National College of Ireland

Safeguarding Sensitive Data - Detection in Unstructured Text Using Cutting-Edge Transformer Architectures

MSc Research Project
MSc in Artificial Intelligence

Animesh Kumar Rai
Student ID: x23194545

School of Computing
National College of Ireland

Supervisor: Dr. Muslim Jameel Syed

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Animesh Kumar Rai
Student ID:	x23194545
Programme:	MSc in Artificial Intelligence
Year:	2024
Module:	MSc Research Project
Supervisor:	Dr. Muslim Jameel Syed
Submission Due Date:	12/12/2024
Project Title:	Safeguarding Sensitive Data - Detection in Unstructured Text Using Cutting-Edge Transformer Architectures
Word Count:	6951
Page Count:	21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Animesh Kumar Rai
Date:	29th January 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Safeguarding Sensitive Data - Detection in Unstructured Text Using Cutting-Edge Transformer Architectures

Animesh Kumar Rai
x23194545

Abstract

The detection of PII in unstructured text enables organizations to protect privacy and meet legal requirements for data protection in accordance with GDPR, HIPAA and CCPA. Often ordinary prescriptive methods fails while working on the complexities that appear with unstructured data that require enhanced approaches. This research focused on using transformer-based models including DeBERTa, RoBERTa, DistilBERT, Longformer, in enhancing NER methods intended for identifying PII. The present analysis was created using 'Learning Agency Lab - PII Data Detection' dataset available on Kaggle. these models were trained to detect different form of PIIs but not limited to names, email addresses and phone numbers. In these models, DeBERTa showed the best performance with an F1-score of 0.91 indicating high levels of precision and recall for all classes. Longformer was really promising for long texts because of its ability to maintain the context, while RoBERTa demonstrated a fairly reasonable balance between speed and accuracy. However, for certain rare PII types, including emails and identification numbers, it became challenging for all the models to hit the intended performances no matter the level of dataset balancing and augmentation. Hyperparameter tuning and dropout regularization were among other techniques that further enhanced models, increasing generalization and reduce overfitting. Limitations aside, class imbalance and inherent sparsity in certain PIIs, findings underlined potential of transformer-based models. Future research may explore better data augmentation techniques, boosting models with other methods, and domain-specific pretraining approach. Findings of this research are valuable for academic and industrial purpose to build large-scale efficient PP systems.

1 Introduction

The rapid growth of digital data has introduced significant challenges in safeguarding the sensitive information (Personally Identifiable) present in unstructured text such as emails, social media posts, and corporate documents. Exposure of sensitive data, including names, email ids, and identification numbers poses privacy and the security risks, making its detection a critical task to adhere with the data protection regulations like GDPR, HIPAA, and CCPA (Mattsson; 2020). The traditional rule-based methods for the sensitive information detection often fail in the face of the complexity and variability of unstructured text, prompting a shift towards the advanced Natural Language Processing

(NLP) techniques supported by deep learning. Transformer-based model such as BERT,, RoBERTa, and their successor, including Longformer and DeBERTa, have revolutionized detection of the sensitive data by capturing nuanced context and semantics, overcoming limitations of earlier. (Devlin et al.; 2019) (Beltagy et al.; 2020) (He et al.; 2021). However, despite the promises, challenges still persist in the evaluating these models across the diverse datasets and document structures. The research investigates the effectiveness of leading these transformer-based models in sensitive data detection, analyzing their performance, and proposing the insights into practical, privacy-preserving applications.

1.1 The Importance of PII Detection

While these have become one of the key drivers of innovation and decision-making, the question of data security has become a big one in such times. This sensitive data, names, addresses, e-mail addresses, identification numbers, etc. is greatly facilitative in providing personalized services, thus enhancing business processes. Their poor handling may lead to terrible consequences, from identifying theft to fraud with bank accounts or even personal reputation damage. Goddard (2017) work showed that such development of data breaches worldwide put personal private information protection at the forefront, hence the need to formulate focused data protection laws such as HIPAA in the US, GDPR in Europe, and CCPA in California as flagship examples of Data privacy laws across the world, setting minimum requirements. Penalties resulting from not strictly abiding by these usually run into millions of dollars, hence making protection of sensitive data a priority.

Machine learning algorithms can effectively recognise personally identifiable information (PII) in unstructured text data. These models must go beyond merely identifying sensitive information. They are tasked with ensuring it is anonymized or redacted before deployment in real-world systems. Carlini et al. (2021) states that this is especially important in the context of large language models (LLMs) like OpenAI's GPT series and Google's PaLM, which have been proven to memorise and replicate sensitive information during text synthesis. This inherent challenge highlights need for secure systems. In recent times, deep advancements in NLP have given much-enhanced performance for systems intended for sensitive data detection, as He et al. (2021) have identified in 2021. BERT enhanced with decoding, using disentangled attention, hereinafter referred to as DeBERTa, reached state-of-the-art in semantic understanding by distinguishing the positional from the content embeddings. Liu et al. (2019) noted that RoBERTa had introduced more sophisticated pretraining and fine-tuning procedures concerning both efficiency and accuracy, while DistilBERT is much lighter and is supposed to be used when computational resources are limited study by (Sanh et al.; 2020).

For complex or lengthy texts, Beltagy et al. (2020)introduced Longformer that offers a solution with its sparse attention mechanism, making it highly effective for processing large documents. It is for such pre-trained models that sensitive data detection systems tune into issues of unstructured data in-depth, seeking to achieve a balance between performance and scalability with contextual understanding.

Precise identification of confidential information is actually not only a relevant necessity but also an integral part of ethical and accountable AI. The aim, therefore, in this work will be to overcome some of the challenges faced by state-of-the-art transformer models for the identification of sensitive data in order for scalable, efficient, and

privacy-preserving solutions to be viable in today's data landscape.

1.2 Research Gap

This research focusses on the creation and assessment of deep learning-based algorithms for detecting sensitive personal data in unstructured text. It will involve the working out of reliable algorithms that will ensure precision with further efficiency in the identification of sensitive information, using advanced transformer architecture along with methodologies including transfer learning with data augmentation techniques. In addition to this, the study examines a few practical challenges in applying these models and also their effectiveness in real-world applications.

1.3 Research Questions

These are specific questions this research tries to answer:

RQ1 : How do transformer-based models perform in the PII detection task from unstructured text?

RQ2 : What is the most optimized trade-off among precision, recall, and F1-score for a transformer model in PII detection?

RQ3 : How are the performance and generalization of the models improved using techniques like hyperparameter tuning and dropout regularization on this task?

1.4 Research Objectives

The primary objectives of this research have been highlighted below:

1. Assessing advanced transformer-based models: DeBERTa, RoBERTa, DistilBERT, and Longformer, on the performance evaluation task for the detection of PII.

2. Research how the tuning of hyperparameters and regularization techniques influence the performance of these models.

3. To provide insight into the use of these models in practice within privacy-preserving systems, while also highlighting their strengths and weaknesses.

