Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 01, January-2020

# An Enhanced Document Referential, Classification And Retention System using Muiltinomial Naive Bayes Algorithm

*O.O. Anyiam[1], L.N. Onyejegbu[2] and F.E. Onuodu[3]
[1-3] Department of Computer Science, University of Port Harcourt,
Port Harcourt, Rivers State, Nigeria

***Abstract*— Securing organizational data is an important information management issue that continues to pose significant challenges for organizations especially in developing countries. Different organizations have made several failed attempts to develop authentic solution that will centrally create and manage document referential, place documents in their respective categories and apply retention on the documents during end-of-live, in a seamless manner. There have been several complaints from users on issues bordering on missing documents, premature destruction of important documents, organization's top-secret documents been found in wrong hands and placement of documents in wrong categories as a result of no well-defined organizational policies. We developed an enhanced online document referential, classification and retention system, which combated these challenges. We used object-oriented analysis and design methodology (OOADM) in our approach. Microsoft ASP.Net and Python technologies were used for this implementation. Automated document referential, classification and retention system provides a platform for easy creation of document referential, placement of documents in respective document categories and automatic application of retention on documents at the end of their lifespan. From the results, the overall accuracy of the classification model is 88% which indicates that most predictions made by the model are correct. The specificity of the individual classes ranges between ninety-eight (98) and one hundred (100) percent, whereas their precision, recall and f1-scores are above 0.5 which is good for prediction. This work could be beneficial to both small, medium and large-sized organizations.**

***Keywords*— *Document, Referential, Retention, Classification***

## I. INTRODUCTION

Placing documents in categories is a rare practice among organizations in developing countries. We tend to chunk out heavy documents on daily basis with little or no attention to what happens to such document throughout its life span. This has exposed many organizations, both in private and public sectors, to great danger of confidential documents reaching wrong hands. In some organizations, documents are short-lived or even stay longer than required, thereby exposing the organization to litigations by parties concerned or fines by regulatory authorities. One of the ways of professionally addressing this problem is by properly categorizing organizational documents, developing policies that govern each category and judiciously enforcing these policies.

Categorization is the process of placing documents into classes. A category is chosen considering the relation between the subject of the category and the document belonging to it [1]. Lots of information are being generated and stored in various hardcopy and electronic forms. There is need for a proper classification of these documents in order to apply accurate retention policy meant for each document. Document Retention is the holding (period) of records/documents for further use [2].

Before a document can be placed in a class, a document type referential (or schedule) must first be created. [3] observed that a good policy comprises of a schedule that has the retention periods for all documents types and a framework that is used in administering it.

Document referential and retention concerns a cycle of organizational activities which include:

i. The creation of document type referential that clearly states the various categories of documents and how to handle documents in a particular category.

ii. The acquisition of documents from one or more sources and placement of such documents in the accurate category;

iii. Its ultimate disposition through archiving or deletion.

Many times in an organization, documents outlives those that created them. The person who created a document may be transferred to another location, sacked or may even resign. Another staff who picks that document needs to have a proper understanding of the type of document he is handling and with the help of a referential apply correct company policy for such a document.

As documents and database records generated by organizations increases enormously, the complexity of effectively handling these documents by users also increases. There is a problem of having a single tool that can standardize document types, effectively place documents into their appropriate categories and apply organizational retention policies on these documents, especially those in storage media and records in multiple databases and varying DBMS platforms. Some works have been done to overcome these problems. However, much is still required. We have developed an enhanced system for document referential, classification and retention for organizations.

Document referential, classification and retention system targets organization of different sizes especially big organizations that chunks out big electronic documents and database records regularly.


(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 01, January-2020

## II. RELATED WORKS

[4] developed an automated system that detects articles that are relevant to disease outbreaks using Machine Learning classifiers. The experiment recorded daily averages of areas under ROC curve is 0.841 for Naive Bayes and 0.836 for SVM classifier (equivalent to 95% confidence interval). The experiment did not explore other classification algorithms and not tested for large datasets.

[5] proposed a two-phased feature selection method and Naive Bayes classifier for Indonesian news classification. The method showed 86% accuracy and lowered the complexity of Maximal Ma 0rginal Relevance for Feature Selection (MMR-FS). [6] reviewed supervised machine learning classification techniques, based on a number of machine learning application-oriented papers. Naive Bayes ranked high in needing less training data, using little storage space; robust to missing values and quick training.

[7] carried a study that compared classification algorithms. The work concluded that the performance of algorithms depends on domain. No one algorithm fits all classification domains. Though he recommended the Random Tree and Logistics Model Tree. It suffered from limited coverage. [8] designed a framework that is based on generating mock examples for self-labelled classification. This framework improved the classification capabilities of self-labeled techniques.

