

# Enhancing Personally Identifiable Information Recognition Using Compact Transformer With Hybrid Sequential Architecture

Pattarapoom Nokkaew<sup>1</sup>, Khantharat Anekboon<sup>2\*</sup>

<sup>1,2</sup> Department of Computer and Information Science, King Mongkut's University of Technology  
North Bangkok, Bangkok, Thailand

<sup>1</sup> <https://orcid.org/0009-0000-9083-1534>; <sup>2</sup> <https://orcid.org/0009-0005-9844-1975>

\*Corresponding Author: Khantharat Anekboon

**Abstract—** The rapid expansion of unstructured data across organizational repositories has significantly increased the risk of sensitive information leakage and regulatory non-compliance. Named Entity Recognition for Personally Identifiable Information plays a crucial role in automated data protection systems by identifying and classifying sensitive entities such as personal names, identification numbers, addresses, and financial information. While traditional rule-based approaches offer interpretability, they often suffer from limited generalization and low recall. In contrast, deep learning-based approaches have demonstrated superior performance but are frequently constrained by high computational complexity and memory requirements. This research investigates the effectiveness of compact Transformer architectures for PII-NER, focusing on the DeBERTa v3 xsmall model as the primary backbone. With approximately 22 million backbone parameters, DeBERTa v3 xsmall provides a favorable trade-off between model size and representational power. To further enhance sequence modeling capabilities, this study integrates Bidirectional Long Short-Term Memory, Bidirectional Gated Recurrent Unit, and Conditional Random Fields layers on top of the Transformer encoder. Experiments were conducted on the English portion of the AI4Privacy PII dataset containing approximately 43,000 annotated samples. The experimental results demonstrate that the hybrid architecture combining DeBERTa v3 xsmall + BiGRU + CRF achieves the best performance, with an F1-score of 0.9139 after 15 training epochs. The findings indicate that integrating compact Transformer backbones with sequential modeling and structured prediction layers significantly improves label consistency and entity boundary detection while maintaining computational efficiency. This study highlights the practical feasibility of deploying lightweight yet high-performance PII-NER systems in real-world environments with limited hardware resources.

**Keywords—** Compact Transformer, Named Entity Recognition, Personal Identifiable Information, Hybrid Sequential Architecture.

## I. INTRODUCTION

The exponential growth of digital information within organizational repositories has intensified concerns regarding data privacy and regulatory compliance [1]. Sensitive information embedded in unstructured text such as emails, contracts, reports, and customer records poses substantial risks if not properly identified and protected. Personally Identifiable Information (PII) refers to any data that can directly or indirectly identify an individual, including names, phone numbers, addresses, identification numbers, financial data, and authentication credentials.

Automated detection of PII using Named Entity Recognition (NER) has become an essential component of modern data protection pipelines. Early NER systems relied primarily on rule-based [2-4] or pattern-matching techniques, such as regular expressions [5-6]. Although these approaches offer interpretability and deterministic behavior, they often suffer from low recall, limited adaptability, and poor generalization to unseen data variations [7].

Recent advancements in deep learning, particularly Transformer-based language models, have significantly improved the state-of-the-art performance in NER tasks. Pre-trained language models enable contextual representation learning, capturing long-range dependencies and semantic relationships across tokens. However, large-scale models

typically require substantial computational resources, high memory capacity, and extended inference time, which restrict their deployment in real-world production systems.

Therefore, this research aims to explore whether compact Transformer architectures can provide competitive performance for PII-NER while maintaining efficiency suitable for consumer-grade hardware deployment.

