Journal Homepage: - www.journalijar.com
INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)
Article DOI: 10.21474/IJAR01/9886
DOI URL: http://dx.doi.org/10.21474/IJAR01/9886

RESEARCH ARTICLE

CDR (CONFIDENTIAL DATA RECOGNITION) BOT.ANALYSIS AND CLASSIFICATION OF DATA BASED ON THE LEVEL OF CONFIDENTIALITY USING MACHINE INTELLIGENCE.

Akshay Kumar C.R and Chaitra.K.
Software Engineer at CGI, B.E (ECE), Bangalore, India.

Manuscript Info

Manuscript History
Received: 12 August 2019
Final Accepted: 14 September 2019
Published: October 2019

Key words:-
Machine Learning (ML), Data Classification bot, Text classifier, SAAS, User Interfaces (UI), Sharepoint.

Abstract

Now a days, Analysis, Classification and Maintenance of huge amount of data based on its complexity and confidentiality level is a major challenge faced by all the organizations. Now we have developed a smart intelligent system with special importance on user friendly Interface, software principles with powerful data analytics algorithms for business which can effectively interpret human ideologies on classification of data. Hence our application allows user to classify huge amount of data and further apply various polices to categorize the data based on its level of confidentiality and suggests the users to apply certain retention policies to retain confidential data with high Security for longer duration.

The CDR bot presented in the paper smartly analyses the confidentiality level of the data in a document and provides graphical interpretation on the level of confidentiality of the data in document on User interface. Hence user can apply certain retention polices as per business requirements of individual organization to retain the document.

This CDR bot is a SAAS model and has been optimized using Machine Learning Algorithm coupled with important Software Engineering Concepts.

This paper is focused on analysis, classification of data based on its confidentiality level which can be adapted by all organizations as a technical solution to classify huge amount of data via CDR Bot.

Introduction:-

Machine Intelligence is an integrative field of Artificial Intelligence focusing on effective interaction between Human and Machine.

Data protection is the major challenge in all public and private organizations. Many of the organization will use various Data Leakage Prevention System (DLP) for protection of confidential data, policies will be applied on the data to prevent data leakage etc. But ancient DLP system used demands for huge maintenance cost as it requires individual infrastructure, security policies and resources to maintain the system. But Data Leakage Prevention System (DLP) still has the threat of loosing the data or data theft which is a huge risk for any organization.

Corresponding Authors:-Akshay Kumar C.R.
Address**:-**Software Engineer at CGI ,Bangalore,India.

To provide an efficient, easy recognition and classification of confidential data as per human ideologies on the level of confidentiality of data we are proposing a new tool called CDR (Confidential Data Recognition) Bot which adapts the principles of Machine Learning algorithm.

Working Principle of CDR BOT.

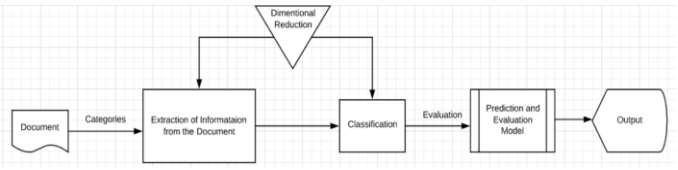

Fig.1: -Principle of CDR BOT.

Literature Survey
This section deals with the text of a scholarly paper which includes current knowledge including theoretical and methodological contribution to this project.

The paper mentioned below are developed by few scholars has provided us the ideologies to develop and improvise the project.
[1] Sheena Angra , Sachin Ahuja "Machine learning and its applications" : -Paper aims at analyzing and extracting some of the useful information and develop algorithms based on the analysis. This is achieved through Data Mining and Machine Learning.
[2] Yang Xin ; Lingshuang Kong ; Zhi Liu ; Yuling Chen ; Yanmiao Li ; Hongliang Zhu "Machine Learning and Deep Learning Methods for Cybersecurity" :- Data security is a major challenge apart from accumulation and classification of data.This paper majorly deals with application of Machine learning (ML) and deep learning (DL) methods for network analysis of intrusion detection and provides a brief tutorial description of each ML/DL method.
[3] Human and Machine Interaction (HMI) via User Interface plays a vital role in software development process.
[4] Existing confidential data identification methods can be divided into two categories: Content based, and Behavior based methods. The content-based method includes rule-based and classifier-based approaches.

