Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 5, No. 3, May : 2023

INSURANCE FRAUD PREVENTION

Ms.S.A.Suje, Assistant Professor, Dept. of AI&DS, SNS College of Engineering.
Ms.Sneha.R, Student, Dept. of Information Technology, SNS College of Engineering.
Ms.Rani.J, Student, Dept. of Information Technology, SNS College of Engineering.
Mr.Naveen.V, Student, Dept. of Information Technology, SNS College of Engineering.
Mr.Chandru.V, Student, Dept. of Information Technology, SNS College of Engineering.

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 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 BLSTM 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 BLSTM technique within the paradigm of Artificial Intelligence.

Keywords: Data Security, Fraud Prevention, Insurance Fraud, Financial Fraud, Privacy.

I. INTRODUCTION

Protecting Sensitive Information (SI) is a far greater operational challenge for organizations. SI refers to information that doesn't identify an individual but is related to an individual and communicates information about that person that is private or could potentially harm the individual should it be made public. SI includes NPI, PII and PI Data. The challenge with traditional data security tools like Data Loss Prevention in protecting SI is that many of those data elements exist in common usage without being related to an individual. It's also very difficult to program a content analytics engine to find information that is in scope with the General Data Protection Regulation (GDPR) without finding large volumes of information that aren't in scope at the same time.Breaches of sensitive data can happen in various ways. Those that garner the most attention are large-scale breaches, which are often caused by incorrect technical configurations or a lack of due care on an industrial scale. But far more frequently, information is compromised on a small scale due to a user being careless or lacking awareness about the sensitivity of data they're handling. In these cases, data classification can help reduce the risk of a breach significantly. Data classification allows a user to tag data by selecting a classification from a list. Many people are familiar with classification schemas used by governments and military organizations, which classify information by levels of secrecy. For example, classifications may include public, sensitive, secret and top secret. The most effective data classification tools are very flexible, allowing for multiple levels of classification and offering customizable fields.

II. EXISTING SYSTEM

- **User-Based Document Management Mechanism in Cloud**
  The document plays an important role in the development of cloud computing. The user obtains and shares information by the electronic document. It is rich in content and various in representations. But the challenge to security is also brought. For the secure requirement for document in cloud, firstly, propose a novel user-based document secure management mechanism which introduces the re-encryption. The re-encrypted key will be generated according the access control conditions, so that the encryption of document creation will be combined with access control.
- **Automatic Authenticity Verification of Printed Security Documents**
  A particular class of security documents has been considered for the present experiment. Bank cheques, several kinds of tickets like lottery tickets, air tickets, etc., legal deeds, certificates, mark sheets, postal stamps, etc. all these documents fall under the same class as far as security is concerned. Criminalists efforts for generating fraudulent version of such documents are on the rise. This study attempts to develop a general framework for automatic authenticity verification of such security documents. The proposed method first computationally extracts the security features from the document images and then the notion of authenticity vs. duplicity is defined in the feature space. Bank cheques are taken as a reference for conducting experiment. Support vector machines (SVMs) are used to verify authenticity of these cheques.
- **Machine Authentication of Security Documents**
  This method first computationally extracts the security features from the document images and then the notion of genuine vs. duplicate is defined in the feature space. Bank cheques are taken as a reference for conducting the present experiment. Support Vector Machines (SVMs) and neural networks (NN) are involved to verify authenticity of these cheques.
- **Document Encryption Through Asymmetric RSA Cryptography**
  Cryptography with asymmetric keys is the strongest data security technique to use. One of the most widely used asymmetric cryptography is the RSA (Rivest-Shamir-Adleman) algorithm. The type of document that is encrypted is the most commonly attached document when sent e-mails. The document types are
  .docx, .pptx, .xlsx, .pdf, .jpg and .mp4. In the encryption process, a public key and a private key will be generated which can be sent separately by sending encrypted digital documents. The decryption process for digital documents is carried out from the receiving end of the document using a private key generatedin the encryption process.
- **Printed Document Authentication Using Watermarking Technique**
  Forgeries related with official printed documents can easily be performed with the aid of today's advanced electronic devices such as scanners and computers. The forged documents are usually undetectable by human eyes and it is with this regard that there is an urgent need to find solutions to the threat of counterfeiting of such documents. With watermarking technique, information that is used to determine the validity of printed document can be embedded in the document. The embedded information can be imperceptible to human eyes; thus, the process of forgery is made harder for the attackers. To authenticate the owner of the document, the embedded watermark is extracted from the watermarked document. However, in the verification process the printed document may suffer from printing and scanning (PS) distortion and as such it is necessary to resolve the noise and unwanted rotation as well as any degradation made by printing and scanning. This system uses a watermarking technique to address this issue.

