Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7641924

IDENTIFICATION AND PROCESSING OF PII DATA, APPLYING DEEP LEARNING MODELS WITH IMPROVED ACCURACY AND EFFICIENCY

Mainak Mitra
Product and Program Delivery Head,Snuvik Technologies

Soumit Roy
Architect Data Analytics, TCS

Abstract
Data governance involves an important aspect of data privacy for the enterprises to compliant with the data privacy standards like GDPR as the data usually involves sensitive personal identifiable information (PII). Across the cross-region collection and distribution of data, the de-identification and anonymization of PII data is mandatory for the security and privacy. In this paper, the potential machine learning and deep learning models are explored for the development of natural language processing (NLP) based large language model (LLM) for the automatic detection of PII data and its masking for implementing data privacy. Support Vector Machines (SVM), Random Forest (RF), Logistic Regressions (LR), Long Short-Term Memory (LSTM), and Multi-Layer Perceptron (MLP) models are trained on features extracted using Term Frequency-Inverse Document Frequency (TF-IDF) approach, for evaluating the performance in text classification of PII data. The implementation of a detection and masking of PII in presentation layer of data is proposed for improved data-anonymization.

INTRODUCTION
Data governance is more than mere management, it is day-to-day data management as it involves implementing policies and procedures to guarantee enterprise data availability, privacy, usability, integrity, and security. It's crucial for organizations looking to ensure that their data remains trustworthy so that it can be used to make business decisions, optimize operations, create new products and services, and improve profitability. Along with these aspects, the privacy and security of important data are also essential to be compliant with data protection regulations like GDPR [1]. A well-designed data governance program requires a well-defined set of procedures and protocols to ensure consistency across departments for data privacy and security, especially in enterprise reporting and analytics. Enterprises are highly responsible in this regard as they collect the personal data of individuals which is always required to be secure and private according to the General Data Protection Regulations (GDPR)[2]. The privacy standards are strictly followed to ensure that data is used properly so that potential misuse of sensitive and personal data can be avoided. Data security often hinges on traceability leading to unsure outcomes on how and who is using one's data[3]. In the modern cloud-based data warehouses, distributed systems are used to handle and manage data which increases the complexities and security challenges as data from different regions is made available in different regions. This distribution of data across regions requires essential privacy and security measures to ensure the privacy aspect of data governance[4].

Considering the significance and requirements of data privacy, the protection of Personal Identifiable Information (PII) is essentially required. PII data includes sensitive information like names, addresses, account details, and any other such personal identification data. Among other challenges in data governance, the privacy and security regarding PII data is one of the most important challenges. Traditionally, PII data security has relied on manual processes involving labor resources which is an error-prone approach and contains inadvertent disclosure risks[5]. Also, this manual process was manageable up to a limited size of data only but now the size of data is too large to be processed manually for tagging PII data and specifically securing it. Replacing these manual approaches, the approaches of machine learning based prediction systems were introduced for handling personal data[6]. The data warehouses contain data from different regions and they are accessed in different regions. The GDPR standards vary from region to region which means that in one region the data handling may be GDPR compliant but in another region, it may not be according to the GDPR standards of that region. So, there is a dire need for a system that masks and secures the PII data across regions.

In this paper, the potential ML-based approaches for improved data masking are proposed and evaluated. The proposed ML approach involves training of model on comprehensive data of different regions to develop a Large Language Model (LLM) which is known to be capable of classifying the PII data from other data so that it can be masked according to the relevant transformation logic[7]. As mentioned earlier, the GDPR standards vary from region to region, therefore, the data must also be masked accordingly. The designed and trained LLM will be able to identify whether the incoming data is PII, if it is then it must be masked when distributing. Also, according to our proposed approach, this data must be masked in the presentation layer when being distributed so that other operations on this data can stay efficient. The enterprise-level distributed data systems require efficient data governance regarding the privacy of PII data so there is a need for an automated LLM that can distinguish PII data from other data across different regions and mask the PII data accordingly.

