modifying the VSM with the Latent Semantic Indexing technique which makes the vector space model more accurate. The Latent Semantic Indexing technique is about dimensionality reduction, that is, the dimensionality of the VSM is reduced using matrix operations. Essentially, the VSM is represented as a matrix and then reduced into three different matrices as in (1):

$$X = TSD^T \qquad (1)$$

TABLE I. COMPARISON OF K-MEANS AND SOME TEXT MINING ALGORITHMS

Spectral	K-Means	Hierarchical
Computationally Demanding	It is not computationally Demanding	Computationally Demanding as it consumes memory and takes more time.
The clustering is hard as it does not allow for different levels of membership or fuzzy logic	Although the clustering is hard, it can be easily modified to allow fuzzy logic as in the fuzzy k-means clustering algorithm	It allows fuzzy methods
	The clustering Algorithm performs better	The clustering Algorithm performs poorly

Where:

X = the original vector space model or matrix

T = the matrix of singular vectors or a matrix whose columns are the eigenvectors of the $\boldsymbol{TT^T}$ matrix which is the matrix(T) multiplied by its transpose

S = the diagonal matrix of singular values or a matrix whose diagonal elements are the singular values or eigenvalues of T

D^T = the matrix of right singular vectors or the transpose of the eigenvectors of T.

LSI projects an original vector space or term-document matrix into a small factor space [13]. LSA also offers a close enough approximation to human knowledge and so is better as a way of representation [14]. The surprising thing about Latent Semantic Indexing is that with this method, the document can make deductions on its own. The mathematical technique can draw out indirect inference which the Vector Space Model cannot. Besides, experiments have shown that the Latent Semantic Indexing technique improves the performance of document clustering.

D. Clustering Phase

Here, the document is clustered using the K-Means clustering algorithm. The K-Means algorithm is a simple algorithm for clustering documents.

E. The Crime Analysis Stage

This phase uses algorithms derived from simple human logic and mimics the way in which researches in the real world find information.

Information derived from newspaper archives served as the textual documents which were clustered. We used the website called www.newspaperarchives.com and accessed some newspaper articles from various American newspapers that talked about crime and violence in Nigeria.

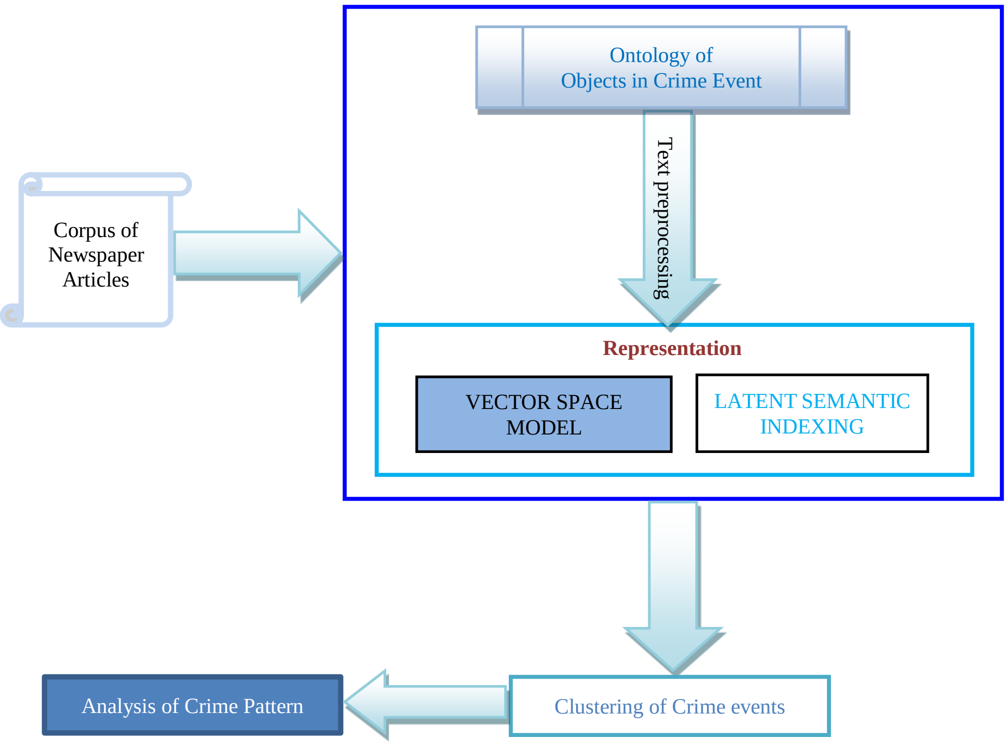


Figure 1. Knowledge Miner Framework (KMF)

IV. RESULT AND DISCUSSION

In this work, the ontology of basic entities in the context of violence occurrence was built with Protégé software as a medium of knowledge representation. The Ontology was hierarchical in nature. A graphical view of the ontology is shown in Fig. 2.