[9] proposed a framework for records retention in relational database systems. This framework applied retention on views in the database. It has no automatic classification. The use of database views has deficiencies since it's not all views are updateable, integrity preservation, cost and what happens to existing databases. Also, it considered records in the databases only and does not allow multiple DBMS platforms. [10], extended RDBS to automatically enforce privacy policies. The work monitored privacy obligations enterprise-wide using an elaborate central obligation monitoring system. A systematic way of scheduling events throughout all corporate data repositories such that the execution of these events will ensure compliance with all privacy obligations. It did not cover retention on others records.

The goal of this paper is to develop an automated document referential, classification and retention system that will be used to standardize document categories, place documents into appropriate categories and apply retention on these documents using organizational policies that were defined in the referential. The system automatically applies retention on records in multiple databases and different DBMS platforms (SQL Server and Oracle) based on the classification and retention schedule. It also places documents into the standardized document types using user-defined options and multinomial naive Bayes algorithm

## III. METHODOLOGY

The methodology adopted by the researcher is Object Oriented Analysis and Design Methodology (OOADM). OOADM involve the identification of critical objects of the document referential, classification and retention system by breaking them down into smaller sub-systems and recursively applying software processing on the identified objects.

### *A. Architecture of the Proposed System*

The architectural design in Fig. 1 describes the components integrated to bring about the workings of the document referential and retention application. It captured the major functional building blocks needed to understand the process of building the document retention system. These components are explained below:

*Generate Contents:* This refers to a collection of different inputs made into the system by staff and non-staff of the organization. It involves records stored in database and conventional files like word and excel document stored in directory.

*Document Referential or Retention Schedule:* This is an elaborate set of document types stored which the classification and retention modules references. It contains government and organizational policies on what should happen to documents during its life cycle. For example, final treatment of a document at end-of-live could be to destroy the document or archive for historical or legal purposes.

*Multiple Data Source:* The data source describes how the document retention module gets its data. It is a connection setup that enables the retention module access the records in the database and conventional files on disks.

*Files in Directory:* This refers to documents in various file formats that are stored in known paths in the disc.

*Records in Database:* Records in database refers to relational records stored in various DBMS platforms like SQL Server and Oracle. The retention system captures the connection strings to these records applies retention on datasets at end-of-live.

*Document Classification:* This module involves placing of each document into a particular document type or class using user-defined or naive Bayes algorithm.

The document classification service performs the calculation of placing a document into existing category using multinomial naive Bayes model given as:

$$P(c|x) = \frac{P(x|c)\ P(c)}{P(x)}$$

$$P(c|x) = P(x_1|c)\ x\ P(x_2|c)\ x\ \ldots\ x\ P(x_n|c)\ x\ P(c)$$

Where:

P(c|x) is the posterior probability of class (target) given predictor (attribute).

P(c) is the prior probability of class.

P(x|c) is the likelihood which is the probability of predictor given class.

P(x) is the prior probability of predictor.

*Document Retention:* This module enforces the retention on the records in the database or files stored in directory using the preferences and criteria slated earlier by the user. This may involve exclusion of documents or migration into new formats.


(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 01, January-2020

*Retention Information and Logs:* This contains basic information about the retention system and also a log of activities that are going on within the system.

*View Retention Reports:* Reports can be generated using the logs and other information that are kept in retention system database itself.

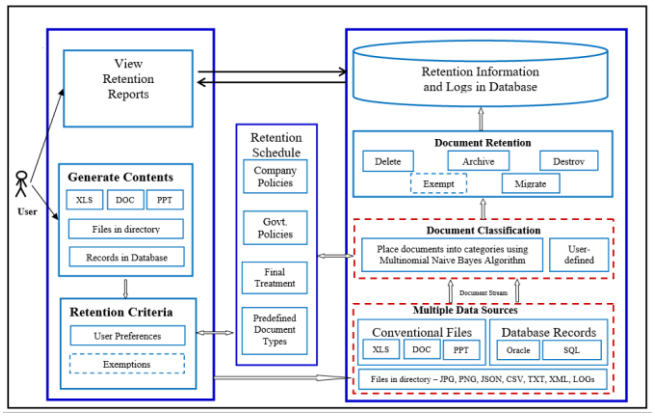


Fig. 1. The Architecture of the Proposed System

## B. *Use Case Diagrams of the Proposed System*

Fig. 2 is the Use Case diagram that shows a list of actions or event steps that defines the interaction between the actors and the enhanced document referential, classification and retention system.

