IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, June 2022
Impact Factor: 6.252

AI Based Self Learning Intelligent Information Leak Protection System for TI Companies using LSTM

K. Pazhanivel[1], E. Mounika[2], R. Shilpa[3], S. Sakthi[4]
Assistant Professor, Computer Science and Engineering[1]
Student, Computer Science and Engineering[2,3,4]
Anjalai Ammal Mahalingam Engineering College, Thiruvarur, India

Abstract: *Preventing the leak of sensitive information, also popularly known as data leak or data loss to an unauthorized recipient, is the primary goal of an organization's information security system. A data leak can occur through multiple channels. While it may not always be possible to prevent it entirely, measures can be taken to minimize the possibility of the occurrence. Like all other financial institutions, TI companies collect sensitive personal information of their customers for business purposes. This information is often categorized into three primary types; NPI, PII, and PI are the designated types in the descending order of sensitivity. The detection of sensitive documents and redaction of sensitive information is required if it is needed to be shared. Inspection of such digital documents to find any sensitive information is by far a human-driven process, and thus time-consuming and costly. An and costly. An intelligent and robust system is required where the content is analysed by state-of-the-art data mining, statistical and machine learning techniques from various data dimensions. An AI based self-learning Intelligent Information Leak Protection System using LSTM is proposed in the project that mines and extracts information and categorizes the document images, to SD or NSD, based on the presence of NPI and PII semantic signatures without any explicit rule configuration. The system is designed to be used proactively as an early warning system to tag the SD images while resting in the data store. It can also act as a real-time checkpoint for the information loss by the documents in transit or use. The proposed model prescribes an information loss protection mechanism using a binary classifier based on the state-of-the-art LSTM technique within the paradigm of Artificial Intelligence.*

Keywords: Critical Infrastructure Security, Network Security, Application Security, Cloud Security, Information Security, Disaster Recovery / Business Continuity Planning Storage Security, End-user Education, etc.

I. INTRODUCTION

Preventing the leak today's touchy information, additionally popularly referred to as facts leak or data loss to an unauthorized recipient, is the number one purpose present day a company's statistics safety device. A statistics leak can arise via multiple channels. at the same time as it is able to now not always be viable to save you it completely, measures can be taken to limit the possibility modern-day the prevalence. like every different economic establishment, TI organizations gather touchy private records of their customers for business functions. This information is cutting-edge categorized into three primary sorts; NPI, PII, and PI are the designated sorts within the descending order brand new sensitivity.

The detection trendy sensitive documents and redaction latest touchy records is required if it's far needed to be shared. Inspection latest such virtual documents to find any sensitive facts is by some distance a human-driven manner, and as a consequence time-eating and pricey. A wise and sturdy system is needed where the content material is analysed through 49a2d564f1275e1c4e633abc331547db facts mining, statistical and device latest techniques from various facts dimensions.

An AI primarily based self-contemporary clever records Leak protection device the usage of LSTM is proposed in the undertaking that mines and extracts statistics and categorizes the record images, to SD or NSD, based at the presence modern day NPI and PII semantic signatures without any express rule configuration. The machine is designed for use proactively as an early warning system to tag the SD photos while resting inside the records save. it may also act as a real-time checkpoint for the facts loss by means of the documents in transit or use.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, June 2022
Impact Factor: 6.252

The proposed version prescribes an statistics loss protection mechanism using a binary classifier primarily based on the 49a2d564f1275e1c4e633abc331547db LSTM approach within the paradigm ultra-modern artificial Intelligence a self-contemporary clever statistics Leak safety system is proposed in the present observe that mines and extracts information and categorizes the document photos to SD or NSD, based on the presence ultra-modern NPI and PII semantic signatures with none explicit rule configuration. The system is designed to be used proactively as an early warning system to tag the SD pics while resting within the facts save.it is able to also act as a real-time checkpoint for the statistics loss by using the files in transit or use, as shown in determine. The classifier is trained and examined with various TI-related manufacturing record pictures tagged as SD and NSD below human supervision. 49a2d564f1275e1c4e633abc331547db class algorithms are compared with the ANN-primarily based approach contemporary the MLP version. The latter has executed with significantly higher accuracy in trendy the sensitive content material from the samples and classifying the documents with accurate, touchy statistics.

II. RELATED WORK

DLPs are specially designed expert systems that can detect monitor, and take preventive actions against any possible information leak based on predefined policy rules [9]. These systems are also popularly identified with names such as Information loss/Leak Prevention system, Extrusion Prevention System, Content Monitoring System, Filtering System etc Many commercial and in-house systems have been designed and deployed in the recent past to address the threat of information leak. The variations of the systems are attributed to the variations of the type of data leak incidents, capabilities required, data types, data monitoring channels, data states, etc.