Information derived from newspaper archives served as the textual documents which were clustered. We used the website called www.newspaperarchives.com and accessed some newspaper articles from various American newspapers that talked about crime and violence in Nigeria. Fig. 3 presents a snap shot of an excerpt from news documents containing description of violence scenarios or reports.

In these documents, there are various news reports about violence in different parts of Nigeria and the aim of this work was to find the most commonly occurring phenomena in the news. Phenomena such as the state in which it most occurs or the type of violence that mostly occurs in the time period specified. The output was transformed into a visual model so as to make it easy for the user to analyze it. The output after the document is clustered is shown in Fig. 4.

For our experiment or test, we selected 7 states namely Abuja, Bauchi, Borno, Cross River, Enugu, Oyo, Zamfara. From the result, it is deduced that violence takes place mostly in Abuja, Bauchi and Borno states (based on the periods of selected past news archive).

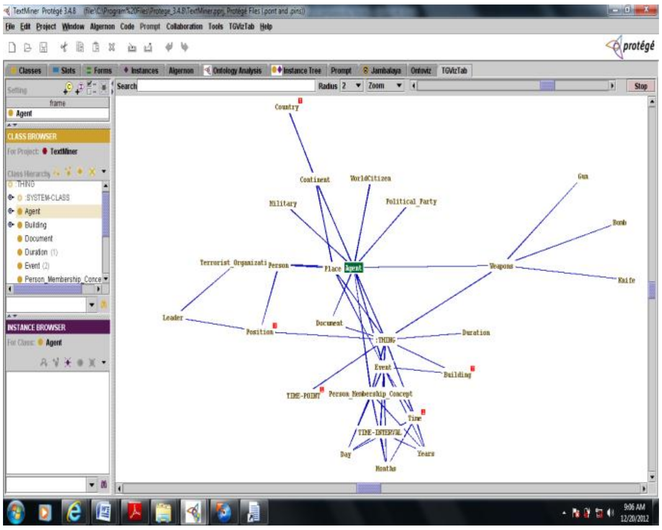
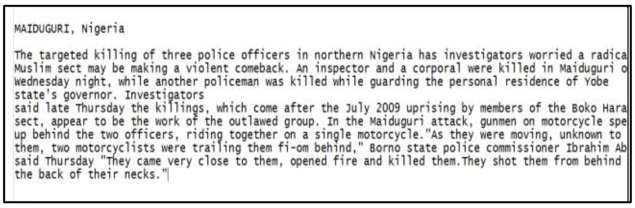


Figure 2. Ontology of relevant objects in news corpus



```
MAIDUGURI, Nigeria

The targeted killing of three police officers in northern Nigeria has investigators worried a radica
Muslim sect may be making a violent comeback. An inspector and a corporal were killed in Maiduguri o
Wednesday night, while another policeman was killed while guarding the personal residence of Yobe
state's governor. Investigators
said late Thursday the killings, which come after the July 2009 uprising by members of the Boko Hara
sect, appear to be the work of the outlawed group. In the Maiduguri attack, gunmen on motorcycle spe
up behind the two officers, riding together on a single motorcycle."As they were moving, unknown to
them, two motorcyclists were trailing them fi-om behind," Borno state police commissioner Ibrahim Ab
said Thursday "They came very close to them, opened fire and killed them.They shot them from behind
the back of their necks."
```