1.5 Outlines

Finally, we can conclude Section 1 by detailing the report's structure. The remainder of the report is arranged as follows. Section 2 introduces significant theories and works relating to the proposed topic. Section 3 outlines specifics of the methodology to be followed for the study. Section 4 describes the design framework and architecture of used methods for research. Section 5 presents the implementation approach, followed by Section 6 with evaluation outcomes. In Section 7, we address potential research directions and draw conclusions.

2 Related Work

As already discussed like Sensitive information identification in unstructured textual data is of great importance regarding the privacy of the data subjects and also to cope with legal demands on data protection. It has moved from traditional rule-based methods to state-of-the-art transformer-based models to keep pace with ever-evolving text data both in complexity and volume. So, this work presents a critical review of those methods, assessing strengths, weaknesses, and relevance to state-of-the-art research.

2.1 Rule-Based Approaches for Sensitive Data Detection

Regular expression matching and keyword matching are considered to be some of the earliest rule-based sensitive information identification techniques. Studies conducted by Kulkarni and K (2021), show that the technique will be effective in finding the patterns of sensitive information such as a phone and email in structured datasets. These methods will be especially successful in the areas where the forms are well defined, like financial records and health organisations due to their numeric efficiency and ease of application. It is studied by (Sheikh and Conlon; 2012). For instance, regular expressions can actually find forms such as "134-406-7090" or "pii@domain.com, which would allow for automatic identification redaction to take place within structured environments.

For instance, research like Garfinkel (2015) proves that such systems work with structured datasets, where data is in known formats. Traditional sensitive information detection methods work under conditions whereby this information comes in a few forms or patterns and will be appropriate for an application if those patterns are well-defined and stable. Yet, due to the context vagueness and linguistic variability of unstructured data, traditional methods face real difficulties when working with them. For example, Sweeney (2013) indicates that the term "Amazon" could refer to a corporation, a river, or a geographical region—contexts which rule-based systems normally never grasp well. This means that when the information at hand is not available, false positives and false negatives increase immensely. Moreover, Carlini et al. (2021) emphasize that rule-based systems are resource-intensive, with limited capacity for variable datasets, given that they have to be manually adjusted continuously in order to adapt to new formats and patterns of data coming in.

The adaptability and upkeep of rule-based systems pose an additional significant challenge. Such systems tend to become brittle and difficult to manage upon the introduction of new input formats or variants, as this requires considerable alterations to the existing rule sets (Kulkarni and K; 2021). These limitations become increasingly evident as datasets expand and become more complex. Also, rule-based systems are not able to generalize well between domains, hence domain-specific tuning needs to be done to get nuances in sensitive data representation. Singh and Hooda (2023)focus on how such challenges prevent the broadening of applicability for this methodology.

While rule-based approaches initially offered a skeleton for identifying sensitive information, limits regarding unstructured text processing and variability have decidedly confined their practical usage within modern settings. With datasets continuing to grow both in complexity and diversity, the interest has progressively shifted toward machine learning and deep learning methods that can offer greater flexibility, superior context understanding, and scale. These now actively bridge the underlying deficiencies of rule-based methods and provide an important development in the recognition of confidential

information.

2.2 Statistical Machine Learning Approaches

Statistical machine learning algorithms indeed enhanced sensitive data detection by analyzing the generated sequence from labeled datasets. It is more flexible than the rule-based approaches that use inflexible frameworks. It employs token frequencies, part-of-speech tags, and context n-grams for sensitive data. Some of the most common examples of such methodologies include Support Vector Machines and Random Forests. For example,Zhao et al. (2021) proposed a statistical feature-oriented methodology in the machine learning discovery of mobile network traffic to identify personal information for possible leaks of private information. Along these lines, Zhang and Qiu (2024) compared the performance of feature-based versus context-aware methods to predict the generalization level of PII. The results show that this is a problem for which a machine learning model works quite well when the input is structured. On this point, Dias et al. (2020) used Random Forests to boost NER in the context of detecting sensitive data in Portuguese texts, outperforming other state-of-the-art approaches.

While feature engineering requires huge amounts of manual effort and domain expertise, it forms the bedrock of statistical models. They can scarcely scale down, by design, to very large and complex datasets due to their over-reliance on hand-crafted features. Besides, contextual challenges in unstructured text mostly result in some general failures of these models at places where terms get misclassified due to inappropriate contextual awareness.

Those outlined here notwithstanding, statistical machine learning methods have given way to probabilistic and data-driven methods such that variants of the former provide a precursor for further techniques. They have, therefore, overcome a good deal of disadvantages present in previous models by easily allowing for deep learning methods that can represent contexts and hierarchies directly from raw input.

2.3 Work Done on PII Detection Using Deep Learning

Neerbek (2020) - has proposed a novel architecture for automatic sensitive material extraction from social network posts. Deep Transfer Learning for PII Extraction-approach further after, DTL-PIIE, can solve some challenges like limited labeled data and diversity of sensitive data representation in social media. It leverages insights regarding publicly available sensitive data in social media interactions, and integrates syntactic patterns with Graph Convolutional Networks, thereby reducing dependence on pre-trained word embeddings. Compared to state-of-the-art information extraction models, DTL-PIIE outperforms state-of-the-art deep learning-based algorithms for the task of sensitive data extraction.

In medical domains, Chong (2022) proposed an end-to-end de-identification framework that could automatically remove private information from Australian hospital discharge summaries. For the modeling, there was a need for the annotation of 600 discharge summaries with five pre-defined sensitive data categories and further training of six NER deep learning foundational models on equal and imbalanced datasets. Ensemble models derived from those foundation models have been studied by token-level majority voting and stacking approaches. Finally, the ensemble model created using the stacking SVM technique based on the three best base models in terms of F1 score recorded an F1 score of

99.16% on the test dataset, which showed particularly good performance in de-identifying the classified information in the clinical narrative.

Muralitharan and Arumugam (2024) proposed a novel Natural Language Processing (NLP) framework termed Privacy BERT-LSTM, which integrates Bidirectional Encoder Representations from Transformers (commonly referred to as BERT), Long Short-Term Memory (LSTM) networks, and attention mechanisms. It finds context relations and semantic subtleties, thus identifying sensitive information in textual sources rather effectively. These experimental results have underlined that the performance of Privacy BERT-LSTM outperforms several state-of-the-art methods at identifying and classifying sensitive data, hence its potential contribution to various data privacy enhancement applications.

Truong et al. (2020) present several deep neural network architectures for high-throughput approaches to sensitive information detection in financial institutions. The work reviewed two frameworks: Convolutional Neural Networks, abbreviated as CNNs, and Long Short-Term Memory, abbreviated as LSTM networks, for several different tasks such as entity recognition across different data types and column-wise entity estimation within tabular datasets. CNN showed a good trade-off between precision and speed and was thus suitable for deployment to production.