## **II. BACKGROUND**

The NER is a sequence labeling task in which each token in a sentence is assigned a label according to a predefined tagging scheme. In the context of PII detection, the label space typically includes multiple sensitive entity types such as PERSON, EMAIL, PHONE, ADDRESS, CREDIT\_CARD, and ID\_NUMBER. Accurate identification of entity boundaries and consistent label transitions are critical for PII-NER systems, as misclassification may result in data leakage through false negatives or unnecessary masking of non-sensitive content through false positives. Consequently, model precision, robustness, and contextual understanding play a central role in ensuring reliable real-world deployment. Although large transformer-based language models achieve strong performance in NER and other natural language understanding tasks, their practical deployment is often constrained by latency, memory, and computational limitations. Compact transformer architectures have therefore emerged as effective alternatives, particularly for resource-constrained environments such as mobile devices, embedded systems, small servers, and edge devices where real-time PII detection is required. Models such as DistilBERT [8] and ALBERT [9] reduce model size through knowledge distillation, parameter sharing, and factorized embedding parameterization while maintaining competitive performance [10]. Additional compact models further explore architecture slimming and bottleneck design. MobileBERT [11] employs a bottleneck structure and rebalanced self-attention and feed-forward layers to significantly accelerate inference while preserving language understanding capability. TinyBERT [12] applies a two-stage knowledge distillation process during both pre-training and task-specific fine-tuning, enabling smaller configurations (e.g., 4-layer or 6-layer variants) to retain a high percentage of the performance of larger BERT models. Meanwhile, BERT-Tiny [13] represents an extremely lightweight alternative with only a few million parameters, suitable for experimentation or ultra-low-latency scenarios, though typically not for high-accuracy production settings. In this research, DeBERTa v3 xsmall [14] is selected as the primary backbone model due to its architectural innovations and parameter efficiency. It introduces a disentangled attention mechanism that separates content and positional representations, adopts an ELECTRA-style replaced token detection pre-training objective, and incorporates gradient-disentangled embedding sharing to enhance efficiency. Despite containing approximately 22 million parameters, DeBERTa v3 xsmall demonstrates competitive performance across multiple natural language understanding benchmarks. Its balance between compactness and contextual modeling capability makes it particularly well suited for PII-NER systems operating under resource constraints while still requiring high reliability and real-time responsiveness. [14]

Beyond compact Transformer architectures, recurrent neural networks (RNNs) remain highly relevant for sequence labeling tasks such as NER due to their inherent capability to model temporal dependencies. LSTM introduces gating mechanisms input gate, forget gate, and output gate to mitigate the vanishing gradient problem commonly observed in traditional RNNs. Bidirectional Long Short-Term Memory (BiLSTM) [15] has been widely adopted in NER systems [16-17] because of its ability to capture long-range contextual information in both forward and backward directions. By processing the input sequence

bidirectionally, BiLSTM enables the model to incorporate both past and future contextual cues when predicting entity labels for each token. This property is especially beneficial in PII-NER tasks, where entity boundaries and semantic meaning often depend on surrounding words. Consequently, BiLSTM layers are frequently integrated on top of Transformer encoders to further refine contextual representations and enhance sequence-level coherence.

Similarly, Bidirectional Gated Recurrent Unit (BiGRU) [18] serves as a computationally efficient alternative to BiLSTM while retaining the ability to model bidirectional contextual dependencies. GRU simplifies the LSTM architecture by combining the input and forget gates into a single update gate and employing a reset gate to control information flow. This streamlined structure reduces the number of parameters and computational overhead while maintaining competitive performance in sequence modeling tasks [19]. In NER applications, BiGRU can effectively capture local and medium-range dependencies, making it suitable for scenarios where model efficiency and faster convergence are prioritized [20]. When integrated with Transformer backbones, BiGRU layers may enhance token-level representation refinement, particularly in resource-constrained environments where minimizing additional parameters is desirable.

In addition to recurrent layers, Conditional Random Fields (CRF) [21] are commonly employed as a structured prediction layer in sequence labeling tasks. Unlike token-level classification methods that predict labels independently, CRF models the dependencies between adjacent labels and enforces valid transition constraints across the entire sequence. In the context of the BIO tagging scheme, certain transition such as an Inside (I) tag without a preceding Begin (B) tag of the same entity type are invalid. A CRF layer learns transition scores between labels and jointly decodes the optimal label sequence using global sequence-level optimization, typically via the Viterbi algorithm. This structured decoding mechanism significantly improves entity boundary consistency and reduces fragmented or logically inconsistent predictions. In PII-NER tasks where precise boundary detection is critical for compliance and data protection, the integration of a CRF layer on top of contextual encoders such as Transformers and BiLSTM/BiGRU has been shown to enhance robustness and overall F1 performance [16-17].