LITERATURE REVIEW

Considering the importance of PII data privacy, there has been multiple researches that yielded many useful approaches of implementing privacy of PII data. The approach of de-identification of data using Safe Harbor method helped in de-identifying the data by shuffling, replacing and masking methods but later the technique proved inefficient as modern methods of re-identification were able to identify the data again[8]. In 1997, the anonymized data of Massachusetts state employee health insurance and in 2007 the data of Netflix subscribers was re-identified by using other data available on internet with the help of re-identification methodologies[9][10]. For making the anonymization securer, the modern and comprehensive approaches were identified that involved, suppression, generalization, controlled granularity, masking and encryption techniques using different methods[8]. Among these methods, anonymization of the PII data is also done using de-identification techniques of k-anonymity, L-diversity and T-closeness. K-anonymity property of data involves making the data anonymized up to a level when the information cannot be perceived at least k-1 individuals whose data is included. In this technique, the suppression and generalization of data is done for some value of k[11]. In homogenous data, the k-anonymity level is not equal to protecting corresponding values which are generalized and suppressed[12]. The enhanced version of k-

anonymity is L-diversity which include promotion of intra-group diversity for sensitive information. The efficiency of this technique depends upon the range of sensitivity attributes. If they are not diverse, the anonymity is limited and traceable[13]. The third technique of T-closeness is enhanced version of earlier mentioned techniques in which the granularity of data is further decreased in data representation[13]. In this approach, an equivalence class is set which basis upon the distance of a sensitive class attribute with all the attributes of data. T-closeness denotes the distance between these classes. Kaur et al. proposed a layered approach that involved multiple layers at different levels of data transactions involving masking and encryption in the third layer of data privacy and security in their model. They proposed use of machine learning and natural language processor algorithms to mask and encrypt the PII data of patients in a healthcare system. The researchers suggested masking of data using EHR masking techniques and for adding an extra and securer layer, they suggested crypto, matrix and specifically AES, RSA and SHA-256 algorithms[14]. This approach is with added security but the encryption and proposed masking can make the data transactions costly because it becomes mandatory to decrypt the data with shared keys before using it into analytics. The potential of machine learning can be used in the anonymization of PII data at different levels. i.e. from identification of PII data to the anonymization and de-identification of anonymized data[11]. The reviewed techniques are considerably helpful in securing PII data but each of this comes with some limitations depending on the diversity of data and complexity of approaches. In the process of de-identification and anonymity, using the complex techniques of encryption cause the ETL processes complex as the data becomes less useful in analytics without decryption. For the enormous amounts of data, the processes of de-identification and re-identification becomes costly, making the transactions time taking for ETL and analytics[11]. Hence, there is a need of an approach which is cost effective and require least transformation of data but strongly ensuring security and privacy of PII data. There is a strong need for a cost-effective automated technique that can be implemented on the presentation layer of data. This will help reducing transformation of data at source and storage level and in the presentation layer, the data is always masked and anonymized. The power of machine learning models can be used to generate language models that can automatically detect and mask the PII data when presenting.

RESEARCH METHODOLOGY

For the detection of PII data, the application of machine learning is proposed for development of a large language model that can detect the PII data and mask it. For this purpose, a dataset containing text data of 10000 rows is considered[15]. For the training and efficient performance of applied models, supervised machine learning technique is used by labeling the data for having PII data in the text. Before labeling, the data is cleaned and preprocessed by performing string tokenization, stemming and removing stop words. For labeling of huge amount of text data included in the dataset, the regular expression is used for the phone numbers, emails, account numbers and license numbers while regarding names and addresses, the data is labeled manually. Using these techniques, the feature extraction is performed and the data is labeled for having PII data. The rows containing PII data are labeled true else they are labeled false. The prepared dataset is randomized to avoid overfitting and biasness.