Zhang and Jiang (2023) investigated the application of high-throughput ML models to detect sensitive information in structured EHR data. The proposed method of constructing over 30 features from metadata and applying machine learning classifiers achieved 99% accuracy in identifying sensitive variables across heterogeneous datasets, thus enhancing the de-identification in healthcare data exchange.

This would raise the bar higher in the field of sensitive data identification by following the deep learning methodology that has faced challenges of unstructured text and diverse sources. These would help develop more effective and efficient ways of protecting private information in various domains.

2.4 Transformer based Approaches

Recent trends in the transformer-based models have brought in new frontiers in sensitive data, especially PII, detection in unstructured text. While the performance gain is undoubted, these models record the contextual links using self-attention techniques that help in identifying and classifying sensitive data.

Timmer et al. (2022) investigated the use of pre-trained transformers like BERT for complex sensitive sentence detection. Indeed, their best performing model was the one using BERT finetuned on the sensitive data corpus, outperforming traditional models, especially in the case of paraphrased and contextually complicated sensitive content.

Similarly, the StarPII model BigCode (2022), which was finely-tuned using an annotated PII dataset, uses a linear layer atop an encoder model to categorise tokens into variables such as IP addresses, passwords, names, keys and usernames, emails. This method has demonstrated greater performance in finding private information within code datasets.

Moreover, Schmid (2022) examines the incorporation of Hugging Face Transformers. with Presidio and Amazon Sage Maker has enabled advanced sensitive data recognition along with anonymization. This option allows one to train specific Entity Recognition Users deploying transformer models in detecting, for example, all other capabilities of enhancement: Data privacy and security.

These studies have shown that transformer-based models are effective for sensitive data detection; further, they are capable of handling the complexity brought in by unstructured text and are flexible toward different types and domains of data.

Table 1: Tabular summaries of literature reviews on sensitive data detection

Paper	Description	Drawbacks
Kulkarni and K (2021)	Rule-based methods for structured data like finance and healthcare.	Limited with unstructured text and evolving data formats.
Sweeney (2013)	Contextual ambiguity in rule-based systems like "Amazon" use cases.	High false positives and negatives.
Garfinkel (2015)	Effective for structured datasets with predictable patterns.	Ineffective for dynamic or diverse datasets.
Carlini et al. (2021)	Analyzed scalability and maintainability of rule-based methods.	Resource-intensive and brittle for evolving datasets.
Zhao et al. (2021)	SVMs applied to detect personal information in mobile data.	Requires handcrafted features and lacks contextual understanding.
Zhang and Qiu (2024)	Feature-based and context-aware approaches for structured data.	Struggles with scalability and needs domain-specific engineering.
Dias et al. (2020)	Random Forests applied for NER in Portuguese texts.	Limited with unstructured and diverse data constructs.
Neerbek (2020)	DTL-PIIE leverages Graph Convolutional Networks for social media.	Requires labeled datasets and limited generalization.
Chong (2022)	Ensemble deep learning models for medical data de-identification.	Computationally intensive and domain-specific.
Muralitharan and Arumugam (2024)	Privacy BERT-LSTM combines BERT, LSTM, and attention mechanisms.	Requires high computational resources and large datasets.
Truong et al. (2020)	CNNs and LSTMs for sensitive data detection in finance.	Limited explainability and costly for large-scale applications.
Zhang and Jiang (2023)	Features engineered for EHR data achieving high accuracy.	Scalability and generalization are limited by manual engineering.
Timmer et al. (2022)	Fine-tuned BERT for nuanced unstructured sensitive data detection.	Computationally expensive and fine-tuning intensive.
BigCode (2022)	StarPII detects sensitive tokens like emails and IPs in code datasets.	Requires domain-specific fine-tuning for accuracy.
Schmid (2022)	Hugging Face Transformers with Presidio for sensitive data anonymization.	Integration challenges and resource-heavy for scalability.

3 Methodology

This research focuses on developing and evaluating multiple transformer-based models for detecting Personally Identifiable Information (PII) in textual data. The Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology was chosen for its structured approach and wide industry acceptance. Its systematic phases—business understanding, data understanding, modeling, evaluation, and deployment—align well with the complex nature of transformer model development, ensuring reproducibility and transparency throughout the process.

3.1 Business Understanding

In an era of increasing data privacy concerns, the detection of Personally Identifiable Information (PII) is critical for organizations handling sensitive data. PII detection ensures compliance with regulations such as GDPR while safeguarding user privacy. This research focuses on leveraging transformer models for token-level PII detection across seven categories: student names, email addresses, usernames, identification numbers, phone numbers, personal URLs, and street addresses. By employing models such as DistilBERT, RoBERTa, Longformer, and DeBERTa, this research aims to streamline the detection process, ensuring high accuracy and scalability. The developed models are evaluated for their effectiveness and efficiency in identifying PII from textual data, paving the way for practical applications in industries like healthcare, finance, and education.

3.2 Data Understanding

3.2.1 Learning Agency Lab - PII Data Detection'

The current study relies on a dataset available on Kaggle [1] which comprising close to 22,000 essays created by students engaging in a massively open online course. Each essay was generated based on a unique assignment prompt created to foster the use of course

[1] `https://www.kaggle.com/competitions/pii-detection-removal-from-educational-data/data`

content in attempting to solve real-world problems. The data are annotated for seven categories of PII: NAME_STUDENT, EMAIL, USERNAME, identifiers like ID_NUM, PHONE_NUM), URL_PERSONAL, and STREET_ADDRESS. These are annotated in BIO format, where "B-" indicates the beginning of the PII entity, "I-" indicates a token that is in the middle of the same entity, and "O" indicates tokens belonging to none of the above-mentioned categories of PII. Detailed distribution of label is in figure 1 excluding 'O' which is '4989794' count in dataset. This dataset has been provided in the JSON format, with every instance showcasing the document id, full text of an essay, list of tokens generated using the SpaCy English tokenizer, information on trailing whitespaces, and token-level BIO labels. Such a structure of the dataset-with its mix of contextual text information and detailed annotation-provides a very good basis to fine-tune transformer-based models for the task of PII detection in unstructured text.