### **III. METHODOLOGY**

#### **A. Dataset**

This study utilizes the English portion of the AI4Privacy PII dataset [1], consisting of 43,500 annotated examples. Each sample contains token-level annotations following the BIO tagging scheme. The dataset covers a diverse set of PII entity types, enabling evaluation across multiple sensitive categories. The dataset was divided into training, validation, and testing subsets ratio 80:10:10 to ensure unbiased evaluation.

#### **B. Model Architecture**

The proposed framework consists of two primary components (Transformer Encoder and Sequential and Structured Prediction Layers), The Several architectural variants were evaluated (shown in Fig. 1).

Config 1: DeBERTa v3 xsmall (baseline)

Config 2: DeBERTa v3 xsmall + BiLSTM

Config 3: DeBERTa v3 xsmall + BiGRU

Config 4: DeBERTa v3 xsmall + BiLSTM + CRF

Config 5: DeBERTa v3 xsmall + BiGRU + CRF

In the baseline configuration, the contextual embeddings produced by the DeBERTa v3 xsmall encoder are passed through a dropout layer for regularization. This dropout operation aims to reduce overfitting and improve generalization during fine-tuning. The embeddings are fed into a fully connected layer for token-level classification. The decoder selects the maximum value across label logits for each token, and the model is trained using CrossEntropyLoss under the BIO tagging scheme. In this configuration, each token is classified independently without explicitly modeling label transition dependencies across the sequence.

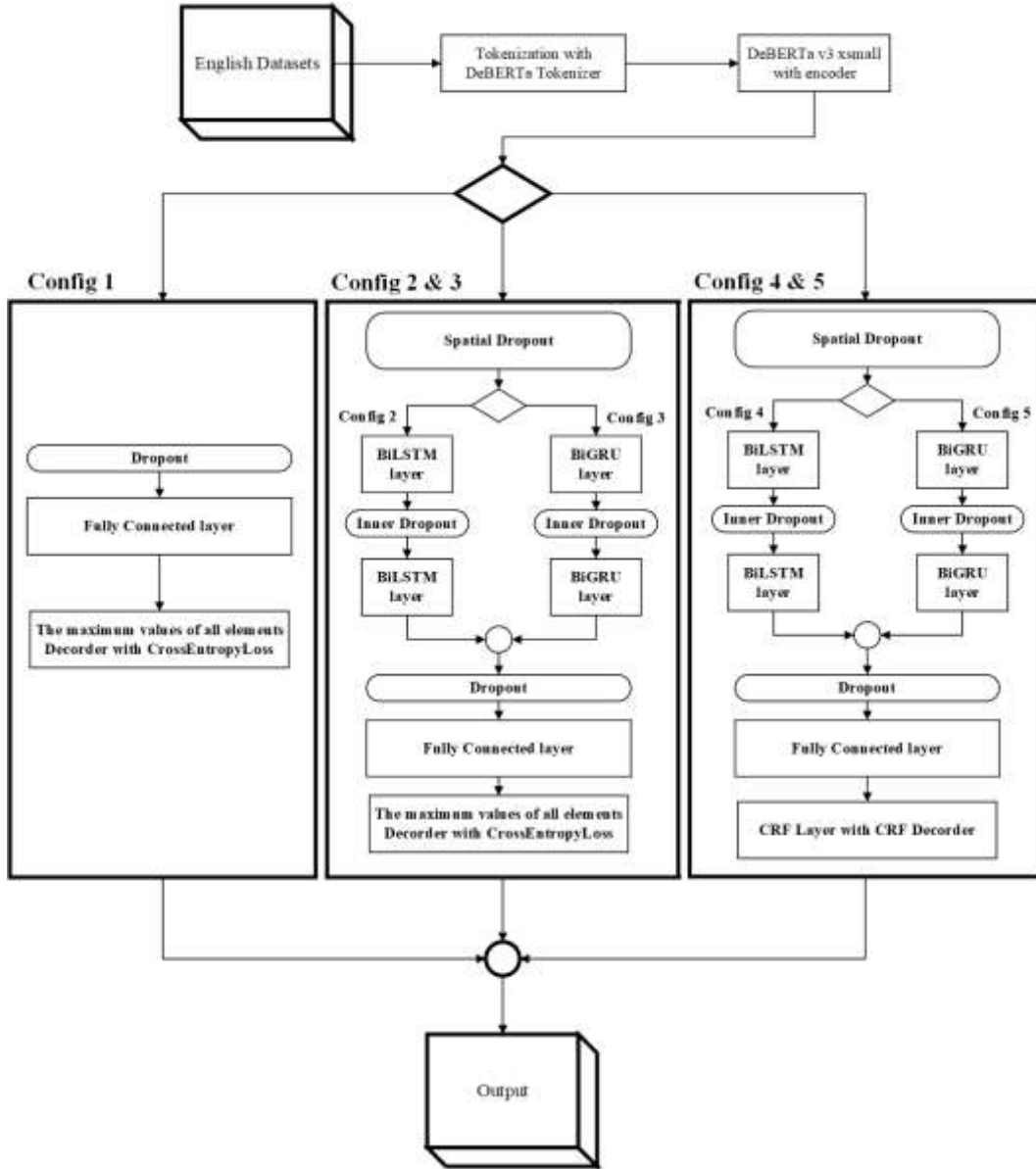
In Configurations 2 and 3, the baseline model is extended by introducing bidirectional recurrent layers after the Transformer encoder. Specifically, the contextual embeddings are first processed with Spatial Dropout [22], which randomly drops entire embedding channels to improve robustness in sequence modeling. The regularized embeddings are then passed through a stacked bidirectional recurrent structure consisting of two BiLSTM layers (Config 2) or two BiGRU layers (Config 3), with an Inner Dropout layer applied between them. This design allows deeper sequential feature extraction while mitigating overfitting. Following recurrent processing, a dropout layer is applied before the fully connected layer performs token-level classification. Similar to the baseline model, decoding is conducted by selecting the maximum logit value per token, and the model is optimized using CrossEntropyLoss without structured sequence constraints.