In content-based approach, directly focuses on data values. Data values contain the use of confidential words, regular expression, text classification and information retrieval

In classifier-based approach various classifications and machine learning techniques are used [12, 13], such as SVM (Support Vector Machine) and Naïve Bayes [14, 15, 16]. In this approach the content of the document is represented as vectors [15]. Vectors are generated using terms and their frequencies of the documents. These vectors classify, whether the documents are confidential or not.

Key Challenges
As discussed earlier confidential data recognition and maintenance are mainly advantageous and huge boon for industries. But also have some key challenges for usage of these.

1. **Infrastructure**: -Highly secure and restricted environment is required in websites to accommodate large amount of data and prevent data theft.
2. **Cost**: - Cost of Implementation of these technology in their official website.
3. **User Interface**: - Smooth and Straightforward user interface is a main challenge of software engineering. As per our analysis 52% of the users will be unhappy only because of User Interface.
4. **Self-Robust Intelligence: -** Ability to read and recognition of data in document. Prediction of the confidentiality level of the document based on the content.
5. **Security: -** User related information and data of the document need to be highly protected. Especially customer trust and satisfaction are important. On demand from the customer end retention policies has to be applied to maintain the data for a longer duration. Exposure of documents to only authorized group of users based on policies applied on the document.
6. **Robustness: -** According to the study we made approximately 50% of web users expects site to be loaded in 2 seconds or less and they will try to avoid the site if it is not loaded in 3 seconds.

Proposed Methodology:-

User had a query regarding the data classification and recognition of confidentiality of data in the document. Even though the client is using the Sharepoint for content management and websites as well.

They were facing a huge challenge/problem in maintenance of the document and retention of the document based upon the confidentiality. And they needed a clarification and solution to overcome these problems. Thus we came up with the CDR Bot application which answers all the queries and hence solve user's problem.

As per the user requirements and Business Standards we have deployed our application CDR Bot on Sharepoint Platform for content management. CDR Bot is a SAAS model hence can be deployed to any of the other websites or portals etc. to classify confidential data and retain it based on its confidentiality level.

Robustness:
Speed of the web application is high. For user convenience various API endpoints has been created to increase the performance. Normal file operation as been used as per requirement to optimize the performance of the tool.

Quality and Design:
Professional web design techniques and color standards has been followed for better User Interface in the front end.

Text classification and identification of the confidential data has been implemented using Machine Learning algorithms. Cross platform integration and unit testing will be done accordingly.

Security:
The security is the critical issues in the real world. Because failure and leakages of the data will lead to financial loses. This is one of the reasons to avoid database. Malicious activities and file will be attended quickly.

Implementation
User Interface: -
User Interface is a key part of all web application which will attract the user to use the application and to navigate through out the application system.
We are using HTML, CSS, JavaScript and AJAX for the front-end design as well as client-side component of the application for functional implementations.

Functionalities: -
Since this application and tool has been proposed for the users where they are using Microsoft Sharepoint for their website portals as well as content management etc. So, functionalities will be explained in context of Sharepoint but there is no dependency of this tool in Sharepoint as stated above this can be hosted any of the website and platform and can be improvised further for that respective portals.

Since CDR Bot was used in Sharepoint for demonstration auto population of the documents in that site collection feature has been added to CDR Bot and on selection of the document analytics of CDR Bot was working. And we have an option of uploading the document and for the document which was uploaded by the user will be scanned accordingly and the result will be provided.

In the same way CDR Bot can be hosted for all the website and portals like Sharepoint, Box for Business, Google Drive, Confluence etc.

Real Time Application Dashboard

Fig 2:-Text Classification Dashboard

On Selection of the document or once the document is uploaded through the ajax call Web API will trigger the python executable file. Entire document will be scanned using python script and text classification machine learning algorithms are used for Confidential data recognition and text classification. One image will be created based upon the scanned data in document which will help us to recognize the type of data in that document please refer picture 2 for the image.

Picture 2:-CDR Cloud Image

In addition to the above functionalities as per the user requirement, based upon the level of confidentiality CDR Bot will suggest the user regarding the label or Retention polies or legal policies which can be applied for that document and this requirement is on demand of user requirement since they were using Sharepoint.

And this can be used for multiple other functionalities as well and this is not restricted only for Sharepoint CDR Bot can be customized based upon the user requirements.

