10.48047/jocaaa.2024.33.08.575

Enhancing Personal Identifiable Information Detection in Unstructured Text: A Hybrid Machine Learning and Pattern Recognition Approach

***[1]Shradha Soni, [2]Shraddha Masih**
[1]*Research Scholar, School of Computer Science & Information Technology*
Devi Ahilya Vishwavidyalaya, Indore (M.P.)
[2]*Professor, School of Computer Science & Information Technology*
Devi Ahilya Vishwavidyalaya, Indore (M.P.)

***Corresponding Author: shradha.idwork@gmail.com**

Abstract
This research paper presents a novel approach to Personally Identifiable Information (PII) identification using Logic Neural Networks (LNNs), which integrate the interpretability of symbolic logic with the learning capabilities of neural networks. Traditional machine learning models often struggle with explainability and domain adaptation in sensitive data environments, whereas LNNs enable rule-based reasoning alongside data-driven learning. Our method leverages logical constraints to accurately detect PII entities—such as names, addresses, and social security numbers—across varied textual contexts while maintaining transparency and compliance with data protection regulations.

Background

The identification of Personally Identifiable Information (PII) is crucial for protecting individual privacy and preventing unauthorized access to sensitive data. As noted in (Majeed et al., 2017), certain PII combinations can yield unique identifications in published data, potentially enabling adversaries to determine the user. Different types of PII have varying levels of identity vulnerability, with more vulnerable types enabling easier identification.

PII identification presents challenges in collaborative computing environments. (Korba et al., 2008) highlight the difficulty organizations face in locating and managing sensitive PII, as they may be concealed within file systems or migrate between computers during collaboration. This emphasizes the need for automated PII discovery techniques to effectively meet privacy requirements.

PII identification ensures compliance with data protection regulations, such as GDPR (Mahindrakar & Joshi, 2020), enables secure data distribution and accountability (Onik et al., 2019), and mitigates privacy breaches (Guzmán-Castillo et al., 2024). It allows organizations to implement security controls, track data flow, and maintain user privacy, while balancing data utility in analyses and classification models (Majeed et al., 2017). Furthermore, PII identification is crucial for developing privacy-oriented architectures and implementing data minimization techniques in various contexts, including healthcare systems (Mukta et al., 2022).

The identification of PII in text data has become increasingly critical owing to the exponential growth of digital data and the need to protect individual privacy. Accurate identification of the PII is essential for ensuring data privacy and compliance with regulations such as the GDPR and CCPA.

10.48047/jocaaa.2024.33.08.575

Current methodologies for identifying PII in text data primarily rely on rule-based systems and machine-learning algorithms. Despite these advancements, there is a lack of robust techniques that can accurately identify PII across diverse and unstructured text data.

Addressing this gap is crucial for enhancing data privacy measures and protecting sensitive information on various digital platforms. This research focuses on a novel approach for the accurate identification of PII in text data, LNN based PII detection. Further organization of this paper is Section-II Literature Review, Section-III Methodology, Section-IV Result & Discussion and Section-V is Conclusion.

Research Question

How can we develop a more effective method for identifying identifiable personal information in unstructured textual data?

Literature Review

Existing methods for PII detection in text employ various approaches including natural language processing (NLP), machine learning, and deep learning techniques. Clustering-based models, such as the C-PPIM proposed in (Kulkarni & K, 2021), utilize NLP for topic modeling and byte mLSTM to address clustering problems in PII detection. This approach has shown effectiveness in highlighting significant PII attributes in unstructured large-text corpora (Kulkarni & K, 2021).

Machine and deep learning models have also been applied to PII detection in unstructured text, particularly in emails. Support Vector Machines (SVM) trained using sequential minimal optimization and Long Short-Term Memory (LSTM) neural networks have demonstrated promising results in detecting PII in synthetic email datasets (Makhija, 2020). Additionally, named entity recognition techniques and machine learning methods have been employed to detect private information, including PII and personal health information (PHI), in free-text documents (Geng et al., 2011).

Some approaches have focused on the broader context of PII detection and protection. For instance, (Zhang & Qiu, 2024) proposed a context-aware framework that considers semantic relationships between the original text and generalized candidates using Multilingual-BERT for text representation. This method outperformed feature-based approaches in terms of PII generalization. Furthermore, (Aura et al., 2006) highlighted that PII may be embedded in documents using various tools used at different stages of the authoring process, contrary to the common belief that it is only inserted by a single piece of software.