Disadvantages

- The encryption result has a size larger than the original file size because it has been encoded in another form according to the RSA algorithm. The longer and bigger the input size, the longer it will take required for encryption.

- Not intelligent enough to identify a new type of leak or an existing leak which is not configured in the policy.
- Not detective in nature.

III. PROPOSED SYSTEM

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. Binary classifier model to identify the composite n-gram (n D 1 to 3) features of the documents and produce a decision boundary. LSTM mines and extracts tagged information and categorizes the document images, to SD orNSD.

- **LSTM – Long Short-Term Memory**

LSTMS are a special kind of RNN which is capable of learning long-term dependencies. LSTMs are designed to dodge long-term dependency problem as they are capable of remembering information for longer periods of time. Long short-term memory (LSTM) units (or blocks) are a building unit for layers of a recurrent neural network (RNN). A RNN composed of LSTM units is often called an LSTM network. A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell is responsible for "remembering" values over arbitrary time intervals; hence the word "memory" in LSTM. Each of the three gates can be thought of as a "conventional" artificial neuron, as in a multi-layer (or feedforward) neural network: that is, they compute an activation (using an activation function) of a weighted sum. Intuitively, they can be thought as regulators of the flow of values that goes through the connections of the LSTM; hence the denotation "gate". There are connections between these gates and the cell. The expression long short-term refers to the fact that LSTM is a model for the short-term memory which can last for a long period of time. An LSTM is well-suited to classify, process and predict time series given time lags of unknown size and duration between important events. LSTMs were developed to deal with the exploding and vanishing gradient problem when training traditional RNNs.

The popularity of LSTM is due to the Getting mechanism involved with each LSTM cell. In a normal

RNN cell, the input at the time stamp and hidden state from the previous time step is passed through theactivation layer to obtain a new state. Whereas in LSTM the process is slightly complex, as you can see in the above architecture at each time it takes input from three different states like the current input state,the short-term memory from the previous cell and lastly the long-term memory.

These cells use the gates to regulate the information to be kept or discarded at loop operation before passing on the long term and short-term information to the next cell. We can imagine these gates as Filters that remove unwanted selected and irrelevant information. There are a total of three gates that LSTM uses as Input Gate, Forget Gate, and Output Gate.

Input Gate

The input gate decides what information will be stored in long term memory. It only works with the information from the current input and short-term memory from the previous step. At this gate, it filters out the information from variables that are not useful.

Forget Gate

The forget decides which information from long term memory be kept or discarded and this is done by multiplying the incoming long-term memory by a forget vector generated by the current input and incoming short memory.

Output Gate

The output gate will take the current input, the previous short-term memory and newly computed long- term memory to produce new short-term memory which will be passed on to the cell in the next time step. The output of the current time step can also be drawn from this hidden state.

Advantages

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 5, No. 3, May : 2023

- The primary advantage of this technique is that it allows an automated way of learning the patterns hidden inside the ocean of data and is capable of learning from the decisions.
- Simple and Easy manageability.