The user generates documents while performing his day-to-day activities and assign preferences (if required). The classifier places these documents into document types based on the user preferences and retention schedule created by the administrator or an Information Management Officer (IMO). The retention algorithm is triggered daily, based on scheduled time to identify document that has reached end-of-life (or met certain criteria) and decides either to delete, exclude, archive or recommend for migration into another format or new platform. The use case trigger for the proposed system is the user's activity of generating documents and assigning preferences to such documents where necessary.

## C. *Data Flow Diagram of the Proposed System*

We used data flow diagram to diagrammatically show how data flows within the enhanced document referential and retention system. In Fig. 3 we represented the proposed document referential and retention system which has one external entity i.e. user of the system namely the Organization and the data flowing in and out of the system is the documents details.

Fig. 4 shows the Level-1 DFD which models the details of the proposed system. It shows how the system is divided into processes. Each process handles one or more of the flows of data, either to or from the Organization.

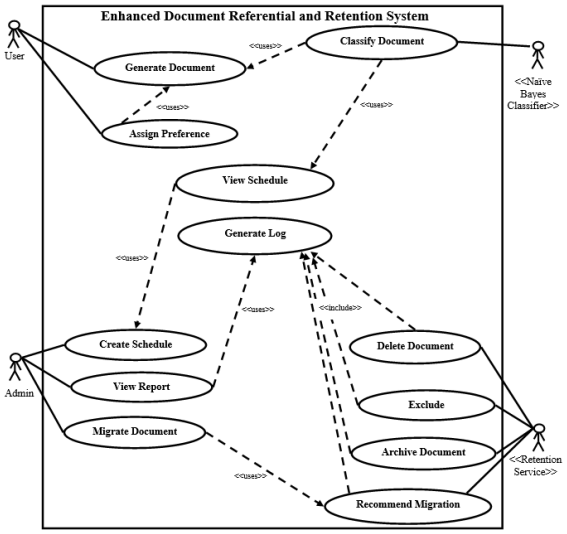


Fig. 2. Use Case Diagram of the Proposed System

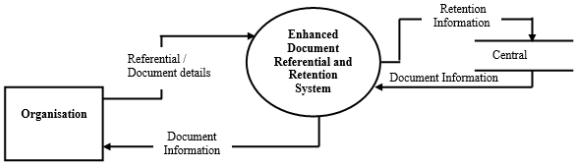


Fig. 3. The Context Diagram of the Proposed System

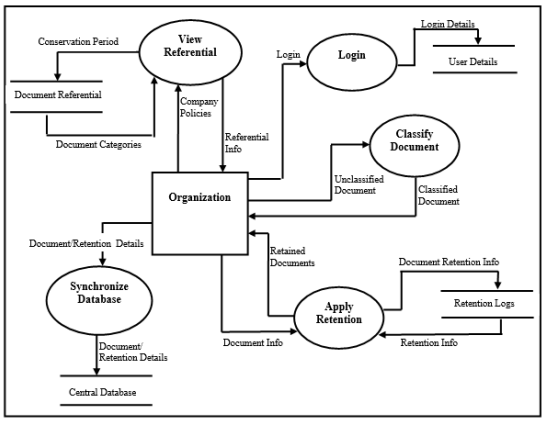


Fig. 4. The Level-1 Diagram of the Proposed System

## D. *Algorithm of the Proposed System*

The steps used in the automated document referential, classification and retention system include the following:

1. Confirm that user exists in the system

2. If user exists, grant him access to create contents in multiple document formats (like .doc, .xls, .ppt), directories with multiple files of different format and records in databases of different DBMS like SQL Server and Oracle.

3. For each content created by the user, provide user with the retention criteria interface and obtain user preferences on what should happen to the contents during its life cycle.

4. Ensure each user preferences matches with what is contained in the organization's retention schedule.


(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 01, January-2020

5. Establish a connection to the document retention system using a multiple data source interface that allows conventional files, database records and different file formats in directory.

6. Pass the contents through a Naive Bayes classifier using the steps below [11]:

i. Load the Training set as $T_s$

ii. Let Class $C_i$ = Folders in $T_s$

iii. Let $D_{Trg}$ = set of labeled documents contained in each folder in Ts

iv. Set $D_{Trg} = \{w_1, w_2 \dots w_n\}$ where $D_{Trg}$ is list of words from Documents in

Training set and $w_n$ is the nth word in the $D_{Trg}$

v. Total w in $C_i$ = Count ($W_i$ in each class)

vi. Total w in $T_s$ = Count ( $W_i$ in Training set)

vii. $P(C_i)$ = (Total documents in $C_i$) / (Total documents in $T_s$)