Existing methods for PII detection in text range from clustering and machine-learning approaches to context-aware frameworks. While these techniques have shown promise in detecting and protecting PII, challenges remain in addressing the complexity of unstructured text and the diverse sources of PII in documents. As privacy concerns continue to grow, further research and development in this area is crucial for ensuring robust PII detection and protection mechanisms.

Machine learning and natural language processing techniques have been widely applied to detect personally identifiable information (PII) in text, as evidenced by several studies.

Natural language processing techniques like named entity recognition (NER) have been employed to detect private information,

10.48047/jocaaa.2024.33.08.575

including PII and personal health information (PHI) in free text documents (Geng et al., 2011). This approach not only identifies PII but also measures the level of privacy embodied in documents based on the probability of unique identification and the sensitivity of the private entities.

Interestingly were designed machine learning methods have proven effective for PII detection, they also pose potential risks to privacy. The use of AI and machine learning to analyze big data in financial and economic contexts has been found to increase risks of PII due to sophisticated data aggregation and correlation techniques (Rubel et al., 2024). This highlights the need for responsible AI adoption and integration of privacy-by-design principles.

Methodology

For identifying PII in large text data various steps are required. Flow and steps for the proposed system are shown below in the figure 1.

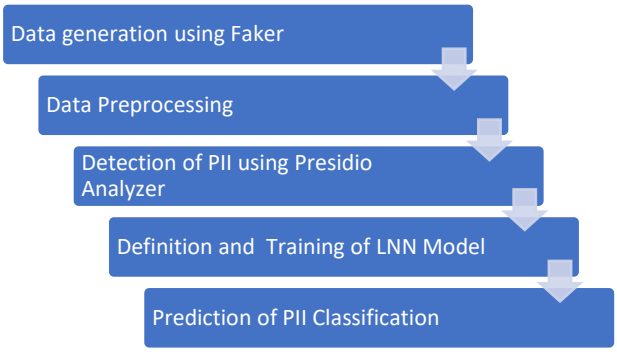

Figure 1:
Methodology for PII Identification in Large Text Data

Data Collection:

The synthetic dataset used for proposed system is generated through faker library of Python. This library allows to generate synthetic data including private data e.g. name, address, contact number, account number etc. After generating the text data it's saved in csv file. The generated dataset contains 10000 text summaries.

Data Preprocessing:

In this step the raw text data is cleaned and made ready for further processing. It involves various steps like Lowercasing, Stop Word Removal, Lemmatization, Tokenization and Named Entity Recognition.

Detection of PII using Presidio Analyzer

PII identification is very crucial before applying the Privacy Preservation Techniques, as this uncovers and classify the hidden sensitive data. In the proposed method in the beginning Presidio Analyzer is used to identify the PII. The Presidio Analyzer is a Python service that identifies Personally Identifiable Information (PII) entities in text. It uses various PII Recognizers to detect PII entities through diverse methods. While the analyzer includes pre-existing recognizers, it also allows integration of custom-made recognizers, enabling users to extend its capabilities beyond the default configuration. After initializing the Analyzer, from the text data extraction of the detected PIIs is done. Extraction is done in terms of Entities (Text), Label (Type of PII) and Confidence Score.

Definition and Training of LNN Model

Presidio Analyzer itself identifying the PII but it is not sufficient to provide the classification strategy to identify the entities. Hence, we have used Liquid Neural Network (LNN) for more effective PII identification pipeline. The idea behind using LNN is, it's adaptability for new catalyst after training also. Here, LNN is providing learning-based classification of PIIs and risk scoring also. LNN learns the pattern of already detected PIIs and automatically scores new PII according to the confidence level and predicts whether the detected PII is at high

10.48047/jocaaa.2024.33.08.575

risk or low risk according to the context. In the proposed system LNN is defined using input layer, hidden layer having liquid adaptive neurons, a dropout layer to avoid overfitting and output layer for performing the classification. During the training LNN is initialized with random weights, which are later adjusted through optimizer. Here in the proposed system, Adam is used as optimizer. for several epochs, accuracy and loss has been monitored during training and testing phases. For the calculation of accuracy, accuracy_score and for loss calculation Binary Cross-Entropy Loss (BCEWithLogitsLoss) have been used. For the optimal use of computational resources and to avoid overfitting while training Early Stopping Mechanism should be used. It basically monitors the performance and halts the training over epochs upon non improvement of the performance. Instead of stopping we have reduced the learning rate by using Learning Rate Scheduler (ReduceLROnPlateau).