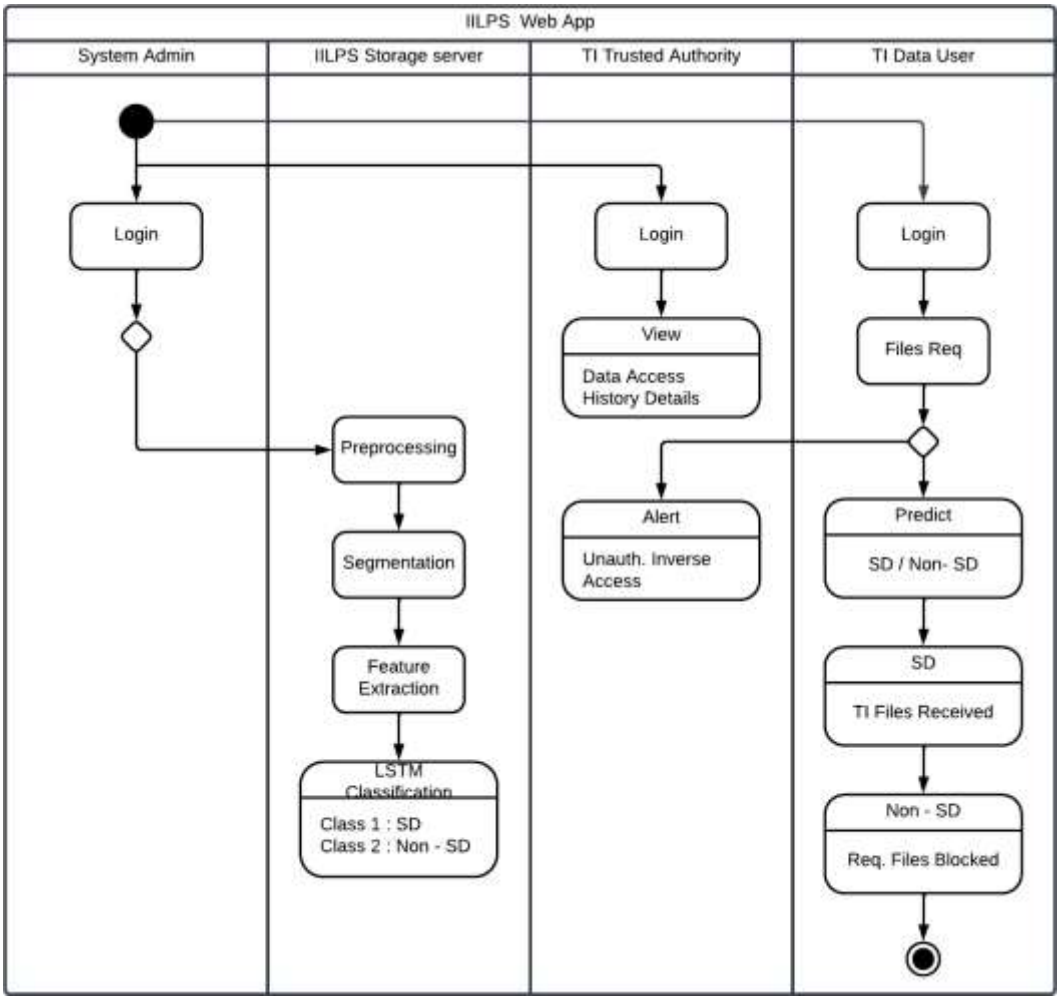
IV. HARDWARESPECIFICATION

- Processors: Intel® Core™ i5 processor 4300M at 2.60 GHz or 2.59 GHz (1 socket, 2 cores,2 threads per core), 8 GB of DRAM
- Disk space: 320 GB
- Operating systems: Windows® 10, macOS*, and Linux*

V. SOFTWARE SPECIFICATION

- Server Side : Python 3.7.4(64-bit) or (32-bit)
- Client Side : HTML, CSS, Bootstrap
- IDE : Flask 1.1.1
- Back end : MySQL 5.
- Server : Wampserver 2i
- LSTM DLL : Keras

VI. ACTIVITY DIAGRAM



VII. MODULES DESCRIPTION

1. IILPS Web Dashboard

The IILPS (Intelligent Information Leak Protection System) web dashboard is a user-friendly and interactive interface designed to display the results of the machine learning algorithms used to detect and prevent information leaks in TI (Technology and Information) companies. The dashboard is built using Python Flask and MySQL for the backend and Plotly Dash for the front-end data visualization. The design and development of the IILPS web dashboard involves several steps, including:

Building the backend with Python Flask: The next step is to build a secure and scalable backend for the web dashboard using Python Flask. This involves implementing authentication and authorization mechanisms for user access control and developing Flask routes and views for the web dashboard.

Creating the data visualization with Plotly Dash: The next step is to create interactive and user-friendly data visualization components for the web dashboard using Plotly Dash. This involves understanding the basics of data visualization and dashboard creation with Plotly Dash and implementing data visualizationand dashboard components for the IILPS web dashboard.

Integrating the backend and the front-end: The final step is to integrate the backend with the front-end using Flask and Plotly Dash. This involves developing the necessary communication channels between the backend and the front-end, integrating the machine learning models with the web dashboard, and deploying the web dashboard and IILPS system in a secure and scalable manner.

2. IILPS Model

The LSTM neural network is a type of recurrent neural network that is well-suited to sequential data analysis, making it a good choice for analyzing text data. Once the network has been trained on the dataset, it can be used to classify new data and alert users to potential information leaks.