Configurations 4 and 5 further extend the recurrent architecture by incorporating a CRF layer as the final decoding component. The architecture begins with the DeBERTa encoder, followed by Spatial Dropout and two stacked bidirectional recurrent layers (BiLSTM in Config 4, BiGRU in Config 5) separated by Inner Dropout. A final dropout layer is applied before passing the features to the fully connected layer. Instead of performing independent token-level classification, the CRF layer is applied as the final decoding layer to enforce valid label transition constraints and improve sequence-level consistency. Unlike CrossEntropy-based decoding, the CRF performs joint sequence decoding by modeling transition scores between adjacent labels. This structured prediction mechanism ensures that invalid BIO transitions are minimized and that entity spans remain coherent across the entire sequence. The integration of recurrent layers and CRF enables the model to combine contextual representation learning, sequential dependency modeling, and globally optimized label decoding within a unified framework.

### ***C. Training Configuration***

Experiments were conducted using Intel Core i7 (13th Gen) of CPU, 24 GB of RAM and NVIDIA GeForce RTX 4060 Laptop GPU (with 8 GB of VRAM). The models were implemented using PyTorch and trained on GPU acceleration. The AdamW optimizer was employed to provide stable weight decay regularization. A differential learning rate strategy was applied, assigning a smaller learning rate to the pre-trained Transformer backbone and a larger learning rate to the newly added layers. The training process was conducted for 15 epochs.

**Fig. 1 The propose model architectures**



#### IV. EXPERIMENTAL RESULTS

To quantitatively evaluate the effectiveness of the proposed PII-NER models, standard classification metrics were employed, including accuracy (1), precision (2), recall (3), and F1-Score (4). These metrics provide complementary perspectives on model performance, particularly in sequence labeling tasks where class imbalance and boundary consistency are critical. While accuracy measures overall prediction correctness, precision and recall capture the trade-off between false positives and false negatives. The F1-Score, as the harmonic mean of precision and recall, offers a balanced evaluation of detection capability, especially in privacy-sensitive tasks where both missed entities and incorrect detections must be minimized.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

$$\text{F1 - Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

where TP (True Positive) is correctly predicted PII entities.

