How do you build effective data classification and handling of documents

A Preliminary Literature Review

Submitted by: Gihan Liyanage (1142109)

Submitted for: Professor Rosie

Submitted Date: 07/02/2023

Capstone Project RSCH6016

## How do you build effective data classification and handling of documents

**Introduction:**

Effective data classification and handling of documents are crucial components of modern information management systems. In today's data-driven world, organizations face significant challenges in organizing, categorizing, and safeguarding their vast amounts of data. Data classification involves the systematic categorization and labeling of data based on its sensitivity, importance, or other predefined criteria. Document handling, on the other hand, encompasses the processes and practices employed to manage and control the lifecycle of documents within an organization.

Building an effective data classification and document handling system requires careful planning, implementation of robust technologies, and adherence to best practices. This ensures that data is appropriately classified, protected, and accessible to authorized personnel. Effective data classification provides numerous benefits, including improved data governance, streamlined information retrieval, enhanced security measures, and compliance with regulatory requirements.

This literature review explores the existing research on building effective strategies and frameworks for data classification and document handling.

**Research Questions**

**Primary Question:**

The primary question of this literature review focuses on exploring the key factors and best practices for building effective data classification systems and document handling processes. Effective data classification and document handling are crucial for organizations to ensure data privacy, security, and regulatory compliance.

**Secondary Questions:**

* What are the challenges and potential risks associated with data classification and document handling?

Data classification and document handling come with their own set of challenges and potential risks that organizations need to address. Some of the key challenges and risks include Data Complexity, Lack of Standardization, Human Error, Regulatory Compliance, Data Security

* What are the current technologies and tools available for data classification and document handling?

Several technologies and tools are available to support data classification and document handling processes. These include Machine Learning and Artificial Intelligence, Data Loss Prevention (DLP) Solutions, Content Management Systems (CMS), Data Encryption Tools, Data Classification Software, Secure File-Sharing Platforms, Document Management Systems (DMS)

By leveraging these technologies and tools, organizations can enhance their data classification and document handling processes, streamline operations, and improve overall data security and compliance. Furthermore, it is important to conduct a thorough review of current literature and industry sources to identify and evaluate the specific technologies and tools available in the rapidly evolving field of data classification and document handling.

**Background**

Building effective data classification systems and document handling processes requires considering several key factors and implementing best practices. These include accuracy and granularity in categorizing data, consistent and well-defined policies aligned with compliance requirements, stakeholder involvement, automation and machine learning for efficient classification, robust access control and authentication mechanisms, encryption and data protection measures, secure sharing and collaboration methods, proper document retention and disposal procedures, user training and awareness programs, and adherence to industry standards and regulations. By incorporating these factors and best practices, organizations can enhance data security, privacy, and compliance, ensuring effective data classification and document handling throughout the information lifecycle.

**Literature Review**

**Primary Literature:**

The research of “Data classification - the foundation of information security” (Tankard, 2015) examines data classification is an essential component of effective data security and information governance. By assigning protective markings to documents, organizations can identify and safeguard sensitive data, including intellectual property, financial records, and customer information. This classification, coupled with clear policies and processes, strengthens other data protection controls like encryption and data loss prevention. Colin Tankard of Digital Pathways emphasizes that data classification plays a pivotal role in enhancing an organization's overall security posture and ensuring its protection. Recognizing the value and importance of data classification is vital for organizations to mitigate risks and safeguard their valuable assets.

The paper “Improving Document Classification Effectiveness Using Knowledge Exploited by Ontologies” presents a novel document classification model that leverages ontologies for document representation. The model incorporates ontology concepts acquired through exact matching and extraction of related terms. It also introduces a concept weighting scheme based on relevance and importance, resulting in improved classification performance according to experimental results. (Zenun Kastrati, 2017)

In the article of “Building solid foundations: the case for data classification”, Cath Everett highlights the significance of data classification as the foundation of information security. While the discipline is crucial, many organizations, except those in defense and security services, have yet to fully embrace it. Data classification should ideally drive rules for user permissions regarding corporate information. Information security professionals increasingly recognize data classification as the core of all information security activities. The article explores the importance and benefits of data classification, delves into the reasons behind its limited adoption, and proposes constructing a business case for its implementation. Additionally, the article offers insights into the necessary steps organizations must take to ensure successful implementation of data classification projects amidst their complexity. (Everett, 2011)

**Secondary Literature:**

Paper “Challenges Related to Identifying Sources and Document Collection for Big Data Analyses” (Gmiterek, 2019) explores the challenges and possibilities of big data in the context of information science, libraries, and digital document accessibility. It addresses issues related to searching, organizing, and classifying large volumes of data and examines the tools offered by libraries, particularly discovery systems. The review utilizes a critical analysis of literature, along with experiments and observations, to present insights on these topics.