2.1. Dataset Annotation

The first step in building such a system is to collect data from various sources such as emails, documents,chats, and other communication channels. This data can be used to train an LSTM model, which is a typeof recurrent neural network that can learn and predict sequences of data.

2.1.1. Digital Documents

Physical documents are scanned in batches and stored in a digital archive as a heterogeneous document stream, referred to as a digital package.

2.2. Preprocessing

Data Cleaning: Clean the data by removing any irrelevant information such as stopwords, punctuations, and HTML tags. Also, remove any personal information such as names and addresses to maintain privacy. **Tokenization:** Tokenize the data into individual words or phrases, which will be used as inputs to the LSTM model. Tokenization can be performed using techniques such as whitespace or punctuation splitting, or using more advanced methods such as natural language processing (NLP) libraries.

Text Normalization: Normalize the data by converting all text to lowercase, removing any special characters, and converting numbers to their word equivalents (e.g. "3" to "three"). This helps to reduce the complexity of the data and makes it easier for the model to learn.

2.3. Feature Extraction

Identify and extract relevant features from the dataset, such as keywords or metadata, that can help the LSTM model better distinguish between normal and potentially sensitive information.

2.4. LSTM Classification

LSTM Model Training: Train an LSTM model using the pre-processed data and extracted features. The model should be trained to classify documents as secure or non-secure based on their content.

Model Evaluation: Evaluate the performance of the model using a validation set of documents that were not used during training. Use metrics such as accuracy, precision, recall, and F1 score to assess the performance of the model.

Deployment: Deploy the model as part of the self-learning intelligent information leak protection system to classify incoming and outgoing documents in real-time. Any documents classified as non-

secure should be flagged and investigated for potential information leaks.

3. Prediction Module

Threshold Determination: Determine a threshold value for the model that indicates the likelihood of adata leak. This value should be set based on the company's risk tolerance and can be adjusted over time as needed.

Real-time Prediction: Deploy the model as part of the self-learning intelligent information leak protection system to monitor the data flow in real-time. The model should analyze incoming and outgoing data packets and predict the likelihood of a data leak. If the predicted likelihood exceeds the threshold value, an alert should be generated for further investigation. The proactive detection module can help to identify potential vulnerabilities and weaknesses in the company's network and proactively detect and mitigate information leaks before they occur. This can help to prevent data breaches and reduce the riskof damage to the company's reputation and financial well-being.

4. Access Control Mechanism

Two types of access control are used in the proposed SDMS. Role-Based Access Control (RBAC) is used to validate that user perform only authorized actions with the digital documents. This validation is enforced at each module of the SDMS. Mandatory Access Control (MAC) is used to validate that only the SDMS performs read and write operations on the Document Repository. On an access control failure, theSDMS stops any action requested of it.

5. End User Interface

An end-user interface for an LSTM based self-learning intelligent information leak protection system forTI companies should be designed with simplicity, ease of use, and effectiveness in mind. The interfaceshould provide users with access to the information they need to make informed decisions about thesecurity of their data. Here are some key features that could be included in the end-user interface: **Dashboard**: A dashboard can provide an overview of the system's performance and status. It shouldinclude key metrics such as the number of detected anomalies, the number of alerts generated, and theoverall risk score.

Alert Management: A section for managing alerts generated by the system should be included in the interface. This section should allow users to view, filter, and acknowledge alerts. Users should also be able to take action on alerts by marking them as false positives or escalating them to the security team.

Risk Management: A section for managing the overall risk score of the system should be included. Users should be able to view the risk score over time, understand what factors are contributing to the score, andtake action to reduce the risk score.

System Configuration: A section for configuring the system should be included. Users should be able to configure the system's thresholds, adjust settings related to data collection and pre-processing, and customize the types of alerts generated.

User Management: A section for managing users and permissions should be included. This section should allow administrators to manage user accounts, assign roles and permissions, and configure user access tothe system.

Reporting: A section for generating reports should be included. Reports could include metrics such as the number of detected anomalies, the types of alerts generated, and the overall risk score. Reports can be used to demonstrate the effectiveness of the system to stakeholders.

6. End Users

The end users of an AI based self-learning intelligent information leak protection system for TI companies can vary depending on the organization's structure and needs. Here are some potential end users: **Security Analysts:** Security analysts are responsible for monitoring the system's alerts and investigatingpotential security incidents. They use the system to detect and mitigate information

leaks, and ensure thesystem is working effectively.