FP (False Positive) is incorrectly predicted PII entities.

TN (True Negative) is correctly predicted non-PII tokens.

FN (False Negative) is missed PII entities.

The experimental results presented in Table I demonstrate that integrating sequential modeling layers on top of the DeBERTa v3 xsmall backbone consistently improves overall performance compared to the standalone Transformer encoder. The baseline configuration achieved an F1-score of 0.9094, indicating that recurrent modeling can enhance contextual refinement depending on architectural choice.

Further performance gains were observed when incorporating the CRF decoding layer. The Config 4 model achieved an F1-score of 0.9056, while the Config 5 model obtained the highest F1-score of 0.9139. These findings indicate that structured decoding via CRF enhances label transition consistency under the BIO tagging scheme, particularly improving boundary coherence and recall performance.

**TABLE I THE RESULT OF ALL MODEL IN THIS RESEARCH**

| Model                                      | Accuracy | Precision | Recall | F1-Score |
|--------------------------------------------|----------|-----------|--------|----------|
| Config 1: DeBERTa v3 xsmall                | 0.9635   | 0.8936    | 0.9259 | 0.9094   |
| Config 2: DeBERTa v3 xsmall + BiLSTM       | 0.9642   | 0.8892    | 0.9089 | 0.8989   |
| Config 3: DeBERTa v3 xsmall + BiGRU        | 0.9654   | 0.8984    | 0.9223 | 0.9102   |
| Config 4: DeBERTa v3 xsmall + BiLSTM + CRF | 0.9632   | 0.8986    | 0.9127 | 0.9056   |
| Config 5: DeBERTa v3 xsmall + BiGRU + CRF  | 0.9634   | 0.9062    | 0.9218 | 0.9139   |

As illustrated in Fig.2 - Fig. 6, the training and validation loss curves show rapid convergence during the early epochs, followed by gradual stabilization. The gap between training loss and test loss remains minimal throughout the 15 training epochs, indicating stable optimization without significant overfitting. Additionally, the performance curves demonstrate that both F1-score and accuracy increase steadily and plateau after approximately epoch 6–8, suggesting effective learning dynamics and sufficient training duration. The hybrid architecture successfully captures contextual dependencies through bidirectional recurrent modeling, while the CRF layer enforces valid transition constraints, reducing fragmented entity predictions.

Despite the improved performance observed in Config 5, the overall architecture remains computationally efficient due to the compact DeBERTa v3 xsmall backbone. With approximately 22 million backbone parameters, the model maintains a significantly lower

parameter count compared to large-scale Transformer models, making it suitable for deployment in practical enterprise environments equipped with consumer-grade GPUs.