(i.e. Prior probability of a document appearing in each class c)

viii. Load the unlabeled document as $D_{ul}$

ix. Set $D_{ul} = \{w_1, w_2 \dots w_n\}$ where $D_{ul}$ is list of words from unlabeled document and $w_n$ is the nth word in $D_{ul}$

x. $P(C_i|\text{document}) = (P(C_i|w_1, w_2 \dots w_n) / n$

(i.e. Probability of the document to belong to the particular class and n is the total words in the input document)

xi. $P(w_j|C_i)$ = (1+ Frequency of $w_j$ in class $C_i$) / (Total w in $C_i$ + Total w in $T_s$)

xii. $P(C_i|\text{document}) = \max (P(C_i) * P(w_j|C_i)) /n$

(i.e. Assign class $C_i$ to the document if it has maximum posterior probability with that class)

7. Apply retention on the classified document or content:

a. Check if the record has reached its end-of-life.

b. If 'yes', exclude record, delete record permanently, archive for historical purpose or migrate into a new format or technology depending on the document class and user preferences.

c. If 'no', skip the document and continue.

8. Log the retention activities

9. Allow user (administrator) to generate reports based on the retention activities.

*E. Program Flowchart*

The program flowcharts in Fig. 5 shows the program structure, logic flow and operations performed by the proposed system.

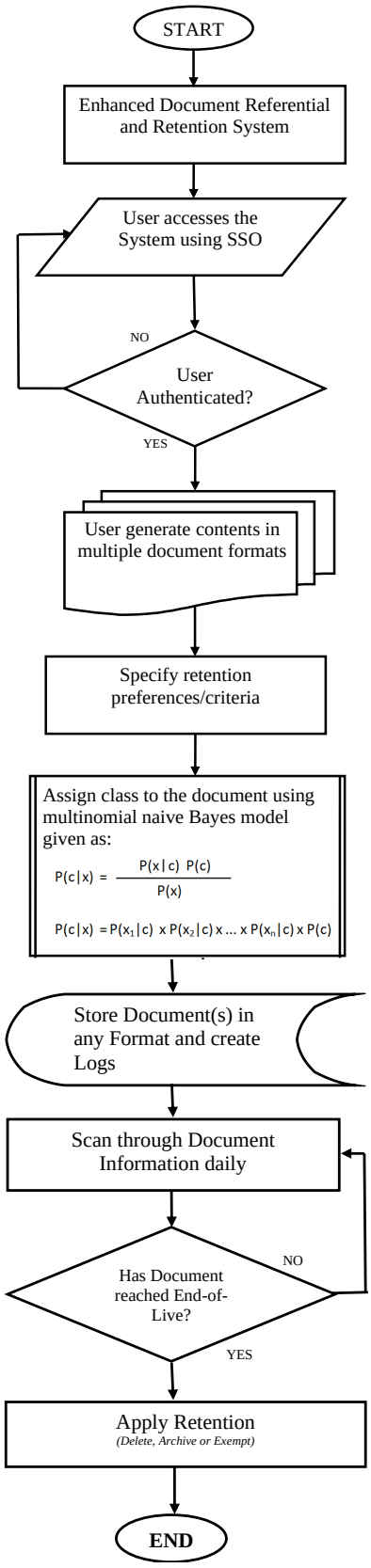


Fig. 5. Program Flowchart


(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 01, January-2020

## IV. RESULTS AND DISCUSSION

The proposed system was implemented using Microsoft Visual Basic .Net, Python and SQL Server T-SQL and Oracle PL/SQL. The IDEs used are MS Visual Studio .Net 2013 and MS-SQL Server Management Studio 2014. They all run on .Net framework which uses Common Language Runtime architecture to manage execution of codes. Python was used for development of the classification service. The classification service is a windows service that performs the automatic classification aspect of the system using multinomial naive Bayes classifier. JavaScript, JQuery, CSS, HTML, JSON and Bootstrap technologies where used for scripting, styling, rendering of documents and remote communication between application and the databases.

### *A. System Testing*

The accuracy of the program was tested with some varying data. The 20NewsGroup dataset [12] which had duplicates removed, was used as training and test dataset. They are pre-processed. Fig. 6 shows that when the document "C:\...\Nora Roberts - Loving Haley.doc" was selected from Open-File dialog box, the status was *"Unclassified. Predicted Class is DT-0016"*. On clicking the classify button and responding *'Yes'* to the ensuing dialog box, the document gets classified as shown in Fig. 8.

Fig. 9 indicates that the status of the same document changed to *"Classified. DT Code: DT-0016. Value: L"* when selected again in the system. This document will be deleted from the system in 10 years because the conservation value for such type of document is 10years.