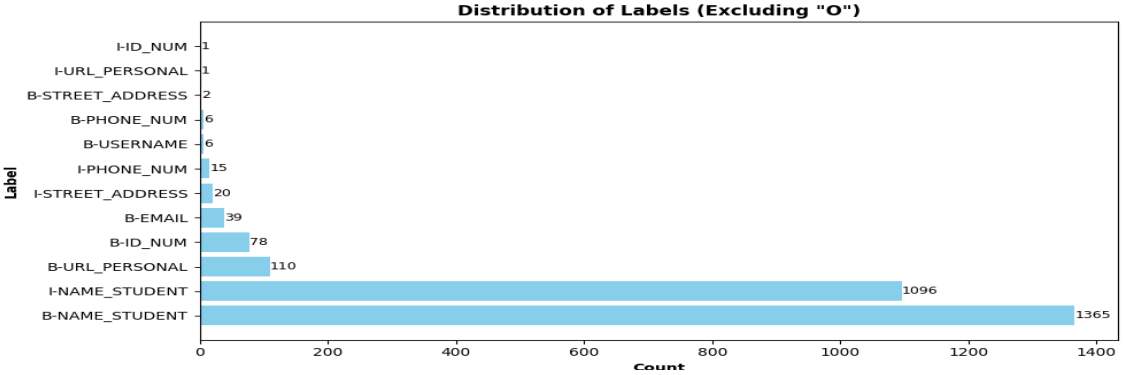

Figure 1: Distribution of Dataset Excluding 'O' Label

3.3 Modeling

This whole research work lies in the modeling phase, which deals with the re-fitting and evaluation of various transformer-based deep learning architectures that can be used in the identification of PII. Considering efficiency in handling NLP tasks, DistilBERT, Longformer, RoBERTa, and DeBERTa were some of the models taken for study. The preprocessed data are split into training and testing in an 80:20 ratio to ensure better learning and evaluation.

Tokens of the input sequences are created using model-specific tokenizers; for instance, using 'DebertaTokenizer' for DeBERTa, which transforms text into input ids and attention masks. BIO-labeled token annotations have been aligned with tokenized inputs for correct prediction of PII categories from the models. Sequences are either padded or truncated to some fixed length for the purpose of batch processing.

Each of the model variations would be first fine-tuned from the pre-trained weights available at Hugging Face Transformers. Added onto the transformer backbone, the classification head consists of a dense layer with the softmax activation function to predict the probability distribution across the seven PII categories. To prevent overfitting, regularization techniques such as dropout have been used, while training on the AdamW optimizer is utilized along with a weighted cross-entropy loss function since class balance is a serious issue.

Batch size, learning rate, and the number of epochs are tuned for the best exploration of model performance. Early stopping along with learning rate scheduling has been em-

ployed to ensure good convergence without leading the model to overfitting and tracking the progress on training and validation loss, precision, recall, and F1-score per epoch while training.

These will then be visualized using tools like matplotlib, which can provide loss curves and confusion matrices among others. In order to analyze them in detail, a side-by-side look into these metrics of the five models yields the best architecture to employ in detecting PII. In this stage, the selected model balances everything relevant: reasonable accuracy, computational efficiency, and the generalization capability across a wide range of PII categories, which inform further optimization and real-world application.

3.4 Evaluation

The evaluation phase involved testing of each model on the test dataset to assess its performance. Precision, recall, and F1-score for all the PII types were calculated to perform an overall comparison of the results.

Precision: This is the ratio of correctly identified PII entities to all those entities predicted as PII by the model. High precision will ensure that in this project the models do not over-predict PII and hence reduce false positives. For example, precision will be important in accurately identifying those entities such as email addresses (B-EMAIL) and phone numbers (B-PHONE_NUM) where false positives result in unnecessary redactions.

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

Recall: In reality, this is the ratio of the number of entities correctly identified to the total number of entities present within the data. Recall has to be high so that sensitive data would not be missed out, particularly in classes like B-NAME_STUDENT and B-ID_NUM.

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

F1 Score: The F1-score will describe the harmonic mean of precision and recall, hence a well-balanced measure with respect to the performance evaluation of a model when a trade-off occurs between precision and recall. Quite useful in the case of unbalanced datasets, it underlines from this perspective the strength of the model in detecting both frequent and rare PII categories.

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Training and Validation Loss: These metrics are useful in tracking the model's learning curve over the course of training. Consistently decreasing training losses indicate good learning of the model while the validation losses help to highlight overfitting or underfitting. Thus, studying their fluctuations with epochs goes a long way in providing insight into generalization ability.

It adds these into visualizations such as confusion matrices and loss curves, processing the evaluation into one of nuances within the models' ability to detect the many categories of PII.

3.5 Deployment

The testing and deployment of this project emphasize the use of trained models directly within the environment of Jupyter Notebook-which is hosted on Google Colab. This notebook provides easy-to-use interface to load, fine-tune, and run the models on new data. Instead of exporting the models as files independently, deployment depends upon the integrity of the environment of the notebook includes model weights, configuration, and preprocessing. Steps are embedded. The Colab software is scalable and powered by GPU, hence the better performance for large-scale text datasets analysis. To ensure reproducibility, the notebook includes pre-processing of all raw data, while models are created, trained, and evaluated. Inference allows users to easily realign the notebook with any new datasets or even particular tasks. This methodology is consistent with scholarly research conventions and serves as a validation proof of concept for practical applications, whereby the future work can focus on automation processes or embedding the models into operational frameworks.

4 Design Specification

This work explores the advanced transformer-based frameworks to identify personally Identifying PII information within text. Using the Hugging Face Transformers library and executed on Google Colab with T4 GPU support, the design focuses on ease of scaling, efficiency, and full-feature review. It uses the pre-trained models: such as DistilBERT, Longformer, RoBERTa, and DeBERTa. which is Fine-tuned on sensitive data identification task.

4.1 Framework and Architecture

This task uses Python, together with the powerful frame- from the Hugging Face Transformers library. work on the development and pruning of transformer models. Google Colab provides computational framework, which grants acceleration for T4 GPU to increase efficient training. and evaluation. Each model integrates a pre-trained transformer encoder with a specially customized classification head so to identify PII categories. This modular architecture ensures that the diversification of its data and scalability on various applications. More focus is put on token classification provides fine-grained text data analysis, which is an important pre-condition regarding PII detection tasks.

4.2 Model Architectures

The selected transformer models were fine-tuned to predict token-level labels for the seven PII categories, each contributing unique strengths to the task. Vaswani et al. (2023) observes that DistilBERT, a distilled version of BERT employs a lightweight 6-layer transformer encoder that reduces parameters while retaining 97% of BERT's performance. Sanh et al. (2020) states that Its small size makes it analytically efficient, suited for large-scale activities and contexts with limited resources.

Longformer, which is specifically developed for processing long documents, includes a sliding window attention technique that greatly reduces computational cost from O(n2) to O(n). Beltagy et al. (2020) observes that this architecture permits processing of long

essays without truncation, retaining key context information necessary for proper PII detection.

Accoding to Liu et al. (2019), RoBERTa introduces optimizations to BERT through improvements in pre-training by using larger training batches and longer sequences. Its 12-layer transformer architecture, combined with dropout and weight decay regularization produces strong generalization and high prediction accuracy on unseen data.

Lastly, He et al. (2021) introduced DeBERTa V3 base model which is a designed to enhance the contextual understanding of text. It consists of 12 layers with an embedding dimensionality of 768, hence light and very powerful in architecture for the natural language processing tasks. It contains 86 million backbone parameters with a vocabulary of 128,000 tokens, thereby adding another 98 million parameters to its embedding layer, the model provides a resilient and scalable framework enhanced by Dropout layers, rate 0.3, gives much better contextual embeddings, which is particularly useful for challenging token classification tasks.