**Fig. 2 Loss and Model Performance of DeBERTa v3 xsmall Model**

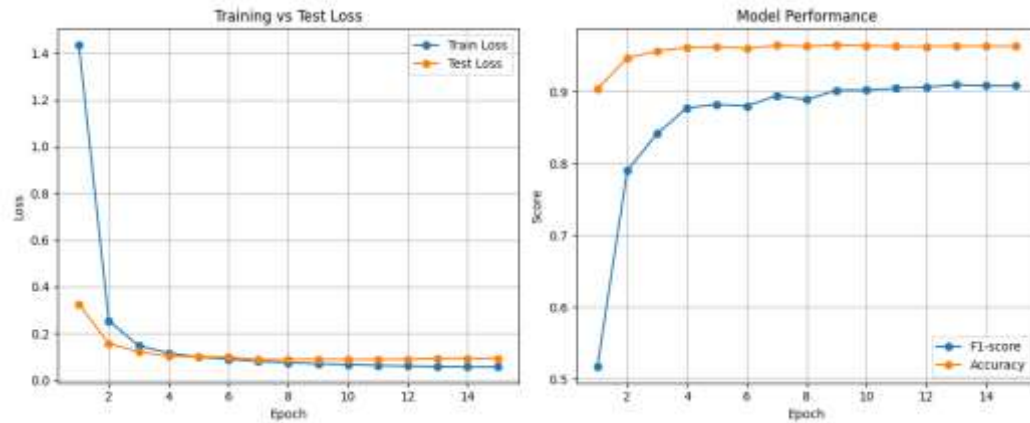


**Fig. 3 Loss and Model Performance of DeBERTa v3 xsmall + BiLSTM Model**

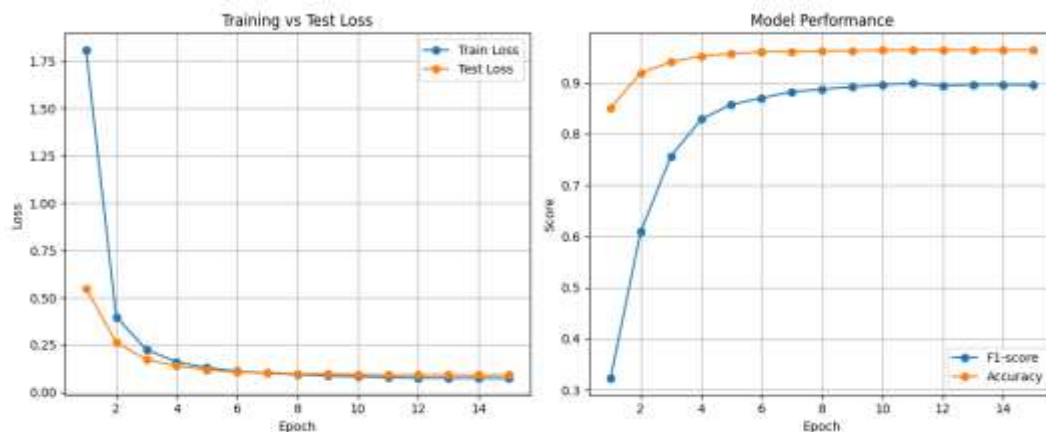


**Fig. 4 Loss and Model Performance of DeBERTa v3 xsmall + BiGRU Model**

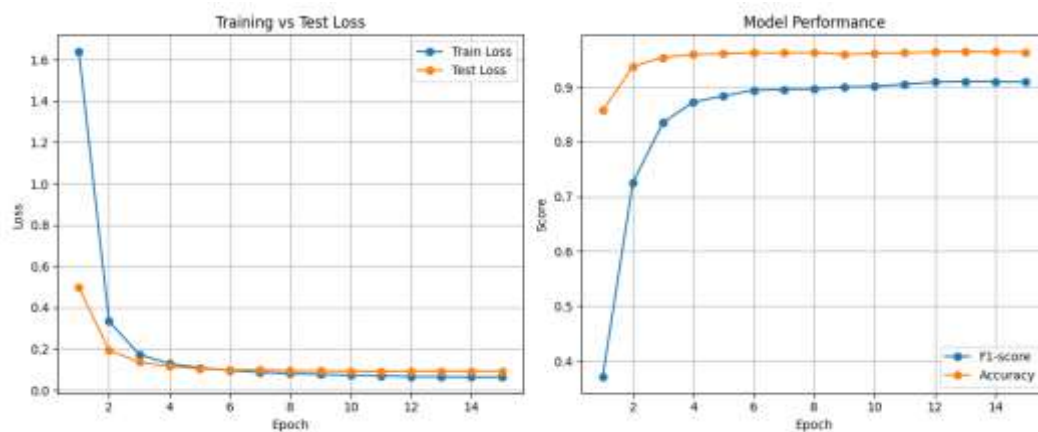
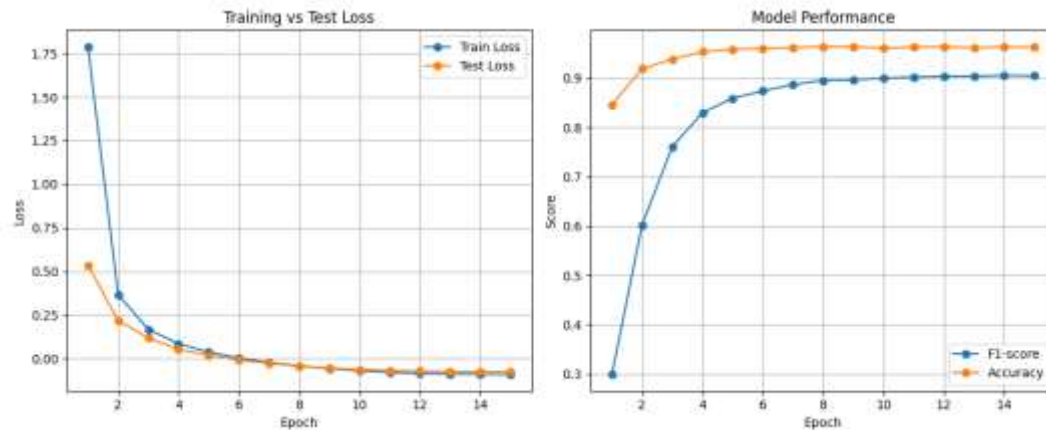
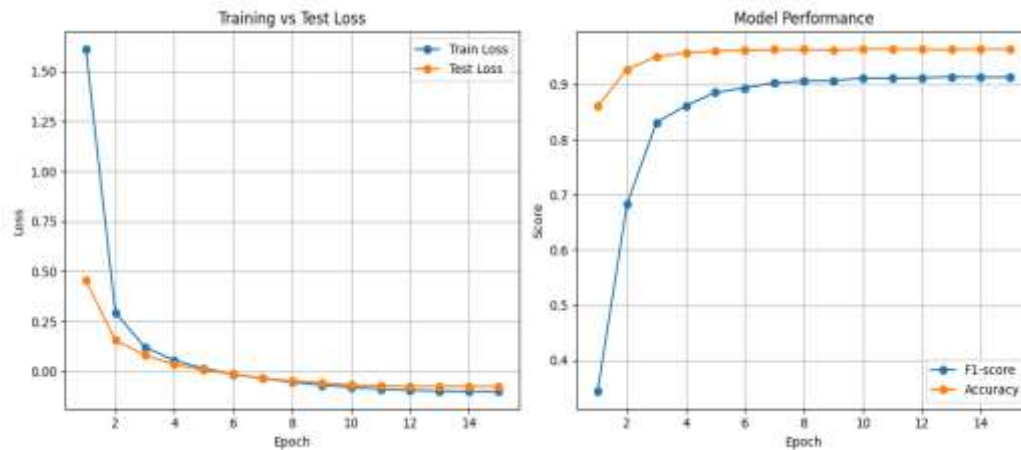


**Fig. 5 Loss and Model Performance of DeBERTa v3 xsmall + BiLSTM + CRF Model****Fig. 6 Loss and Model Performance of DeBERTa v3 xsmall + BiGRU + CRF Model**

## V. CONCLUSIONS

This study examined the effectiveness of compact Transformer-based architectures for Personally Identifiable Information Named Entity Recognition (PII-NER) tasks. Experimental results demonstrate that integrating sequential modeling through BiLSTM and structured decoding via CRF on top of DeBERTa v3 xsmall significantly enhances performance compared to the standalone backbone model. The proposed DeBERTa v3 xsmall + BiGRU + CRF architecture achieved the highest F1-score of 0.9139, confirming that compact Transformer backbones can deliver strong performance when complemented by appropriate sequence modeling layers. These findings highlight that lightweight hybrid architectures provide an optimal balance between predictive accuracy and computational efficiency, rendering them highly suitable for real-world deployment in privacy-sensitive environments.

Future research should focus on extending the model's capabilities to address broader and more practical challenges. First, multilingual expansion using mDeBERTaV3 would enable effective PII detection across diverse languages, supporting global data privacy requirements. Second, techniques for handling long sequences beyond the typical 512-token limitation of Transformer architectures should be implemented to improve robustness in processing extensive documents. Finally, knowledge distillation strategies

could be explored by leveraging the best-performing model as a teacher to transfer knowledge into smaller architectures such as TinyBERT, thereby further reducing inference time and enhancing suitability for deployment on resource-constrained edge devices.