IT Administrators: IT administrators are responsible for managing the system's configuration and settings. They use the system to ensure the system is configured correctly and the data is being collectedand processed accurately.

Business Executives: Business executives use the system to gain insights into the overall security posture of the organization. They use the system's reporting features to understand the system's performance andto make strategic decisions regarding the organization's security.

Compliance Officers: Compliance officers use the system to ensure that the organization is meeting regulatory requirements related to data protection and privacy. They use the system to ensure that the organization is compliant with relevant laws and regulations.

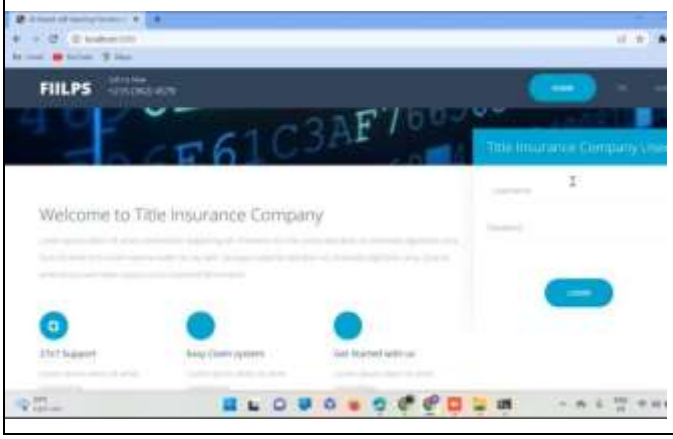
End Users: End users of the organization may use the system to ensure that their activities comply withthe organization's security policies. They may use the system to understand the risks associated with theirwork and to take appropriate actions to protect sensitive data.

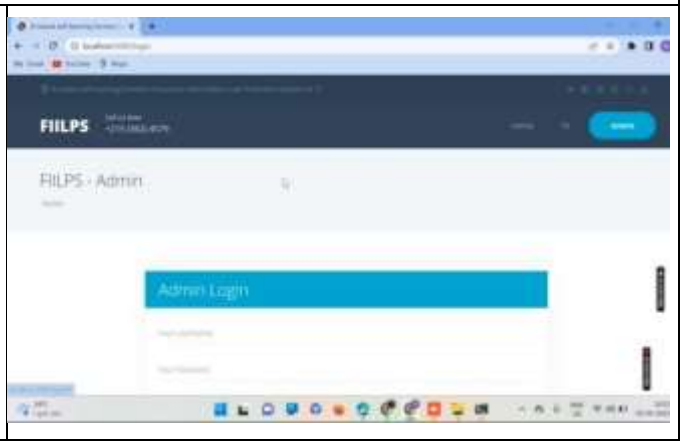
By providing an interface that meets the needs of these end users, the self-learning intelligent information leak protection system can help to improve the organization's overall security posture and reduce the riskof data breaches.