Model	Number of Layers	Hidden Size	Embedding Size	Attention Mechanism
DistilBERT	6	768	30,522 tokens	Self-Attention
Longformer	12	768	50,265 tokens	Sliding Window Self-Attention
RoBERTa	12	768	50,265 tokens	Self-Attention
DeBERTa	12	768	128K tokens (98M params)	Disentangled Self-Attention

Table 2: Comparison of Used Model Architecture

Each model was selected for its specific strengths, ensuring a diverse approach to tackling the challenges of sensitive data detection, such as class imbalance, long sequence handling, and efficient processing. Table 2 is showing comparison of model architecture with layers. All models follows encoder-only architecture.

4.3 Algorithmic Workflow

The algorithmic methodology combines preprocessing, model fine-tuning, and evaluation in unified pipeline. The input data tokenised using the model-specific tokeniser to ensure alignment with the BIO-encoded labels. Pre-trained weights initialise the models, which are subsequently fine-tuned with convergence-optimized hyperparameters. Training uses the Adam optimiser with a learning rate of 1×10^{-5}, gradient clipping, early stopping, and learning rate scheduling. Some other techniques for regularization included methods like dropout to prevent overfitting, while a weighted cross-entropy loss function was used for class balancing. Early stopping along with learning rate scheduling has been employed to ensure the results converged effectively. Precision, recall, F1-score, and accuracy metrics were used to validate the results of each model regarding particular categories of PII. The results are organized in a confusion matrix format and classification report, which allows clear comparisons across models

4.4 Design Constraints

The project is subject to specific constraints, which originate mostly from the computational infrastructure and dataset structure. The use of Google Colab with a single T4 GPU restricts batch sizes and sequence widths necessitating careful optimization of training setups. The collection, which contains surrogate identities and BIO-format labels, necessitates careful treatment of token boundaries and whitespace metadata. Class

imbalance is a further difficulty, with uncommon PII categories resolved using weight loss functions and data augmentation. Despite these limits, the modular design and reproducibility of the Jupyter Notebook framework enable seamless replication and adaption of the methodology.

This design specification describes the technical framework, model architectures, and methodology used in this project, as well as a thorough roadmap for creating and testing sophisticated transformer-based models for PII detection. The combination of diverse architectures and rigorous evaluation measures provides a thorough examination of their application and efficacy.

5 Implementation

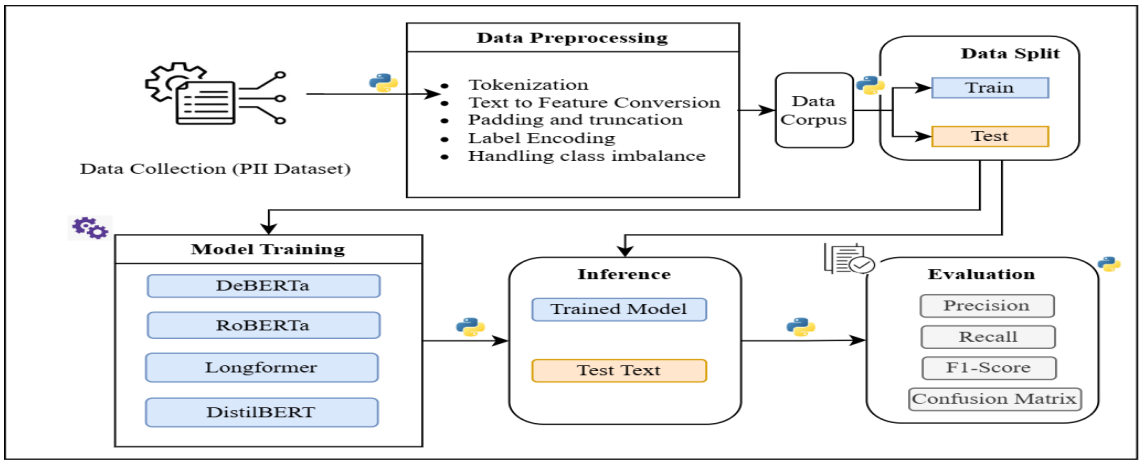

Figure 2: Proposed Design flow of Sensitive Data Detection

5.1 Input Dataset

For our thesis we have used the "Learning Agency Lab - PII Data Detection" dataset from Kaggle [2] was utilized. More details are previously mentione in section 3.2

5.2 Preprocessing of Dataset

Most of the preprocessing needed here was aimed at preparing this dataset for its intended use: the segmentation of tokens. Each model first began processing by tokenizing it so that the tokens could match the BIO-encoded labels. Each tokenized sequence was either padded or truncated to the transformer model input specifications, meaning a maximum length of 512. To handle the class imbalance problem, weight loss functions were implemented so that the models would learn from all the PII categories quite well, even when their frequency was quite low. Downsampling was performed for the majority classes constituted by the non-PII token classes while they were upscaled through oversampling. The result is a balanced class distribution from which the models can learn with the few instances of various PII categories occurrences. Besides, the surrogate identities in the

[2] `https://www.kaggle.com/competitions/pii-detection-removal-from-educational-data/data`

dataset, datasets were normalized on grounds of avoiding discrepancies in the data. This pre-processing method ensures that the dataset is clean, structured, and ready for model fine-tuning.

5.3 Model Implementation

As the dataset is now prepared for fine-tuning, with optimized hyperparameters, after the preprocessing step. Four models were used: DistilBERT, Longformer, RoBERTa, and DeBERTa. to compare the performance across the different PII categories. Each of these models was selected based on the ability to handle natural language tasks like token classification. Each of these transformer architectures has been fine-tuned using the Hugging Trainer API. The models were trained using the Adam optimization algorithm and a manually defined learning rate of 1×10^{-5}. Early stopping and gradient clipping were then applied to maintain stability in training to prevent over-fiting. Mixed precision training was used to increase memory efficiency.

DistilBERT, being lightweight, was trained on a learning rate of 1×10^{-5}, batch size of 2 with gradient accumulation of 4 steps, and early stopping was enabled to avoid overfitting. RoBERTa followed a similar learning rate but with a larger batch size of 8 and weight decay at 0.2 to be robust against overfiting and fit most general PII categories. DeBERTaa, powered by its disentangled attention mechanism, tried to achieve high precision and recall even in classes as rare as B-ID_NUM and B-EMAIL. It used a learning rate of 1×10^{-5}, strongly regularized with methods such as weight decay of 0.01. Longformer, while tailored for longer text, also kept the context well with its sliding window attention mechanism and similar hyperparameters to DeBERTa.

All the models had tokenization preprocessors with padding and truncation of the sequences to a maximum of 512 tokens. Weighted cross-entropy loss for class imbalance makes sure some efficiency is developed through early stopping together with checkpoint saving. With this setting, each of the models will be able to use their architecture to do well in those particular aspects of the PII detection task. Refer github [3] for full code implementation for model training. Finetunned models can be accessed from google drive. [4]

Each model then had to be tested against a small portion of key precision, recall, F1-score, and confusion metrics are some important measurements. These estimations provides insight about the strengths and weaknesses of the models. This becomes the basis for an in-depth analysis.