The evolution of DLPs took place in three stages. In the early years, most of the solutions were developed to monitor and protect sensitive data going out of the network boundary of an organization by deploying software programs at different network egress junctions. Protecting data leakage through removable storage devices such as hard discs, USB devices, etc. were the point of focus in the second stage. Along with the paradigm shift in data storage and analytic structure in the last few years, the DLPs are experiencing the third phase of the evolution. In this phase, unstructured data of various forms are addresses with state-of-the-art statistical, ML, and natural language processing capabilities.

III. METHODOLOGY

A deductive and quantitative technique has been taken inside the present experimental observe to investigate and set up the pro-posed system's (See the workflow of IILPs in parent three under Capability to resolve an actual-time trouble of a reputed. TI enterprise. The agency has approximately one billion scanned document pix in one of its virtual record repositories as of the date of the study. A sizable seasoned-component of those files deliver NPI and PII and have to be secured while at relaxation and while the documents are in transit. The test's purpose became to set up a machine which can act as an early-warning system by tagging the secured documents while at relaxation and identifying a document as touchy earlier than getting used or dispatched outside the agency premise.

The general experimental setup changed into executed in six steps; The Steps are facts series, Sampling, analysis, function representation, floor fact technology, version constructing. The experiments had been carried out in two simultaneous levels, with the sample records randomly break up into eighty percentage as a teach set (forty-six,896 files) and twenty percent as a check set (11,723 files). each phase became in addition subdivided into four experiments with four distinct feature representations of the files TF-IDF characteristic space with unigram, bigram, trigram, and composite n-gram (n = 1,2 and 3) tokens. inside the first phase, the baseline classifiers had been educated and confirmed the usage of four different types of features independently.LR, RF, NB, KNN, and SVM are the 5 popularly used in python.

A. IILPS Web UI

In this module the physical documents are scanned in batches and stored in a digital archive through this web interface for TI companies as a heterogeneous document stream, referred to as a digital package. This web interface has user account and access management design, File access management. To access the data, a library of IILPS analysis tasks is available through this web interface

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, June 2022
Impact Factor: 6.252

B. Dataset Preparation and Exploration

The first step of a data science task is to obtain, gather, and measure the necessary and targeted data from available internal or external data sources, and then compiled into an established system. Sample documents were manually verified by the subject matter experts from the TI domain and tagged as either `Secured' or `Not Secured' based on the presence of NPI and PII attributes. Create manually annotated tagged files using spread sheets and save it as file type csv.

C. Features Extraction

Choosing those features which are somehow correlated to NPI and PII (correlation > 0.1) correlation heatmap is used to list all the correlation coefficients in order to identify multicollinearity.

D. LSTM Classification

The LSTM classifier is trained with various TI-associated production document images tagged as SD and NSD without human supervision. After the classification two folders are created and named automatically as SD and NoNSD and split and locate the files in the specified folder based on the classified result.

E. Decision Making

Automated thresholding of the prediction was one of the vital features of the proposed system. Optimizing the Type 1 and Type 2 error of the binary classifier as needed to receive the system's best prediction capability.

F. Proactive Detection

The system is designed to be used proactively as an early warning system to tag the SD images while resting in the data store. It can also act as a real-time checkpoint for the information loss by the documents in transit or use. System. that can act as an early-warning system by tagging the secured documents while at rest and identifying a document as sensitive before being used or sent outside the organization premise.

E. End User

IILPS Authorized Regulator, TI Companies employees, Server Admin

F. Experimental Analysis

Performance metrics such as overall accuracy, sensitivity, specificity, and F1 score were collected and measured from both phases of experiments.

Algorithm

LSTM: LSTM is a class of recurrent neural network. The three significant parts of the LSTM model include

Forget gate: Removes information that is not, at this point, essential for the finishing of the errand. This progression is fundamental to streamlining the presentation of the network.

Output Gate: Selects and yields vital information.

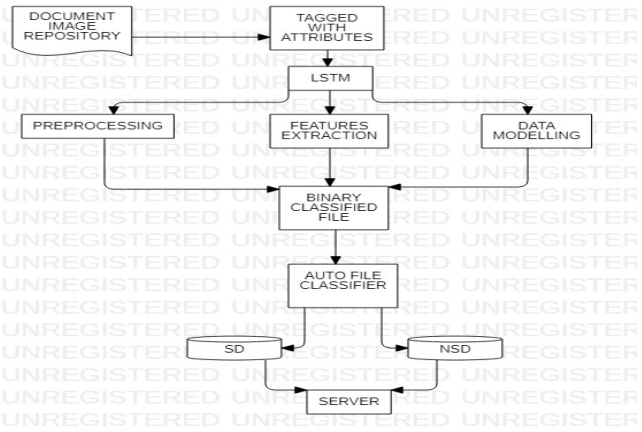
Input gate: Responsible for adding information to the cells. incorporate forecast dependent on past information.