Validation And Demonstration
A. Control Flow
The Control flow fig2 explains how the data transfers

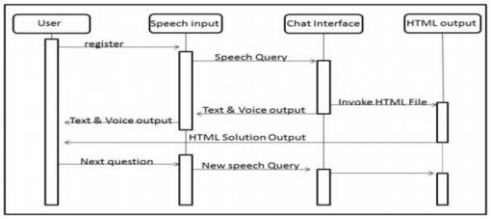

Fig 3:-The Control flow diagram of the system.

internally in the system. user invokes the system. The user should register first (only the user name is stored) Then the admin team as to validate the user has per business approval has to give access for the user who had raised the request.

Once the access has been given to the tool user can start using it and documents lists inside that site will be populated in the dashboard and on click or selection of the document text classification will be performed using machine learning algorithms.

Meanwhile, we will ask the user for feedback for the satisfaction of output received. If the user is happy about solution, we can conclude that the machine has classified and retrieved accurate information as per user needs by training on the limited database. The user query can be added to training set of databases which could further bolster the performance of the Machine learning algorithm.

If the user is unhappy of the response from the system, then we will provide the opportunity to notify the question to the system admin and admin can act accordingly,

Conclusion:-

The basic requirement of all the organizations is accumulation, classification and highly secured maintainability of data. The actual power of an organization lies in the ability to maintain the confidentiality of Data. This software is built to attain that specific purpose for all the organizations irrespective of the fields they work on. This application can be hosted on server reducing the cost of application as this is used as [Software As A Service] SAAS, Uptime of system is 99%,

Robustness, Security are all inbuilt traits of modern-day server.

The technology coupled with the Software Engineering principles and Best Architectural principles can bring down the development cost of these system by huge margin. This application proved to one of the efficient tool with less requirement of human resources , service costs and data loss risk. We have implemented and hosted this tool on Sharepoint Platform as per our Business requirement. Tool is Platform independent and can be deployed on any environment.

Future enhancements are deploying this as a Mobile phone application model since it expands the target reach for large user base. The reinforcement learning, or deep learning algorithms can be used to increase accurate classification of system. The Application after perfection can be enhanced to identify and secure most confidential data on all the platforms.

Citations:-

1. Machine Learning and Its Applications Publisher: IEEEPublished in: 2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC) [https://ieeexplore.ieee.org/document/8070809/references#references]
2. Machine LearningPublished in: IEEE Software[https://ieeexplore.ieee.org/document/7548905/authors#authors]
3. A Comprehensive Study of Text Classification Algorithms
4. Published in: 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI) [https://ieeexplore.ieee.org/document/8125990/authors#authors]
5. Classification of Algorithms in Machine Learning
6. Published in: IEEE Access [https://medium.com/datadriveninvestor/classification-algorithms-in-machine-learning-85c0ab65ff4]
7. Introduction of Classification Algorithms [https://www.edureka.co/blog/classification-algorithms/]
8. 7 Types of Classification Algorithms [https://www.analyticsindiamag.com/7-types-classification-algorithms/]
9. Machine Learning and Deep Learning Methods for Cybersecurity [https://ieeexplore.ieee.org/document/8359287/authors#authors]
10. A taxonomy for combining software engineering andhuman-computer interaction measurement approaches:towards a common framework by JennyPreece and H.DieterRombach
11. The 8 Core Principles of Good Customer Service by Pascal [https://www.userlike.com/en/blog/customer-service-principles]
12. Human-Computer Interaction (HCI) by interaction design foundation [https://www.interaction-design.org/literature/topics/human-computer-interaction]
13. Samuel, Arthur (1959). "Some Studies in Machine Learning Using the Game of Checkers". IBM Journal of Research and Development. 3 (3): 210–229. doi:10.1147/rd.33.0210.

14. W.W Cohen & Y.Singer, (1999)"Context-sensitive learning methods for text categorization", ACM transactions on Information sytermmms, pp141-173.
15. H.Drucker & D.Wu, (1999)"Support vector machines for spam categorization", IEEE transaction on neural networks.
16. I.Androutsopoulos & J.Koutsias,(2000),"An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with person e-mail messages", In proceedings of the 23rd annual international ACM SIGR conference on Research and Development on Information Retrieval,ACM.Athens,Greece,pp160-167.
17. M.Sahami & S.Dumais, (1998)"A Bayesian approach to filtering junk email", AAAI-98 workshop on Learning for text categorization.
18. J.Hovold, (2005),"Naive Bayes spam filtering using word-position-based attributes", in proceedings of the 2nd conference on Email and Anti-spam.