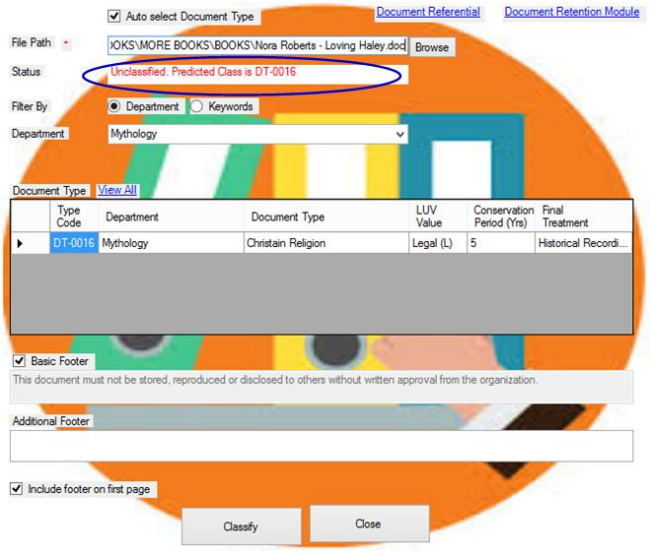


Fig. 6. Selected non classified document

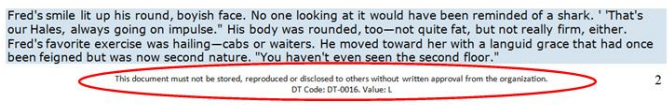


Fig. 8. Showing the footer of the selected document after classification

Fig. 10 shows the implemented document referential. Users can search a particular document type using the search button or the departmental dropdown list.

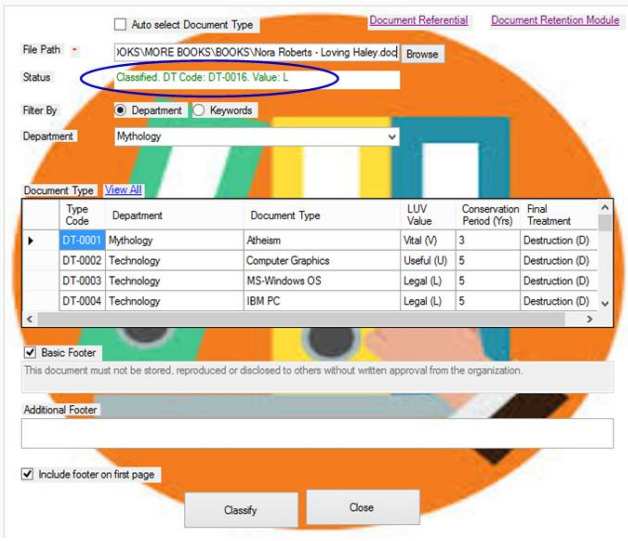


Fig. 9. New Status of the selected Document after Classification

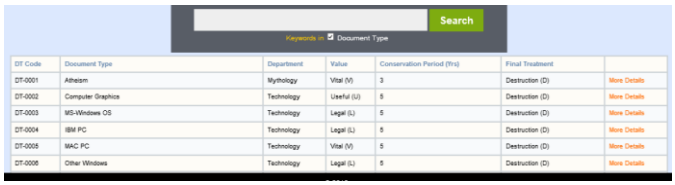


Fig. 10. Document Retention Policies (Referential)

The system has a configuration page for records stored in different databases and different DBMS platforms. It captures the connection strings and other information that will enable the retention service to know how to handle such records. Many databases from two different DBMS platforms, Oracle and SQL Server, were configured and activated for retention using the database records retention configuration module. The database records retention configuration module in Fig. 11, was used to configure a banking software which uses the 'NKPO.mdf' database. The system was able to pull all the existing tables from the database and for each table selected, it listed all the date fields (column name) and identified the primary key. When the Save button was clicked, the setup request was saved in the retention database and listed in a grid. Fig. 12 shows that when the link Delete/Exempt Record was clicked, it displayed all the activated database records retention requests. Clicking on the Select link beside each request displayed the details of the request, its referential details and all the records that have reached their end-of-live and are pending deletion or exemption.