VIII. CONCLUSION

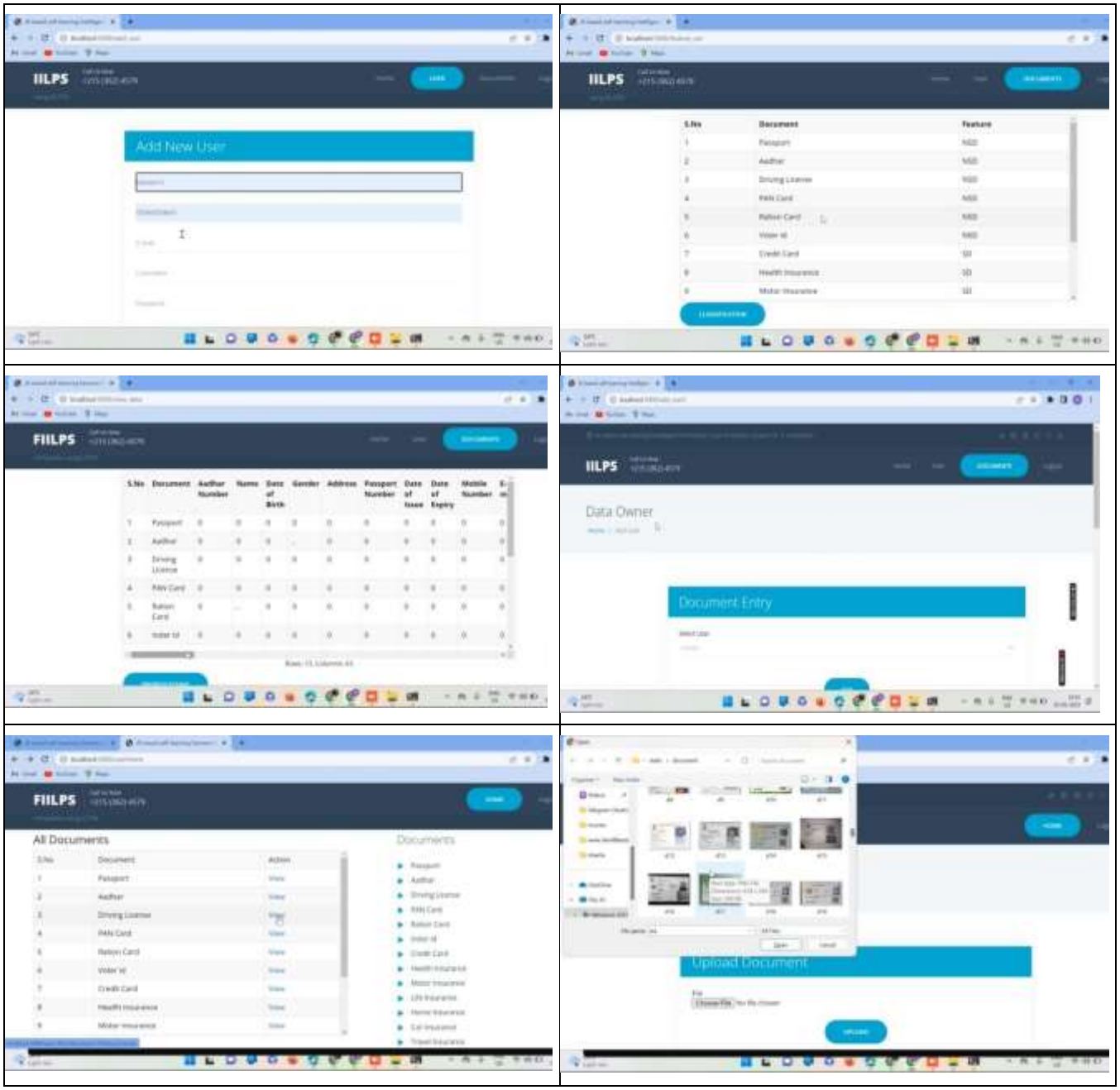
In conclusion, the Self-Learning Intelligent Information Leak Protection System for Title Insurance Companies Documents like PI, PII and NPI using LSTM is a highly effective system that accurately detects potential information leaks and provides real-time alerts to security teams. The system is capable of handling various document formats and large volumes of data, making it a suitable solution for title insurance companies. The testing process, including data validation, training and validation, integration,performance, security, user acceptance, and compliance testing, showed that the system is stable, reliable, and scalable, with high accuracy in detecting potential information leaks. The system's performance was satisfactory, as it was able to handle large volumes of data and network traffic, was stable and reliable under heavy load, and was able to scale up and down as needed. The system was compliant with regulations and industry standards related to data protection and privacy, and the user interface was user- friendly and easy to use. It is also compliant with regulations and industry standards related to data protection and privacy. However, to further improve the system's performance and accuracy, optimizing the LSTM model's hyper parameters, enhancing the system's scalability, and improving the system's ability to handle different types of data formats are recommended. Overall, the Self-Learning Intelligent Information Leak Protection System using LSTM is a promising solution that can help title insurance companies protect sensitive data from potential information leaks and ensure compliance with data protection and privacy regulations.

IX. SCREENSHOTS





Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 5, No. 3, May : 2023



X. FUTURE SCOPE

Here are some potential areas for future scope of the Self-Learning Intelligent Information Leak Protection System for Title Insurance Companies Documents like PI, PII and NPI using LSTM:

1. Multi-language support: The system can be enhanced to support multiple languages, which will enable it to detect sensitive information in title insurance documents written in different languages. This will be particularly useful in multinational companies or organizations that deal with clients and customers from different countries.
2. Integration with blockchain technology: The system can be integrated with blockchain technology to create a secure and immutable record of all title insurance documents. This will provide an additional layer of security and prevent any unauthorized modifications or tampering of documents.
3. Enhancing collaboration and sharing of information: The system can be enhanced to enable collaboration and sharing of information between different departments within the company. This will enable a more coordinated and effective approach to detecting potential information leaks and prevent

sensitive information from being disclosed.