The dataset is then split into x-train, x-test, y-train and y-test component. The ratio between the test and train split is kept 70% for train and 30% for test. After splitting, Term Frequency – Inverse Document Frequency TF-IDF vectorizer model is used to convert text data to vectors. TF-IDF is used to estimate the frequency of words in the dataset to estimate the weights and relevance of the words to the document. The extracted features included frequent vectors and used for training of models. In next stages, the machine learning and deep learning models of SVM, RF, LR, LSTM, MLP and RNN are fitted on the pre-processed dataset and the performance metrics are extracted for evaluation. The applied models are tuned by adjusting hyperparameters of each algorithm for attaining the optimal performance. In terms of scalability and performance, the tree navigation of machine learning models is controlled for suitable complexity and optimized performance and for the Multiple Layered Neural Networks based models, the number of epochs is controlled up to a suitable value for optimal complexity. The overview of complete flow of the methodology is provided in the below Figure: 1 while implementation and performance evaluation of each model is also provided in the next sections.

Figure 1:LLM development for PII detection

Support Vector Machines (SVM)

SVM classifier is supervised machine learning model that distributes different classes of the data points according to the TF-IDF features and maps with reference to an extracted hyperplane. The data points with strong differences are mapped away from hyperplane while the similar data points are mapped closer to the hyperplane. For detection of PII text, the two possible classes are PII data and non-PII data. The function for the binary classification is as follows:

$$f(x) = sign(w \cdot x + b)$$

In the above function 'x' is the input feature vector obtained from target text data, 'w' is the weight vector which is learned by model during training, 'b' is the Bias term and Sign function, classifies the target to be PII or non-PII. SVM performed well with an accuracy value of 85%, F1 score value of 89%, precision value of 87%, and recall value of 92%. The confusion matrix and ROC Curves are provided in below Figure 2 and Figure 3.

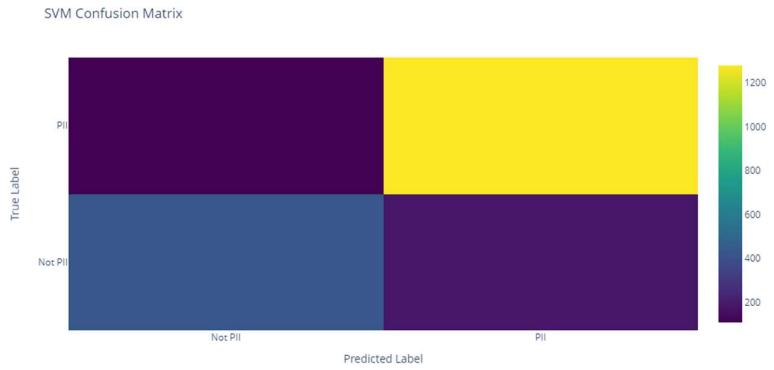

Figure 2: Confusion matrix for SVM

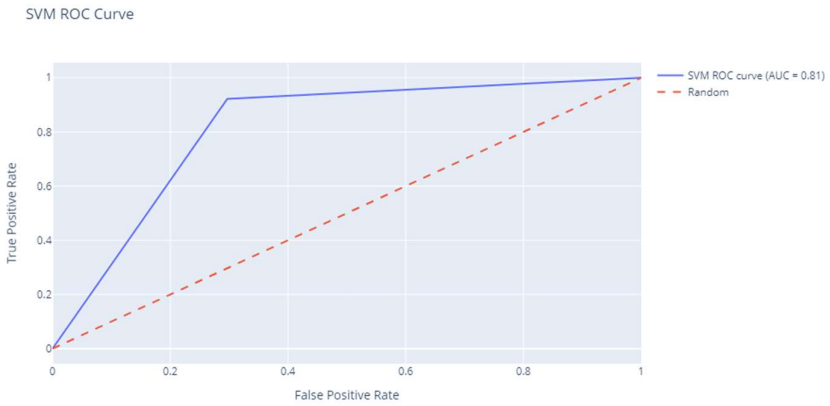

Figure 3:ROC Curve for SVM

Random Forest (RF)