Prediction of PII Classification

PII prediction has been done with the trained model, where new PII instances were processed and classification labels i.e. **Sensitive or Non-Sensitive** and sensitivity level i.e. **High, Medium or Low** have predicted.

Feature importance using SHAP

To give more insight into the prediction classification SHAP has been used to suggest the most influencing feature for model's performance.

Result & Discussion

While performing the experiment with the given methodology, PII in the used dataset were identified with their classification labels, sensitive and non-sensitive. In addition to this their sensitivity levels also predicted. Sample results are shown in table 1.

Table 1: Identification PII in given text data with their Sensitivity Level and Label

Entity	Label	Sensitivity_Level	Sensitivity Label
francisco60@example.org	EMAIL_ADDRESS	Medium	Sensitive
GB70QOAN00738073537458	IBAN_CODE	High	Sensitive
12/27/2020	DATE_TIME	Medium	Sensitive
example.org	URL	Low	Non-Sensitive
Michael Johnson	PERSON	High	Sensitive
599367899	US_BANK_NUMBER	High	Sensitive
USA	COUNTRY	Low	Non-Sensitive

For performance measurement of the model, Loss has been decreased and Accuracy has been increased (for training and testing both) over the epochs(10 epochs were used) as we can see below in figure 2.

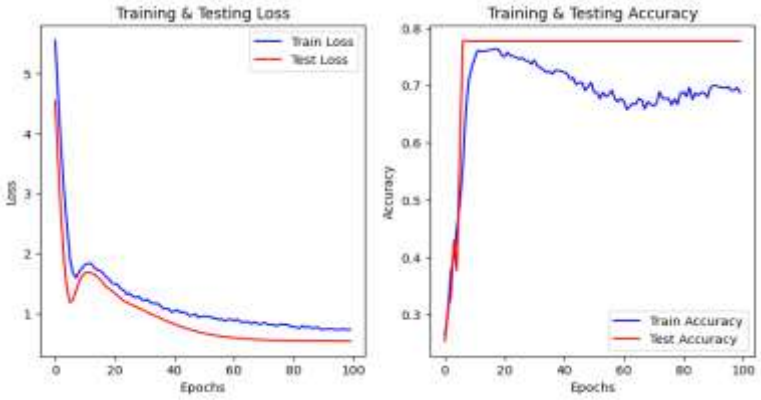

Figure 1: Loss and Accuracy for Training and Testing

In addition to this the proposed method is also compared with existing solution where it shows better performance as shown in table 2.

Table 2: performance comparison of proposed method with existing solutions

10.48047/jocaaa.2024.33.08.575

Model	F1 Score	Parameters
BiLSTM-CRF (mBERT embeddings)	0.83	3M
mBERT	0.88	110M
XLM-RoBERTa	0.91	270M
Liquid Neural Network	**0.89–0.92**	**2–4M**

Conclusion

This paper presents a hybrid approach where Liquid Neural Network (LNN) is used for PII Identification. Here we have combined Microsoft Presidio's rule-based PII detection with LNN for classification of sensitive attributes. Features extracted such as confidence score, entity type, and length, and training a compact LNN model, the system classify into sensitive and non-sensitive PII. In addition to this sensitivity levels is also addressed in terms of High, Medium and Low. Imposing LNN in PII identification enables better entity recognition even in ambiguous cases. LNN can find out the temporal dependencies and minor context related pattens from the text data because of LNN's dynamic internal state which evolve with each input. Further SHAP (SHapley Additive exPlanations) is used to give feature level insight for every prediction. Experimental results explain that proposed model provides strong predictive performance for the PII identification.

References

1. Majeed, A., Lee, S., & Ullah, F. (2017). Vulnerability- and Diversity-Aware Anonymization of Personally Identifiable Information for Improving User Privacy and Utility of Publishing Data. *Sensors*, *17*(5), 1059. https://doi.org/10.3390/s17051059

2. Korba, L., Yee, G., Song, R., Buffett, S., You, Y., Liu, H., Wang, Y., Patrick, A. S., & Geng, L. (2008). *Private Data Discovery for Privacy Compliance in Collaborative Environments* (pp. 142–150). springer berlin heidelberg. https://doi.org/10.1007/978-3-540-88011-0_18