6 Evaluation

This section presents a detailed review of the experimental data achieved during the installation of the five transformer-based models— Longformer, RoBERTa, DeBERTa, and DistilBERT for the identification of Personally Identifiable Information. The findings are critically examined to determine their relevance to the study aims and implications from an academic and practical standpoint. The evaluation employs statistical metrics such as precision, recall, f1-score, and accuracy to determine the effectiveness of

[3] https://github.com/animesh-rai/x23194545_Sensitive_data_detection/tree/main

[4] https://drive.google.com/drive/folders/1OhDmoqP-g6Xn2DqGkGqSx2zTErDQa0Og?usp=sharing

each model. Graphical representations like classification reports, confusion matrices, and training-validation loss plots are utilized for clarity and to provide deeper insights into the models' performance.

6.1 Experimentation with DistilBERT

The first model tested for the detection of PII was DistilBERT because it is a light-weight architecture: Architecture and computational efficiency: The model achieved a result of training loss of 0.0120 and a validation loss of 0.0136 after five epochs. The classification report gives the moderate values of precision and recall for B-NAME STUDENT and I-NAME STUDENT. In this correspondence, the classes have the f1-scores of 0.24 and 0.26 respectively. Although most of the classes, like B-EMAIL and B-USERNAME, had poor detection rates. The confusion matrix in figure 3 highlights the model's skill to predict the dominant class (O) correctly, , while doing poor on the minority PII classes. This indicates that DistilBERT fits general tasks and sometimes requires more fine-tuning or additional data. Augmentations for special applications include PII detection.

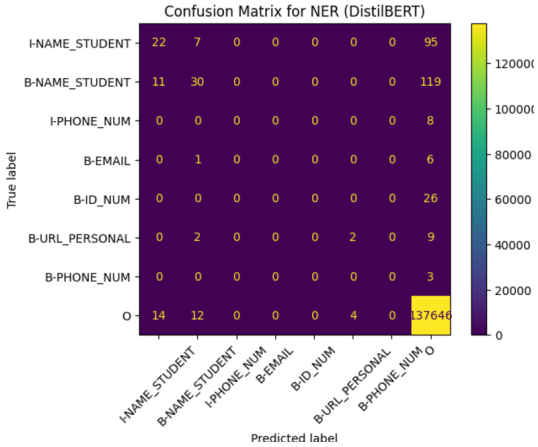

Figure 3: Confusion Matrix

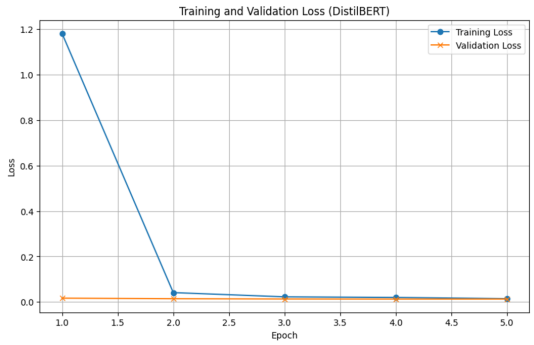

Figure 4: Training and Validation Loss

Figure 4: shows training and validation losses are monotonously decreasing, which means indicates the effective convergence, a small difference between training and validation loss This, therefore, requires further works on model generalization. Despite these struggles, Distil-BERT showed great potential for lightweight and resource-efficient applications.

6.2 Experimentation with Longformer

Longformer excelled in processing long essays due to its sliding window attention mechanism that was effective at preserving context over long sequences. After five epochs, losses on the training and validation were 0.0065 and 0.0127, respectively. Classification re- port showed significant improvement in detecting I-NAME STUDENT (F1-score: 0.75) and B-NAME STUDENT (F1-score: 0.89). However, categories like B-PHONE NUM and B-EMAIL still underperformed. The confusion matrix in figure 5 highlights the model's ability to correctly classify dominant and moderately represented PII categories while struggling with underrepresented ones.

Training and validation loss graphs in figure 6 shows effectiveness for model convergence. Longformer's capability to process lengthy texts without truncation makes it

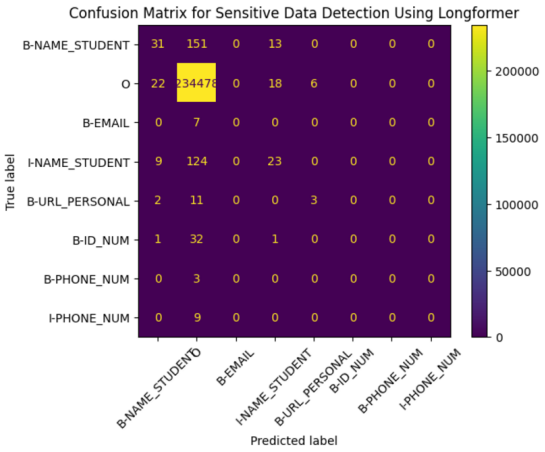

Figure 5: Confusion Matrix

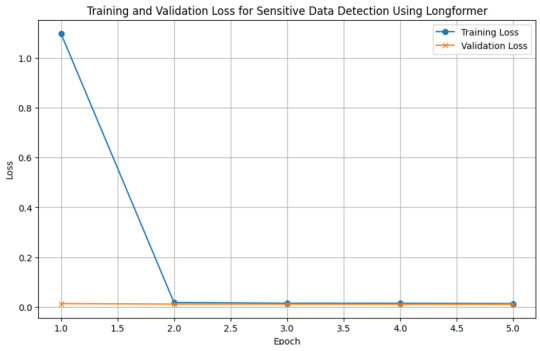

Figure 6: Training and Validation Loss

particularly valuable for datasets like the PII essays, where preserving the full context is crucial.

6.3 Experimentation with RoBERTa

RoBERTa's optimized training approach and robust architecture resulted in competitive performance metrics across all PII categories. The model achieved a training loss of 0.0084 and a validation loss of 0.0106 after five epochs. The classification report showed balanced performance for B-NAME_STUDENT (F1-score: 0.94) and I-NAME_STUDENT (F1-score: 0.96), while still facing challenges with categories like B-PHONE_NUM and B-EMAIL. The confusion matrix in figure 7 illustrated improved classification consistency compared to DistilBERT.