Random forest (RF) with 100 decision trees is used as ensemble learning model for classifying between PII and Non-PII text data. The model is trained on TF-IDF features. A tree is created

for each class and the most frequent class has majority of votes and considered as prediction as described in the equation below:

$$RF\_Prediction(x) = MajorityVote(T1\ (x), T2\ (x), \ \dots, TN\ (x))$$

Here, x denotes the input and T denotes the tree. RF performed well with an accuracy of 81%, F1 score of 86%, precision of 85% and recall of 88%. The confusion matrix and ROC Curve is provided in the Figure 4 and Figure 5 below:

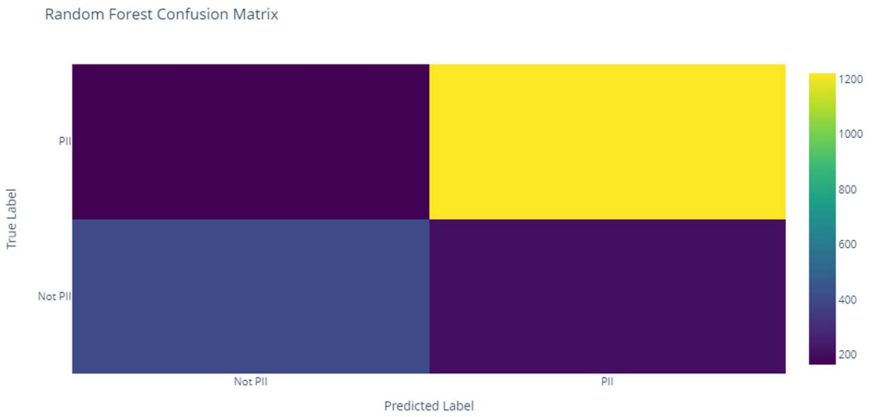

Figure 4: Confusion matrix for RF

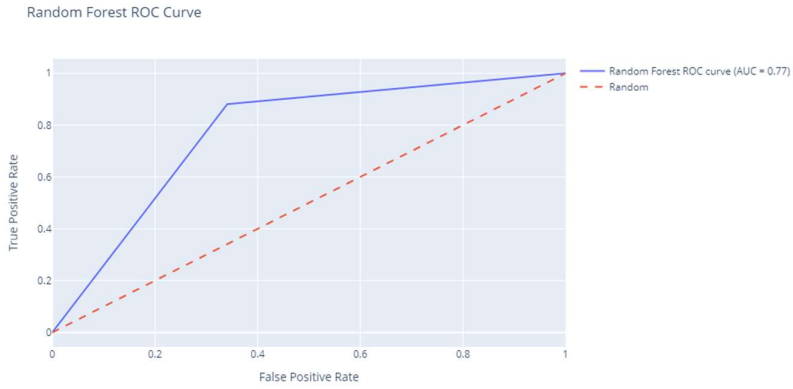

Figure 5: ROC Curve for RF

Linear Regression (LR)

The logistic regression model is trained on TF-IDF features for using its logistic function to estimate the probability of binary outcome for predicting PII or Non-PII class of text data. The LR model for the PII detection is implemented according to the formula given below:

$$P(PII \mid x) = 1 / (1 + e\text{^}\{-(w \backslash cdot\ x + b)\})$$

Here P is the probability that target variable of PII is equal to 1 for the given inputs of features as x. While w denotes the weighted summation of features, b denotes the baseline log-odds which are used along with sigma function to map the value between binary classes of 0 and 1 for PII and Non-PII. The LR model predicted the target PII existence in text data with an accuracy of 81%, recall of 96%, precision of 80% and F1 score of 87%. The confusion matrix and ROC Curve is provided in the Figure 6 and Figure 7 below:

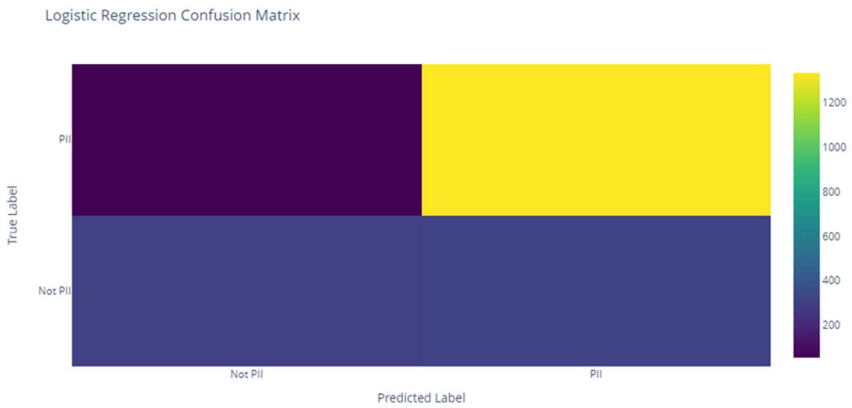

Figure 6: Confusion matrix for LR

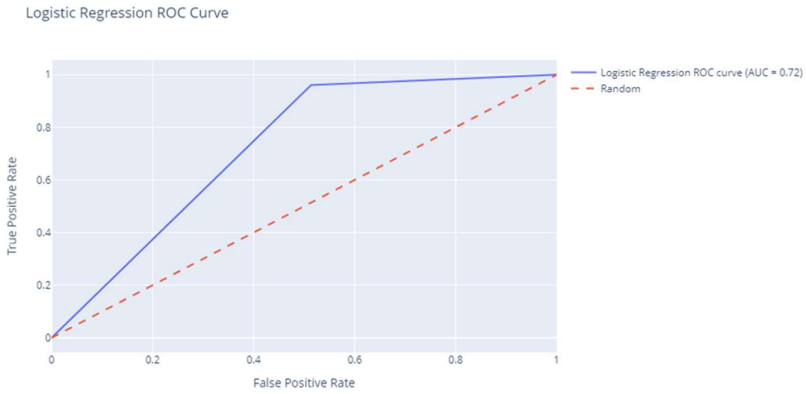

Figure 7: ROC curve for LR

Recurrent Neural Networks (RNN) and Long-Short Term Memory (LSTM)

Recurrent Neural Network (RNN) and RNN based LSTM model is also applied on TF-IDF features extracted from the dataset to predict the PII text. The implementation is performed considering the potential memory-based approach for long-range dependencies and sequence modeling. The below notation describes working of a single neuron where a is the output and considered activation neuron, a sigmoid activation function, w is the associated weight, X is the input and b is the bias term.

$$a = \sigma(W \cdot X + b)$$

The input is multiplied according to the weight and summed up with the biased function which later treated by sigmoid function to get the binary values for classification of PII and Non-PII data. Base RNN model provided an accuracy of 86%, recall of 95%, precision of 86%, and F1 score of 90%. The LSTM performed well and after 3 epochs, the model reached the optimal accuracy of 90%, recall of 91%, precision of 92% and F1 score of 92% for the dataset being used. Figure 8 and Figure 9 represents the accuracy for each epoch for RNN and LSTM respectively.