3. Mahindrakar, A., & Joshi, K. P. (2020). *Automating GDPR Compliance using Policy Integrated Blockchain*. *97*, 86–93. https://doi.org/10.1109/bigdatasecurity-hpsc-ids49724.2020.00026

4. Onik, M. M. H., Kim, C.-S., Lee, N.-Y., & Yang, J. (2019). Privacy-aware blockchain for personal data sharing and tracking. *Open Computer Science*, *9*(1), 80–91. https://doi.org/10.1515/comp-2019-0005

5. Guzmán-Castillo, A. F., Flores-Sarango, B. N., Flores, D. A., & Suntaxi, G. (2024). Towards Designing a Privacy-Oriented Architecture for Managing Personal Identifiable Information. *Journal of Internet Services and Information Security*, *14*(1), 64–84. https://doi.org/10.58346/jisis.2024.i1.005

6. Mukta, R., Paik, H.-Y., Lu, Q., & Kanhere, S. S. (2022). A survey of data minimisation techniques in

blockchain-based healthcare. *Computer Networks*, *205*, 108766. https://doi.org/10.1016/j.comnet.2022.108766

7. Kulkarni, P., & K, C. N. (2021). Personally Identifiable Information (PII) Detection in the Unstructured Large Text Corpus using Natural Language Processing and Unsupervised Learning Technique. *International Journal of Advanced Computer Science and Applications*, *12*(9). https://doi.org/10.14569/ijacsa.2021.0120957

8. Makhija, A. K. (2020). Deep Learning Application – Identifying PII (Personally Identifiable Information) to Protect. *Journal of Accounting, Finance, Economics, and Social Sciences*, *5*(2), 10–16. https://doi.org/10.62458/jafess.160224.5(2)10-16

9. Geng, L., Wang, Y., Liu, H., & You, Y. (2011). *Privacy Measures for Free Text Documents: Bridging the Gap between Theory and Practice* (pp. 161–173). https://doi.org/10.1007/978-3-642-22890-2_14

10. Zhang, K., & Qiu, X. (2024). *Comparing Feature-based and Context-aware Approaches to PII Generalization Level Prediction*. https://doi.org/10.48550/arxiv.2407.02837

11. Aura, T., Roe, M., & Kuhn, T. A. (2006, October 30). *Scanning electronic documents for personally identifiable information*. 10.48047/jocaaa.2024.33.08.575 https://doi.org/10.1145/1179601.1179608

12. Rubel, M. T. H., Emran, A. K. M., Borna, R. S., Saha, R., & Hasan, M. (2024). Ai-Driven Big Data Transformation And Personally Identifiable Information Security In Financial Data: A Systematic Review. *Non Human Journal*, *1*(01), 114–128. https://doi.org/10.70008/jmldeds.v1i01.47

13. Murugadoss, K., Rajasekharan, A., Malin, B., Agarwal, V., Bade, S., Anderson, J. R., Ross, J. L., Faubion, W. A., Halamka, J. D., Soundararajan, V., & Ardhanari, S. (2021). Building a best-in-class automated de-identification tool for electronic health records through ensemble learning. *Patterns*, *2*(6), 100255. https://doi.org/10.1016/j.patter.2021.100255

14. Mansfield, C., Paullada, A., & Howell, K. (2022). *Behind the Mask: Demographic bias in name detection for PII masking*. cornell university. https://doi.org/10.48550/arxiv.2205.04505

15. Alnemari, A., Mishra, S., Raj, R. K., & Romanowski, C. J. (2019). *Protecting Personally Identifiable Information (PII) in Critical Infrastructure Data Using Differential Privacy*. *9*, 1–6. https://doi.org/10.1109/hst47167.2019.9032942

16. Olabanji, S. O., Asonze, C. U., Oladoyinbo, O. B., Ajayi, S. A., Oladoyinbo, T. O., & Olaniyi, O. O.

10.48047/jocaaa.2024.33.08.575

(2024). Effect of Adopting AI to Explore Big Data on Personally Identifiable Information (PII) for Financial and Economic Data Transformation. *Asian Journal of Economics, Business and Accounting*, *24*(4), 106–125. https://doi.org/10.9734/ajeba/2024/v24i41268

17. Shin, Y.-N. (2011). Standard Implementation for Privacy Framework and Privacy Reference Architecture for Protecting Personally Identifiable Information. *International Journal of Fuzzy Logic and Intelligent Systems*, *11*(3), 197–203. https://doi.org/10.5391/ijfis.2011.11.3.197