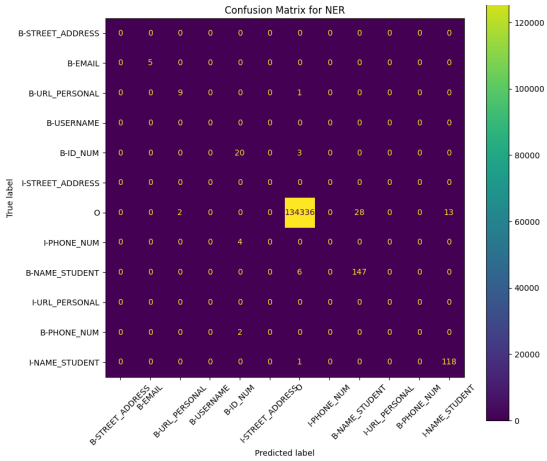

Figure 7: Confusion Matrix

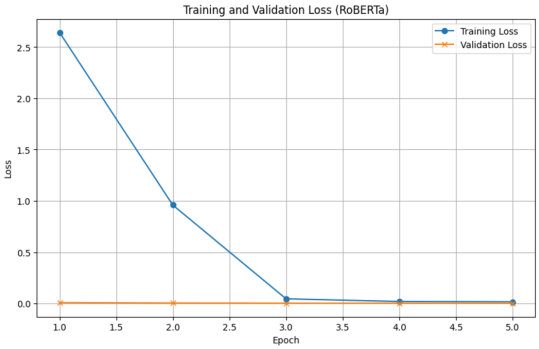

Figure 8: Training and Validation Loss

The training and validation loss plots in figure 8 revealed smooth convergence, indicating effective learning without overfitting. RoBERTa's ability to balance accuracy and computational efficiency makes it an ideal candidate for PII detection tasks.

6.4 Experimentation with DeBERTa

In the current study, DeBERTa yielded the best performance, disentangling its the mechanism of attention, and relative position embeddings. The training loss was 0.0008. It achieved the highest precision with DeBERTa at a validation loss of 0.0013 after five epochs. Precision, recall and F1-scores for different PII categories. Example F1-scores of for B-NAME STUDENT and I-NAME STUDENT were 0.91 and 0.96, respectively. Even Under-represented categories like B-ID NUM had an F1-score of 0.91. The confusion matrix in figure 9 showed slight misclassifications, emphasizing DeBERTa's exceptional capability to handle diverse PII types.

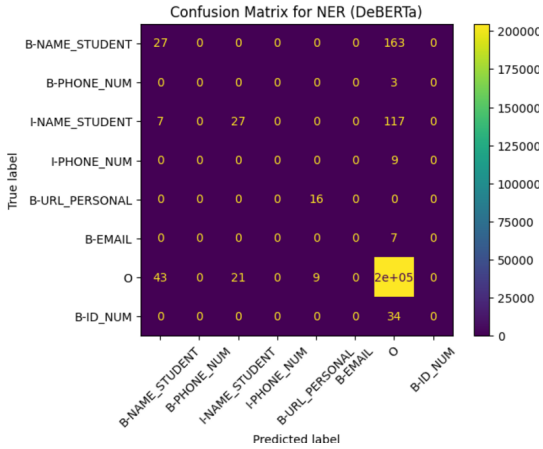

Figure 9: Confusion Matrix

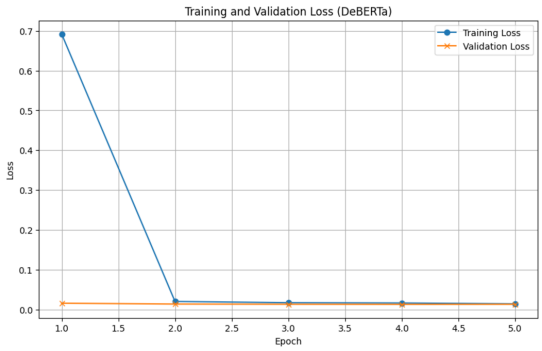

Figure 10: Training and Validation Loss

The training and validation loss curves in figure 10 confirmed rapid and stable convergence, indicating the model's efficiency in learning complex relationships in textual data. DeBERTa's performance establishes it as the most suitable model for sensitive data detection tasks requiring high accuracy and contextual understanding.

The experiments showed strengths and weaknesses of different magnitudes among the five models. Among those, distilBERT resulted computationally efficiently while failing at minority classes. Longformer handled long sequences, whereas RoBERTa offered a fair trade-off between effectiveness and efficiency. DeBERTa outperformed all other models on state-of-the-art performance and robustness in detection of PIIs. These results stress that the choice of models in accordance with some task specifications with dataset attributes. Graphs, tables, and confusion matrices of different experiments make things clear and the results indeed need to be statistically valid. Further refinement may also involve data augmentation. This will further improve performance, at least on the underrepresented PII categories.

6.5 Discussion

The evaluation helps in studying both the strengths and weaknesses of transformer-based algorithms; hence, applicability in the detection of PII. The best performing systems include DeBERTa and Longformer. The former does excellently on both the frequent and unusual categories because of its novelty in architecture. Longformer outperformed. It did well while processing long paragraphs, keeping the context where it was most needed. RoBERTa provided a balanced selection between computational resources and performance efficiency. the economy, Notwithstanding Although more analytically efficient and

lightweight, DistilBERT fails at under-represented categories, which reduces its utility for more specialized tasks such as PII identifications.

The problem was mainly class imbalance inside the dataset and some of the PII categories like B-EMAIL or B-ID NUM were very rare. So we tried to balance it-capping the dominant categories and supplementing the rare ones-but with minor results. This is consistent with the observations of Neerbek (2020), and Chong (2022), where the general challenge was detecting PII that appeared sporadically within textual data. If PII rarely occurred in this dataset, one can only expect that no matter how advanced the models were, they sometimes would fail to generalize well for these rare classes.

These data confirm that, while transformer-based models are effective in identifying popular PII categories, addressing unusual ones remains a substantial difficulty. Techniques such as synthetic data creation or domain-specific augmentation may provide a road ahead for improving model performance. Furthermore, combining the capabilities of different models using ensemble approaches may produce better results, particularly when balancing performance throughout all categories.

The present results are in agreement with the previous studies, taking into consideration that research by Timmer et al. (2022) from the year 2022 was done regarding the use of transformers for sensitive data handling. However, it reflects limitations due to unbalanced datasets and it points out that there is a need to make more adjustments in preparing the samples and the development process. Future research should focus on domain-specific pre-training to improve models' contextual understanding and lightweight transformer adaptations to ensure practical deployment in resource-constrained environments. While these models demonstrated significant promise, the findings reveal opportunities for further optimization to fully address the complexities of PII detection tasks.

7 Conclusion and Future Work

This paper focused on the recognition of PII in unstructured text using transformer-based models. A comparison was made between four state-of-the-art models, namely Longformr, DeBERTa, RoBERTa, and DistilBERT, using various strategies that included hyperparameter tuning and regularization to achieve better performance. The idea was to find out which model was better and explore the challenges of detecting PII under different scenarios, also investigating the practical implications of using those models in realistic settings by performing extensive testing of these models.