4. Real-time analysis and response: The system can be enhanced to provide real-time analysis and response to potential information leaks. This can be achieved through the use of real-time stream processing and automated response mechanisms that can take immediate action when a potential leak is detected.
5. Incorporating natural language processing (NLP) techniques: By incorporating NLP techniques, the system can better understand the context and meaning of the text in title insurance documents. This can improve the accuracy of the system in detecting sensitive information and reduce false positives.

XI. REFERENCE

[1] R.Selvanambi, J.Natarajan, M.Karuppiah, S.K.H.Islam, M. M. Hassan, and G. Fortino,‘‘Lung cancer prediction using higher-or derrecurrent neural network based on glow worm swarm optimization,’’ Neural Comput. Appl., vol. 32, no. 9, pp. 4373–4386, 2020.

[2] M. S. Ahmad and G. Bamnote, ‘‘Data leakage detection and data preventionusingalgorithm,’’Int.J.Comput.Sci.Appl., vol.6, no.2, pp. 394–399, 2013.

[3] A. Akhigbe and A. M. Whyte, ‘‘The Gramm-Leach-Bliley Act of 1999: Risk implications for the financial services industry,’’ J. Financial Res., vol. 27, no. 3, pp. 435–446, Sep. 2004. [4] S. Al-Fedaghi, ‘‘A conceptual foundation for data loss prevention,’’ Int. J. Digit. Content Technol. Appl., vol. 5, no. 3, pp. 293–303, Mar. 2011.

[5] A.A. Mamun,M.K.Hassan, andS.Van Lai,‘‘Theimpactof theGrammLeach-Bliley act on the financialservices industry,’’ J. Econ. Finance, vol. 28, no. 3, pp. 333–347, Sep. 2004.

[6] H.Alhindi,‘‘Aframeworkfordatalosspreventionusingdocumentsemantic signature,’’ Ph.D. dissertation, Dept. Elect. Comput. Eng., Univ. Victoria, Victoria, BC, Canada, 2019.

[7] H. Alhindi, I. Traore, and I. Woungang, ‘‘Preventing data leak through semantic analysis,’’ Internet Things, Jun. 2019, Art. no. 100073. [Online]. Available: https://www.sciencedirect.com/science/article/ abs/pii/S254266051930126X?via%3Dihub.

[8] H. AlKilani, M. Nasereddin, A. Hadi, and S. Tedmori, ‘‘Data exfiltration techniques and data loss prevention system,’’ in Proc. Int. Arab Conf. Inf. Technol. (ACIT), Dec. 2019, pp. 124–127.

[9] S. Alneyadi, E. Sithirasenan, and V. Muthukkumarasamy, ‘‘A survey on data leakage prevention systems,’’ J. Netw. Comput. Appl., vol. 62, pp. 137–152, Feb. 2016.

[10] A. L. Buczak and E. Guven, ‘‘A survey of data mining and machine learning methods for cyber security intrusion detection,’’ IEEE Commun. Surveys Tuts., vol. 18, no. 2, pp. 1153–1176, 2nd Quart., 2016.

[11] R. Chatterjee, T. Maitra, S. H. Islam, M. M. Hassan, A. Alamri, and G. Fortino, ‘‘A novel machine learning based feature selection for motor imagery EEG signal classification in Internet of medical thingsenvironment,’’ Future Gener. Comput. Syst., vol. 98, pp. 419–434, Sep. 2019.

[12] L. Cheng, F. Liu, and D. D. Yao, ‘‘Enterprise data breach: Causes, challenges, prevention, and future directions,’’ Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery, vol. 7, no. 5, p. e1211, Sep. 2017.

[13] M.Datardina and K.Leung,‘‘Informationleakage&datalossprevention,’’ IT Assurance Governance Inf., Jul. 2009.

[14] O. Farràs and J. Ribes-González, ‘‘Provably secure public-key encryption with conjunctive and subset keyword search,’’ Int. J. Inf. Secur., vol. 18, no. 5, pp. 533–548, Oct. 2019.

[15] D. Graupe, Principles of Artificial Neural Networks, vol. 7. Singapore: World Scientific, 2013.