(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 01, January-2020

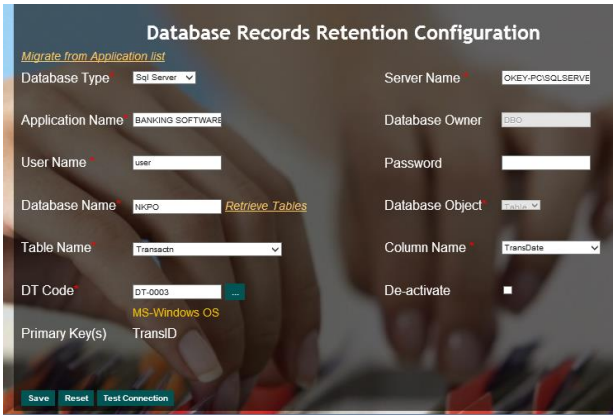


Fig. 11. Records retention configuration for 'NKPO' database

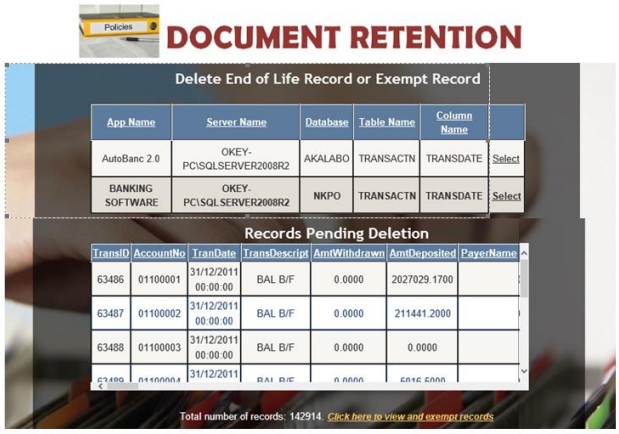


Fig. 12. Records Pending Deletion or Exemption

When the link "Click here to view or exempt records was clicked, a new page showing the Exempt records page was pop-up. Fig. 13 shows two records that were exempted out of 142914 records that have reached end-of-live. When retention was applied all the 142912 were purged from the database leaving on the two records that were exempted.

The tests show that the system was able to configure conventional documents and database records in different databases of varying DBMS platforms. The system was able to apply retention on these records at their end-of-live. Also, the users were able to exempt some records and gave reasons for such exemptions. New extension dates for expiration were as well indicated. When searches were performed on the referential, users were able to locate company policies and descriptions that pertains to various document categories the company created. This means that the system is running properly and will achieve its purpose and objective.

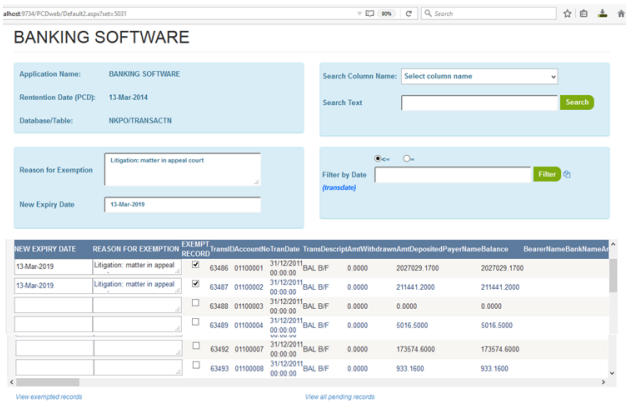


Fig. 13. Exempted Records in the Exemption Page

## B. *Discussion of Results*

Table 1 and Table 2 shows the summary of outcomes from the observation of test documents classified using the proposed document referential and retention system. The overall accuracy of eighty-eight (88) percent was determined by the percentage number of documents that were correctly placed in their exact document classes, against the total number of documents. The 20NewsGroup dataset version used excluded the cross-posts and included only "From" and "Subject" headers. It has 18828 documents (newsgroups posts) on twenty (20) topics [12]. The documents were split into two subsets, the training and test set.

Table 1 is the confusion matrix of the enhanced system which showcases the performance of the classification model on the test dataset. The table shows clearly the actual classes and the predicted classes, thereby helping in determining the number of documents that were correctly or incorrectly predicted by the model at a glance. Table 2 shows the summary of the outcomes in the confusion matrix and the performance evaluations of the model using the metrics: precision, recall, f1-score, specificity and accuracy. The true positive (TP) values shows that 1643 documents out of 1876 documents were correctly predicted. This gave an overall accuracy of eighty-eight (88) percent indicating that most predictions from the model are correct.

The false positive (FP) values shows that 239 documents were classified as positive when they are not. Whereas 217 documents were falsely classified as negative (FN column). The individual true negative (TN) values were gotten by calculating the sum of all columns and rows excluding that class's column and row. The TN values are almost equal to the total number of documents tested. This shows that the model correctly predicted the negative classes. The specificity of the individual classes ranges between ninety-eight (98) and one hundred (100) percent, whereas their precision, recall and f1-scores are above the threshold of 0.5 which shows that the model's predictions are reliable.

The recall values show the proportion of the actual positive classes the model was able to correctly identify as positive. The precision values which are close to one (1), shows the proportion of positive predictions that was actually correct. The results above shows that the proportion of the data points the model says are relevant actually were relevant and its predictions are reliable.