Results have shown that transformer-based algorithms go exceptionally well in detecting PII. The best performance was that of DeBERTa, which was outstanding for both common and unusual PII categories. With a superior detached attention mechanism and relative place embeddings, DeBERTa had the capability of learning deep relations within text and hence was the best for sensitive data detection. Longformer had particularly impressive performance on very long documents, sustaining contextual integrity over longer sequences. RoBERTa had a good balance between performance and computational efficiency while DistilBERT was computationally very light but couldn't handle the underrepresented PII types very well. Even though the dataset balancing techniques were used with capping dominant categories and augmenting rare ones, the inherent sparsity within some of the PII categories, like email addresses and ID numbers, ensured that the models do not generalize well.

The principle goals were achieved by this research, which detailed findings about the results, benefits, and limitations of the investigated models. Besides this, some of the key identified challenges were those connected with class imbalance and computing needs, whose solution can be earnestly sought before deploying them in security systems. This serves to reinforce the general field of sensitive data identification by providing concrete recommendations for scholarly researchers and industrial practitioners alike.

7.1 Limitations

The data imbalancing in the dataset was one of the most important obstacles. Even though the balancing strategies were applied, the scarcity of different types of PII prevented the models from generalizing well. This observation agrees with previous related works such as Neerbek (2020) and Chong (2022), which reported similar challenges when detecting less frequent sensitive information. A further concern was that one was dependent on a single dataset, which by implication reduced generalizability of the conclusions across different domains. The other practical limitation is that fine-tuning such large models as DeBERTa is very computationally costly; hence, it is not that practical in resource-constrained settings.

7.2 Future Work

Future studies should therefore focus on overcoming these identified limitations. Improving uncommon PII categories may call for an update in the dataset composition toward including more diverse and balanced samples, probably generated synthetically. Ensemble methods could yield the best outcome as they combine the powers of several models and effectively reduce their respective deficiencies. This could be furthered by developing lightweight transformer models which would become specialized, say in health or finance, that could then be easily deployed in scenarios where computational resources are at a premium.

These cross-lingual transformer algorithms can also be used to extend the detection of PII to several language datasets. Embedding these models into secure technologies will therefore let organizations automate their processes in compliance with data protection laws such as GDPR and CCPA. Such devices, powered by the most effective models, offer flexible, reliable options to anonymize sensitive data. These findings are opening business opportunities for enterprise-ready systems which will do the identification of PII automatically. The technologies can be applied to different industries, offering the necessary instrumentation for organizations interested in acquiring better protection for sensitive information based on legal and ethical considerations. Further research in this direction has also got to take along the path of bridging the gap between academia and enterprises in better model development and exploring further scenarios.

This work establishes the effectiveness of transformer-based models in the detection of PII in this respect. However, it also brought the areas that needed improvement. Overcoming those limitations by following the suggested future direction improves the accuracy, growth, and applicability of these models to a much greater degree, thereby contributing to the evolution of privacy-preserving technologies.

8 Acknowledgements

I am especially grateful to Dr. Muslim Jameel Syed for his support and encouragement as well as careful feedback I received during the stages of developing my thesis and code development.His expertise and support have been instrumental in shaping this research work.I also appreciate the authors of previous research work on which this study is aimed because their work provided premise for this exercise.

References

Beltagy, I., Peters, M. E. and Cohan, A. (2020). Longformer: The long-document transformer.
URL: *https://arxiv.org/abs/2004.05150*

BigCode (2022). Starpii: A model for detecting personally identifiable information in code datasets.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A. and Raffel, C. (2021). Extracting training data from large language models.
URL: *https://arxiv.org/abs/2012.07805*

Chong, P. (2022). Deep learning based sensitive data detection, pp. 1–6.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding, pp. 4171–4186.
URL: *https://aclanthology.org/N19-1423*

Dias, M., Boné, J., Ferreira, J. C., Ribeiro, R. and Maia, R. (2020). Named entity recognition for sensitive data discovery in portuguese, *Applied Sciences* **10**(7).
URL: *https://www.mdpi.com/2076-3417/10/7/2303*

Garfinkel, S. (2015). De-identification of personal information.

Goddard, M. (2017). The eu general data protection regulation (gdpr): European regulation that has a global impact, *International Journal of Market Research* **59**(6): 703–705.
URL: *https://doi.org/10.2501/IJMR-2017-050*

He, P., Liu, X., Gao, J. and Chen, W. (2021). Deberta: Decoding-enhanced bert with disentangled attention.
URL: *https://arxiv.org/abs/2006.03654*

Kulkarni, P. and K, C. (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**.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach.
URL: *https://arxiv.org/abs/1907.11692*

Mattsson, U. (2020). Practical data security and privacy for gdpr and ccpa, *ISACA Journal* **2020**(3).
URL: *https://www.isaca.org/resources/isaca-journal/issues/2020/volume-3/practical-data-security-and-privacy-for-gdpr-and-ccpa*

Muralitharan, J. and Arumugam, C. (2024). Privacy BERT-LSTM: a novel NLP algorithm for sensitive information detection in textual documents, *Neural Computing and Applications* **36**(25).
URL: *https://doi.org/10.1007/s00521-024-09707-w*

Neerbek, J. (2020). Sensitive information detection: Recursive neural networks for encoding context.
URL: *https://arxiv.org/abs/2008.10863*

Sanh, V., Debut, L., Chaumond, J. and Wolf, T. (2020). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.
URL: *https://arxiv.org/abs/1910.01108*

Schmid, P. (2022). Advanced pii detection and anonymization with hugging face transformers and amazon sagemaker.
URL: *https://www.philschmid.de/pii-huggingface-sagemaker*

Sheikh, M. and Conlon, S. (2012). A rule-based system to extract financial information, *Journal of Computer Information Systems* **52**.

Singh, S. and Hooda, S. (2023). A study of challenges and limitations to applying machine learning to highly unstructured data, pp. 1–6.

Sweeney, L. (2013). Discrimination in online ad delivery, *Commun. ACM* **56**(5): 44–54.
URL: *https://doi.org/10.1145/2447976.2447990*

Timmer, R. C., Liebowitz, D., Nepal, S. and Kanhere, S. S. (2022). Can pre-trained transformers be used in detecting complex sensitive sentences? – a monsanto case study.
URL: *https://arxiv.org/abs/2203.06793*

Truong, A., Walters, A. and Goodsitt, J. (2020). Sensitive data detection with high-throughput neural network models for financial institutions.
URL: *https://arxiv.org/abs/2012.09597*

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. and Polosukhin, I. (2023). Attention is all you need.
URL: *https://arxiv.org/abs/1706.03762*

Zhang, K. and Jiang, X. (2023). Sensitive data detection with high-throughput machine learning models in electrical health records.
URL: *https://arxiv.org/abs/2305.03169*

Zhang, K. and Qiu, X. (2024). Comparing feature-based and context-aware approaches to pii generalization level prediction.
URL: *https://arxiv.org/abs/2407.02837*

Zhao, S., Chen, S. and Wei, Z. (2021). Statistical feature-based personal information detection in mobile network traffic.
URL: *https://arxiv.org/abs/2112.12346*