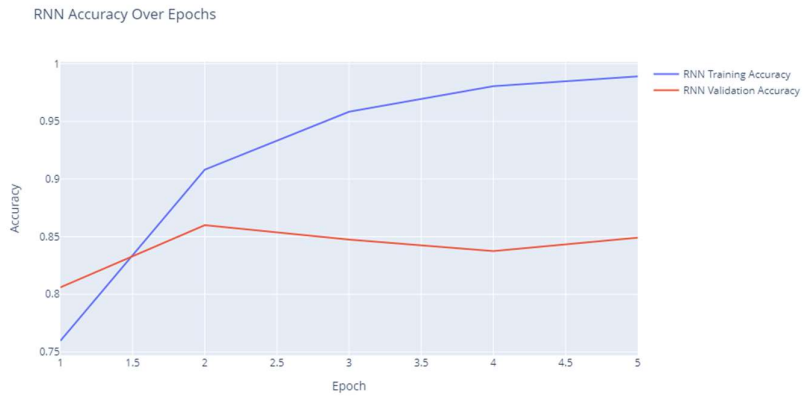

Figure 8: RNN Accuracy Over Epochs

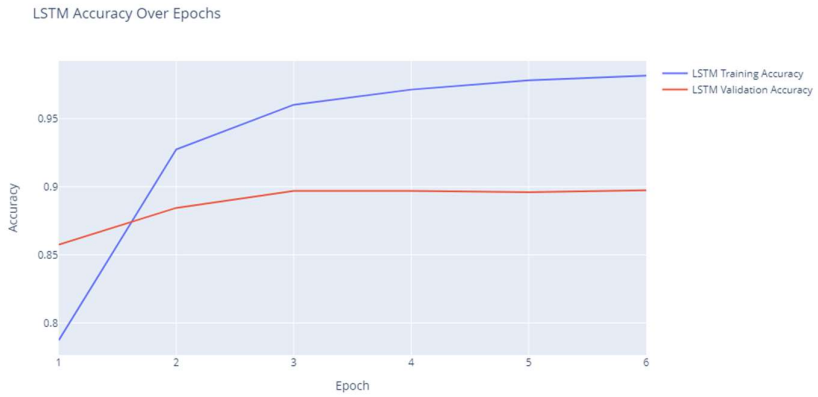

Figure 9: LSTM Accuracy over Epochs

Multi-Layer Perceptron

Multi-Layer perceptron (MLP) model is neural network of interconnected nodes and includes two layers with weights and biases along with the activation function of ReLU for hidden layer and activation function of sigmoid for the target output.

$$\text{Output}=\sigma(W_2 \cdot \text{ReLU}(W_1 \cdot \text{Input}+b_1)+b_2)$$

MLP learns patterns using the hidden layers and activation functions to classify the PII text and Non-PII text. The performance evaluation of MLP model provided an accuracy of 89% in the 6th epoch along with recall of 91%, precision of 86% and F1 score of 90%. Figure 10 presents the accuracy in each epoch.