“A Policy-Driven Framework for Document Classification and Enterprise Security” (E. Nwafor, 2016) highlights the challenges of accessing sensitive resources on ubiquitous devices in enterprise environments. It emphasizes the need for dynamic access control to combat cyber-attacks. The proposed framework aims to automate document classification based on sensitivity levels, reducing manual classification errors. The framework incorporates text mining, user context, and enterprise policies to detect security profiles and access patterns of enterprise resources. Additionally, a prototype system is developed to enhance the effectiveness of the framework. This research contributes to the advancement of automated security measures and access control in enterprises, addressing the evolving landscape of ubiquitous device usage and cybersecurity threats.

The paper of “A Review of Machine Learning Algorithms for Text-Documents Classification” (B. Baharudin, 2010) focuses on the crucial task of automatic document categorization in the context of the abundant electronic documents available today. It emphasizes the importance of employing text mining, machine learning, and natural language processing techniques to effectively classify various types of documents such as e-documents, online news, blogs, e-mails, and digital libraries. The objective is to provide a comprehensive review of the existing literature, highlighting the techniques and methodologies used in text document classification. Additionally, the paper raises awareness about the remaining challenges in the field, particularly regarding text representation and machine learning techniques. Overall, this review contributes to the understanding and advancement of document classification and text mining methods.

The study of “Text Classification for Records Management” conducted by (Franks, 2022) addresses the pressing need for automatic classification of electronic records, driven by the increasing data volumes and digital rights legislation. The traditional approach of using expert systems based on metadata is no longer viable due to the growing diversity of records and the lack of metadata. The study compares the performance of traditional text classification techniques with newer natural language processing technologies using real records data. The results reveal that while advanced Transformer language models demonstrate superior classification abilities, traditional methods still maintain strong performance. The findings were further discussed by a focus group of record managers, who recognized the potential of text classification in managing risk and ensuring compliance. This research represents an important initial step towards the goal of synthesizing narrative from a corpus of records.

“Managing Content with Automatic Document Classification” (R. Calvo, 2004) paper explores the use of machine learning and automatic document classification techniques for managing large volumes of news articles and web page descriptions. By leveraging these techniques, the burden on domain experts is reduced, allowing for more efficient categorization of content. The study utilizes two datasets containing a significant number of Reuters news stories and web sites, employing a Naive Bayes algorithm to classify them into predefined categories. The paper discusses various parameters and design decisions involved in building automatic classifiers, including stemming, stop-words, thresholding, data size, and strategies for enhancing performance using XML document structure. The proposed methodology has the potential to enhance information management in web-based applications and workflow systems by automating document categorization or assisting human experts in the process.

**Coding Process**

The coding process for this literature review involved conducting a systematic search using various academic databases and relevant keywords related to data classification, document handling, and information security. The selected papers were then analyzed and summarized based on their contributions to the research questions.

**Project Timeline**

The following timeline outlines the major milestones and activities for completing the research:

Week 1: Conduct literature search and gather relevant papers.

Week 4: Read and analyze selected papers.

Week 6: Summarize key findings and create template.

Week 7: Create survey for research to gather information.

Week 8: Write the literature review document.

Week 10: Review and finalize the document.

# References

B. Baharudin, L. H. (2010). *A Review of Machine Learning Algorithms for Text-Documents Classification.*

E. Nwafor, P. C. (2016). *A Policy-Driven Framework for Document Classification and Enterprise Security.* IEEE Conferences on Ubiquitous Intelligence & Computing.

Everett, C. (2011). *Building solid foundations: the case for data classification.*

Franks, J. (2022). *Text Classification for Records Management.* Journal on Computing and Cultural Heritage.

Gmiterek, G. (2019). *Challenges Related to Identifying Sources and Document Collection for Big Data Analyses.* e-mentor.

R. Calvo, J.-M. L. (2004). *Managing Content with Automatic Document Classification.* Journal of Digital Information.

Tankard, C. (2015). *Data classification - the foundation of information security.* Netw. Secur.

Zenun Kastrati, S. Y. (2017). *Improving Document Classification Effectiveness Using Knowledge Exploited by Ontologies.* NLDB.