Figure 3. An excerpt from part of the news documents

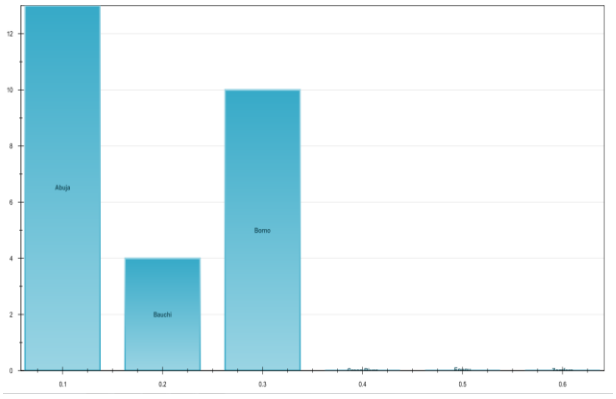


Figure 4. A visual model of clustered news documents

V. CONCLUSION

The knowledge mining framework proposed in this work is implemented for extraction of textual information from media news corpus and/or archives. This framework is a system which shows how text mining can be useful in sociological analysis and for organizing documents for various purposes. It is efficient for mining textual information with some advantage over the existing frameworks with the application of Latent Semantic Indexing (LSI) technology which minimizes errors such as synonyms and antonyms during information retrieval.

The text miner system developed in this work is only a pilot implementation of the generic framework. This framework includes ontology and we will like to mention the fact that ontology is becoming increasingly important in the field of text mining. For future research, we plan to utilize ontology to the core mining stage of the system, such as, clustering, association rule mining and related algorithms.

REFERENCES

[1] I.T. Fatudimu, A.G. Musa, C.K. Ayo, A.B. Sofoluwe, “Knowledge Discovery in Online Repositories: A Text Mining Approach”, European Journal of Scientific Research 2008

[2] R. Feldman and J. Sanger, “The Text Mining Handbook”, Cambridge University Press, 2007.

[3] R. Feldman, I. Dagan, “Knowledge Discovery in Textual Databases (KDT), Math and Computer Sceince Dept. Bar-Ilan University Ramat-Gan, ISRAEL 52900, 1997

[4] S. Bloehdorn, P.Cimiano, A. Htho, S.Staab, Institure AIFB University of Kalruhe, KDE Group University of Kassel, 2004.

[5] R.M. Patton, C.C. Rojas, B.G. Beckerman, T.E. Potock, “A Computational Framework for Search, Discovery and Trending of Patient Health in Radiology Reports”, Computational Sciences and Engineering Division Oak Ridge National Laboratory Oak Ridge, TN, USA, 2011

[6] S. Sudharhar, R. Fanzosi, N. Cristianini, “Automating Quantitative Narrative Analysis of News Data”, JMLR: Workshop and Conference proceedings 17(2011) 63 2nd Workshop on Applications of Pattern Analysis, 2006

[7] D.S. Rajpu, R.S. Thakur, G.S. Thakur, “Rule Generation from Textual Data by using Graph Based Apporach”, International Journal of Computer Applications, 2011

[8] K. Norvag, R. Oyri, “News Item Extraction for Text Mining in Web Newspapers”, Department of Computer and Information Science Nowergian University of Science and Technology, 2004

[9] R. B. Allen, “Improving Access to Digitized Historical Newspapers with Text Mining, Coordinated Models, and Formative User Interface Design”, College of Information Science and Technology Drexel University, Philadelphia, 2010

[10] S. Lee, J. Song, Y. Kim, “An Empirical Comparison of Four Text Mining Methods”, Journal of Computer Information Systems, 2010.

[11] A.K. Ojo, A.B. Adeyemo, Framework for Knowledge Discovery from Journal Articles Using Test Mining techniques, African Journal of Computing & ICT,vol. 5, pp.33-42, 2012.

[12] M. Aiello and A.Pegoretti, “Textual Article Clustering in Newspaper Pages”, Dept. of Information and Communication Technologies, Universita di Trento, 2006.

[13] M. Steinbach, G. Karypis and V. Kumar, “A Comparison of Document Clustering Techniques”, Department of Computer Science and Engineering, University of Minnesota, 2000.

[14] N. Sumathi and V. Chittu, “A Modified Genetic Algorithm Initializing K-Means Clustering”, Global Journal of Computer Science and Technology, 2011.

[15] C. Luo, Y. Li, and M. Chung, “Text document clustering based on neighbors”. Data & Knowledge Engineering, vol. 68, 2009.

[16] T. K. Landauer, P. W. Foltz and D. Laham, “an Introduction to Latent Semantic Analysis”, Department of Psychology, Campus Box 345, University of Colorado, 1998.

[17] S.E. Shaeffer, “Graph Clustering”, Laboratory for Theoretical Computer Science, Hesinki University of technology, Finland, 2007.