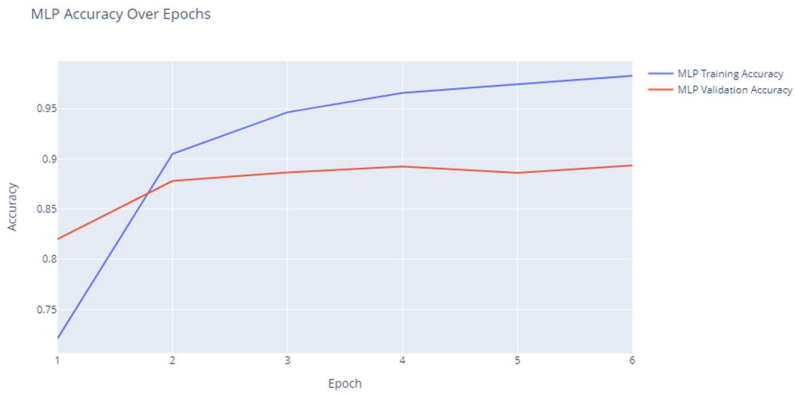

Figure 10: MLP Accuracy Over Epochs

PERFORMANCE EVALUATION AND IMPLICATIONS

The potential of machine learning makes it convenient to automatically label the PII data making it easier to handle securely. Currently, multiple commercial-level approaches are being applied in data governance that use machine learning for sensing PII data but still, these approaches need to be improved as they show disparities when the data is being accessed in different regions. This results in a higher risk of PII data of individuals from particular demographic regions being exposed. The error rates of these ML-based systems are significant when applied in different regions. Commonly, data masking is performed to secure the PII data but this also involves a challenge as in analytics, using the masked data becomes inefficient. Considering these significant challenges in PII data privacy, there is a need for a system that automatically masks the data in the presentation layer in such a way that it is always secure and masked according to the GDPR of the region without limiting the use of data for analytics. Ensuring PII data privacy using an improved and efficient approach can help avoid PII data breaches and make organizations more compliant with privacy standards. In this study, different machine learning and deep learning-based models are applied to evaluate their performance and implementation in the detection of PII data. Among the applied learning models, LSTM, which is RNN based neural network model, outperformed other models in terms of accuracy of predictions. A performance metrics comparison of the applied language models is provided in the Figure 11 and Figure 12.

Model	F1 Score	Precision	Recall	Accuracy
SVM	0.8981741573033708	0.8754277891854894	0.9221341023792358	0.855
Random Forest	0.8672817601135556	0.8539482879105521	0.8810382119682768	0.813
Logistic Regression	0.8780487804878049	0.8087431693989071	0.9603460706560922	0.815
LSTM	0.9251453488372093	0.9326007326007326	0.9178082191780822	0.897
MLP	0.9178429243575823	0.9215116279069767	0.9142033165104542	0.8865
RNN	0.9040438656614119	0.8615284128020901	0.9509733237202596	0.86

Figure 11: Performance Evaluation of Applied Models

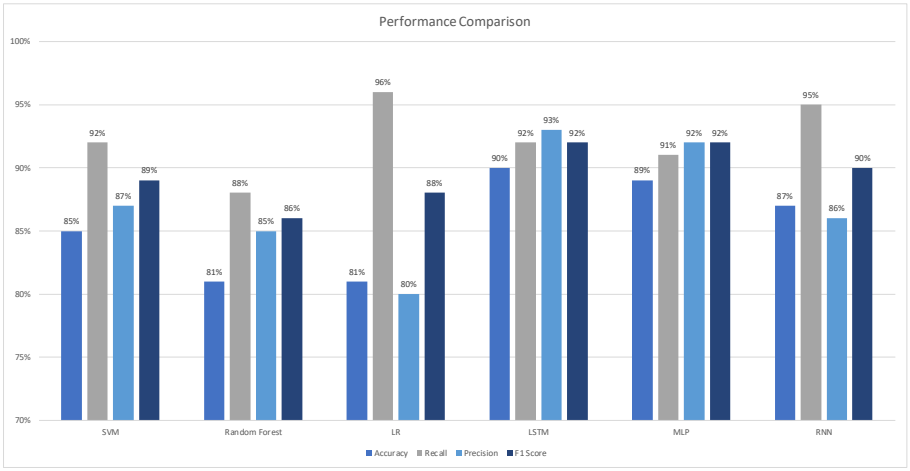

Figure 12: Performance Evaluation of Applied Models

Considering the prediction problem of PII text data, language-based model using LSTM can be preferred for efficient detection. For further improvement and accuracy, the ensembled learning using LSTM, MLP and SVM can be used for the development of a large language model, able to predict PII text. Regarding the practical implementation, in the presentation layer of data, the LLM based prediction model can be used to detect the PII data and once it is detected, it is easier to mask the text labeled by PII detection LLM model. The study included a dataset of limited size but with the training of models on the larger dataset can enhance the accuracy and efficiency of the LLM for accurate prediction of the PII data at enterprise level.

CONCLUSION

The enterprises collect and distribute data from different regions which involves data privacy and security aspects specifically related to the GDPR compliance. Ensuring the protocols and strategies for de-identification and anonymization an automated and accurate approach is required. In this paper, the potential machine learning and deep learning algorithms for the development of large language models are evaluated regarding their performance for prediction of PII data. The results described that the neural-network based models can be used for the development of NLP based large language models that can flag the PII data. Once the data is flagged for being PII, it can be easily masked in the application layer of data during distribution and use for applying anonymity of PII data.