Expected Outcome

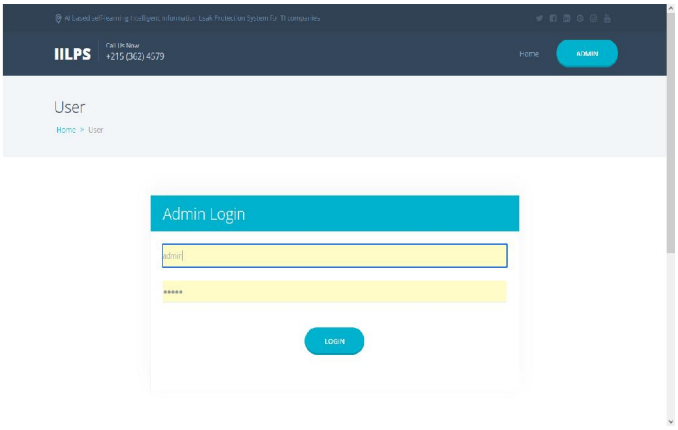
- Prevent the issue of information leaks from a digital document while it is at rest, use, or in transit.
- Minimize the overall cost of the model.

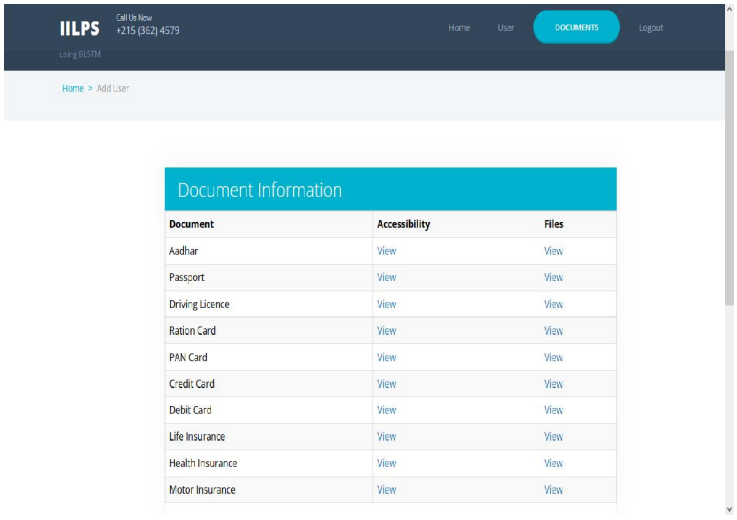
Below diagram shows System Architecture of proposed system

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, June 2022
Impact Factor: 6.252



IV. IMPLEMENATION AND RESULTS





V. CONCLUSION

Even though a diffusion cutting-edge industrial and in-residence DLP solutions were proposed and followed within the beyond by way of a couple of agencies the solutions lacked the cognitive ability and, most contemporary the time, no

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, June 2022
Impact Factor: 6.252

longer scalable in terms brand new the unseen and unknown styles cutting-edge touchy information. Due to many feasible varieties in touchy records and its usually evolving nature, the systems work nicely with predefined styles and fail to locate and act on any new styles intelligently. The beyond research indicated and emphasized the want for statistical, system present day, and facts mining techniques within the discipline modern DLP. on this paper, we proposed a gadget that mines, identifies, and learns the patterns from the target information and can improve itself from the non-stop remarks from the output. The proposed model prescribes an data loss seasoned Detection mechanism.

REFERENCES

[1] H. Alhindi, ``A framework for data loss prevention using document semantic signature,'' Ph.D. dissertation, Dept. Elect. Comput. Eng., Univ. Victoria, Victoria, BC, Canada, 2019.
[2] A. Guha and D. Samanta, ``Hybrid approach to document anomaly detection: An application to facilitate RPA in title insurance,'' Int. J. Autom. Comput., vol. 18, no. 1, pp. 55-72, Feb. 2021.
[3] Y. Lu, X. Huang, Y. Ma, and M. Ma, ``A weighted context graph model for fast data leak detection,'' in Proc. IEEE Int. Conf. Commun. (ICC), May 2018, pp. 1-6.
[4] K. Rishika and V. Damodaran, ``Data leakage detection: Challenges and prevention systems,'' Springer, Singapore, Tech. Rep., 2020.
[5] J. Zheng, ``Pattern matching for data leak prevention,'' US Patent 10 354 088, Jul. 16, 2019.R