(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 01, January-2020

TABLE I. THE CONFUSION MATRIX OF THE CLASSIFICATION MODEL

| ACTUAL \ PREDICTED | 001 | 002 | 003 | 004 | 005 | 006 | 007 | 008 | 009 | 010 | 011 | 012 | 013 | 014 | 015 | 016 | 017 | 018 | 019 | 020 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 001 | 58 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 13 | 3 | 3 | 0 | 0 |
| 002 | 0 | 83 | 5 | 3 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |
| 003 | 0 | 6 | 79 | 6 | 1 | 2 | 0 | 0 | 0 | 2 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 004 | 0 | 2 | 5 | 80 | 4 | 1 | 3 | 0 | 0 | 1 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 005 | 0 | 0 | 2 | 5 | 81 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 3 | 1 | 1 | 0 | 1 | 1 | 0 | 0 |
| 006 | 0 | 3 | 2 | 1 | 0 | 86 | 0 | 0 | 0 | 0 | 1 | 3 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 007 | 0 | 0 | 2 | 5 | 2 | 0 | 79 | 0 | 1 | 0 | 2 | 1 | 3 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| 008 | 0 | 1 | 0 | 1 | 0 | 0 | 3 | 86 | 1 | 2 | 0 | 1 | 1 | 0 | 0 | 0 | 2 | 0 | 1 | 0 |
| 009 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 96 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 010 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 95 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 011 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 97 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 012 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 97 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 013 | 0 | 2 | 1 | 7 | 2 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 76 | 0 | 5 | 1 | 0 | 1 | 0 | 0 |
| 014 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 94 | 0 | 3 | 0 | 0 | 0 | 0 | 0 |
| 015 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 95 | 0 | 0 | 0 | 0 | 0 |
| 016 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 1 | 1 | 94 | 0 | 0 | 0 | 0 |
| 017 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 88 | 0 | 0 | 0 |
| 018 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 93 | 0 | 0 |
| 019 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 4 | 2 | 10 | 4 | 54 | 0 |
| 020 | 7 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 2 | 2 | 12 | 2 | 1 | 1 | 32 |

TABLE II. THE PERFORMANCE EVALUATION SUMMARY OF THE CLASSIFICATION MODEL

| Code | Document Type | No. of Docs | TP | FN | FP | TN | Precision | Recall | F1-Score | Specificity | FPR |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 001 | alt.atheism | 79 | 58 | 5 | 9 | 1788 | 0.8657 | 0.9206 | 0.8923 | 0.995 | 0.005 |
| 002 | comp.graphics | 97 | 83 | 14 | 16 | 1763 | 0.8384 | 0.8557 | 0.847 | 0.991 | 0.009 |
| 003 | comp.os.ms-windows.misc | 98 | 79 | 19 | 20 | 1758 | 0.798 | 0.8061 | 0.802 | 0.9888 | 0.0112 |
| 004 | comp.sys.ibm.pc.hardware | 98 | 80 | 18 | 28 | 1750 | 0.7407 | 0.8163 | 0.7767 | 0.9843 | 0.0157 |
| 005 | comp.sys.mac.hardware | 96 | 81 | 15 | 10 | 1770 | 0.8901 | 0.8438 | 0.8663 | 0.9944 | 0.0056 |
| 006 | comp.windows.x | 98 | 86 | 12 | 8 | 1770 | 0.9149 | 0.8776 | 0.8959 | 0.9955 | 0.0045 |
| 007 | misc.forsale | 97 | 79 | 18 | 10 | 1769 | 0.8876 | 0.8144 | 0.8494 | 0.9944 | 0.0056 |
| 008 | rec.autos | 99 | 86 | 13 | 6 | 1771 | 0.9348 | 0.8687 | 0.9005 | 0.9966 | 0.0034 |
| 009 | rec.motorcycles | 99 | 96 | 3 | 2 | 1775 | 0.9796 | 0.9697 | 0.9746 | 0.9989 | 0.0011 |
| 010 | rec.sport.baseball | 99 | 95 | 4 | 13 | 1770 | 0.8796 | 0.9596 | 0.9179 | 0.9927 | 0.0073 |
| 011 | rec.sport.hockey | 99 | 97 | 2 | 10 | 1767 | 0.9065 | 0.9798 | 0.9417 | 0.9944 | 0.0056 |
| 012 | sci.crypt | 99 | 97 | 2 | 9 | 1768 | 0.9151 | 0.9798 | 0.9463 | 0.9949 | 0.0051 |
| 013 | sci.electronics | 98 | 76 | 22 | 11 | 1767 | 0.8736 | 0.7755 | 0.8216 | 0.9938 | 0.0062 |
| 014 | sci.med | 99 | 94 | 5 | 6 | 1771 | 0.94 | 0.9495 | 0.9447 | 0.9966 | 0.0034 |
| 015 | sci.space | 98 | 95 | 3 | 14 | 1764 | 0.8716 | 0.9694 | 0.9179 | 0.9921 | 0.0079 |
| 016 | soc.religion.christian | 99 | 94 | 5 | 33 | 1744 | 0.7402 | 0.9495 | 0.8319 | 0.9814 | 0.0186 |
| 017 | talk.politics.guns | 91 | 88 | 3 | 20 | 1765 | 0.8148 | 0.967 | 0.8844 | 0.9888 | 0.0112 |
| 018 | talk.politics.mideast | 94 | 93 | 1 | 11 | 1771 | 0.8942 | 0.9894 | 0.9394 | 0.9938 | 0.0062 |
| 019 | talk.politics.misc | 77 | 54 | 23 | 3 | 1796 | 0.9474 | 0.7013 | 0.806 | 0.9983 | 0.0017 |
| 020 | talk.religion.misc | 62 | 32 | 30 | 0 | 1814 | 1 | 0.5161 | 0.6808 | 1 | 0 |
| | **Total** | **1876** | **1643** | **217** | **239** | | | | | | |
| | **Overall Accuracy** | **0.88** | | | | | | | | | |