#### **REFERENCES**

1. T. Areskou, and V. Mörk, "Detection of personal data in unstructured text using deep learning models," 2025.
2. M. D. Drovo, M. Chowdhury, S. I. Uday, and A. K. Das, "Named entity recognition in Bengali text using merged hidden Markov model and rule base approach," In 2019 7th International Conference on Smart Computing & Communications (ICSCC), pp. 1-5, June 2019.
3. W. C. Chang, C. H. Tsou, and A. Huang, "Fast rule-based NER in SpaCy for chest radiography reports with CheXpert's 14 Categories," In 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 1-4, July 2024.
4. N. A. D. Dharviyanti and N. Wilantika, "Rule-Based NER for crime information extraction through online news site," In 2024 International Conference on Information Technology Research and Innovation (ICITRI), pp. 99-104, September 2024.
5. Y. S. Chuang, C. T. Lee, R. Brandon, T. D. Tran, O. Tokede, M. F. Walji, et al., "Extracting periodontitis diagnosis in clinical notes with RoBERTa and regular expression," In 2023 IEEE 11th International Conference on Healthcare Informatics (ICHI), pp. 200-205, June 2023.
6. Z. Toh, W. Wang, M. Lan, and X. Li, "An NER-based product identification and lucene-based product linking approach to CPROD1 challenge: Description of submission system to CPROD1 Challenge," In 2012 IEEE 12th International Conference on Data Mining Workshops, pp. 869-871, December 2012.
7. I. Augenstein, L. Derczynski, and K. Bontcheva, "Generalisation in named entity recognition: A quantitative analysis," *Computer Speech & Language*, vol. 44, pp. 61-83, 2017.
8. V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter", *ArXiv*, 2019.
9. Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," *arXiv*, 2019.
10. M. N. M. M. Noor and M. F. M. Badauraudine, "Evaluating small BERT based models on automated essay scoring task," *Information Research Communications*, vol. 2, pp. 175-181, 2025.
11. Z. Sun, H. Yu, X. Song, R. Liu, Y. Yang, and D. Zhou, "Mobilebert: a compact task-agnostic bert for resource-limited devices," In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 2158-2170, July 2020.
12. X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, et al., "Tinybert: Distilling bert for natural language understanding," In *Findings of the association for computational linguistics: EMNLP 2020*, pp. 4163-4174, November 2020.
13. I. Turc, M. Chang, K. Lee, and K. Toutanova, "Well-read students learn better: On the importance of pre-training compact models," *arXiv*, 2019.
14. P. He, J. Gao, and W. Chen, "Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing," *arXiv*, 2021.
15. A. Graves, S. Fernández, and J. Schmidhuber, "Bidirectional LSTM networks for improved phoneme classification and recognition," In *International conference on artificial neural networks*, pp. 799-804, September, 2005.

16. R. Panchendrarajan, and A. Amaresan, "Bidirectional LSTM-CRF for named entity recognition," In Proceedings of the 32nd Pacific Asia conference on language, information and computation, 2018.
17. K. L. Thant, K. Nongpong, Y. K. Thu, T. Aung, K. H. Wai, and T. M. Oo, "myNER: Contextualized Burmese named entity recognition with bidirectional lstm and fastText embeddings via joint training with pos tagging," In 2025 IEEE International Conference on Cybernetics and Innovations (ICCI), pp. 1-6, April, 2025.
18. X. Liu, Y. Wang, X. Wang, H. Xu, C. Li, and X. Xin, "Bi-directional gated recurrent unit neural network based nonlinear equalizer for coherent optical communication system," Opt. Express, vol. 29, no. 4, pp. 5923–5933, February 2021.
19. L. Anand, "LSTM-and GRU-Based architectures for text sequence modeling," 2022.
20. J. Deng, L. Cheng, and Z. Wang, "Self-attention-based BiGRU and capsule network for named entity recognition," arXiv, 2020.
21. J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," In Proceedings of the Eighteenth International Conference on Machine Learning (ICML '01), pp. 282–289, 2001.
22. J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using convolutional networks," In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 648-656, 2015.