REFERENCES

[1] G. Barta, "Challenges in the compliance with the General Data Protection Regulation: anonymization of personally identifiable information and related information security concerns," *Knowledge--economy--society business, Financ. Technol. as Prot. Support Soc.*, pp. 115–121, 2018.

[2] M. Goddard, "The EU General Data Protection Regulation (GDPR): European regulation that has a global impact," *Int. J. Mark. Res.*, vol. 59, no. 6, pp. 703–705, 2017.

[3] T. Linden, R. Khandelwal, H. Harkous, and K. Fawaz, "The privacy policy landscape after the GDPR," *arXiv Prepr. arXiv1809.08396*, 2018.

[4] N. Gruschka, V. Mavroeidis, K. Vishi, and M. Jensen, "Privacy issues and data

protection in big data: a case study analysis under GDPR," in *2018 IEEE International Conference on Big Data (Big Data)*, 2018, pp. 5027–5033.

[5] B. A. Beckwith, R. Mahaadevan, U. J. Balis, and F. Kuo, "Development and evaluation of an open source software tool for deidentification of pathology reports," *BMC Med. Inform. Decis. Mak.*, vol. 6, no. 1, p. 12, 2006, doi: 10.1186/1472-6947-6-12.

[6] O. Uzuner, T. C. Sibanda, Y. Luo, and P. Szolovits, "A de-identifier for medical discharge summaries.," *Artif. Intell. Med.*, vol. 42, no. 1, pp. 13–35, Jan. 2008, doi: 10.1016/j.artmed.2007.10.001.

[7] X. Li, J. Feng, Y. Meng, Q. Han, F. Wu, and J. Li, "A Unified {MRC} Framework for Named Entity Recognition," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Jul. 2020, pp. 5849–5859, doi: 10.18653/v1/2020.acl-main.519.

[8] G. Nelson, "Practical Implications of Sharing Data: A Primer on Data Privacy, Anonymization, and De-Identification," 2015.

[9] D. Barth-Jones, "The 'Re-Identification' of Governor William Weld's Medical Information: A Critical Re-Examination of Health Data Identification Risks and Privacy Protections, Then and Now," *SSRN Electron. J.*, 2012, doi: 10.2139/ssrn.2076397.

[10] A. Narayanan and V. Shmatikov, "Robust De-anonymization of Large Sparse Datasets," in *2008 IEEE Symposium on Security and Privacy (sp 2008)*, 2008, pp. 111–125, doi: 10.1109/SP.2008.33.

[11] P. Jain, M. Gyanchandani, and N. Khare, "Big data privacy: a technological perspective and review," *J. Big Data*, vol. 3, no. 1, p. 25, 2016, doi: 10.1186/s40537-016-0059-y.

[12] K. El Emam *et al.*, "A globally optimal k-anonymity method for the de-identification of health data.," *J. Am. Med. Inform. Assoc.*, vol. 16, no. 5, pp. 670–682, 2009, doi: 10.1197/jamia.M3144.

[13] Y. Sei, H. Okumura, T. Takenouchi, and A. Ohsuga, "Anonymization of sensitive quasi-identifiers for l-diversity and t-closeness," *IEEE Trans. dependable Secur. Comput.*, vol. 16, no. 4, pp. 580–593, 2017.

[14] P. Kaur, M. Sharma, and M. Mittal, "Big Data and Machine Learning Based Secure Healthcare Framework," *Procedia Comput. Sci.*, vol. 132, pp. 1049–1059, 2018, doi: 10.1016/j.procs.2018.05.020.

[15] "King-Harry/NinjaMasker-PII-Redaction-Dataset · Datasets at Hugging face." [Online]. Available: https://huggingface.co/datasets/King-Harry/NinjaMasker-PII-Redaction-Dataset.