The line graph in Fig. 14 was used in visualizing the observations. It shows at a glance the individual performances of each classes on the evaluation metrics - precision, recall, f1-score and specificity.

## V. CONCLUSION

This paper has discussed the document referential, classification and retention system and how organization can take advantage of it to protect their documents and as well, avoid ligations from concerned parties, fines from regulatory authorities, exposure to information theft or confidential documents reaching wrong hands. The implementation of this system will result to organized document referential that is easily accessible to members of the organization and will present a platform that allows for easy document classification and automatic monitoring/application of retention when document reaches end-of-live.

The study has shown the possibility of applying retention on conventional documents and records in multiple relational databases and two different RDBMS platforms based on the classification and retention schedule. Development of a single system that can create and manage document referential, provide easy way to place documents into their respective document classes and automatically apply retention on documents at its end-of-live.


Fig. 14. Graph of Documents Classes against individual Precision, Recall, Specificity and F1-Score

## VI. REFERENCES

[1] Arruda M., Prinzing M. and Rana S. (2003). "Documents, what Documents?" Business Law Today, US, p 23. In Howell R. and Cogar R. (2003). "Records retention – an essential part of corporate compliance". American Bar Business Law Newsletter, Vol. 19, p 1.

[2] Concklin J., Cook G. and Demond D. (2007). "Records retention manual". Proceedings of 80th annual conference of California Association of School Business Officials, California, Vol. 5, pp 9-10.

[3] Ataullah A. (2008). "A framework for records management in relational database systems". Thesis research, Department of computer science, University of Waterloo, Canada, p 3.

[4] Falai A., Arif Z., Gosaria C. and Prabowo S. (2017). "Indonesian news classification using naive bayes and two-phase feature selection model". IJEECS, Vol. 8, No. 3, pp 610-615.

[5] Kotsiantis S. (2007). "Supervised machine learning: a review of classification techniques". Informatica, Greece, Vol. 31, pp 250, 262.

[6] Lang K. (1995). "Newsweeder: learning to filter netnews". proceedings of the twelfth international conference on machine learning, pp 331-339. In Jason R. (2008). "The 20 Newsgroups data set". Retrieved from http://qwone.com/~jason/20Newsgroups/

[7] Naik C., Somaiya K., Kothari V. and Rana Z. (2015). "document classification using neural networks based on words". IJARCS, Vol. 6, No. 2, p183.

[8] Torii M., Yin L., Nguyen H., Mazumdar T., Liu F., Hartley M. and Nelson P. (2011). "An exploratory study of a text classification framework for internet-based surveillance of emerging epidemics". NCBI, US, Vol. 80, No. 1, pp 56–66.

[9] Triguero I., Sáeza J., Luengob J., Garciac S. and Herreraa F. (2014a). "On the characteri-zation of noise filters for self-training semi-supervised in nearest neighbor classification". Elsevier, US, Vol. 132, 30-41.

[10] Zakaria Z. (2015). "predicting performance of classification algorithms". International Journal of Computer Engineering and Technology, India, Vol. 6, No. 2, pp 19-28.

[11] Rakesh A., Paul B., Tyrone G., Logan S., and Walid R. (2005). "Extending relational database systems to automatically enforce privacy policies". IEEE-ICDE, New York, Vol. 64, p 1.

[12] Jasneet K. and Seema B. (2016). "News classification using naïve baye's classifier". International Journal of Advanced Research in Computer Science and Software Engineering Research, India, Vol 6(4), p 698


(This work is licensed under a Creative Commons Attribution 4.0 International License.)