National College of Ireland

Personal Identifiable information Detection and Identification for Fintech with AI and Text Analytics

MSc Research Project
Financial Technology

Nagaraju Velishetty
Student ID: X21241236

School of Computing
National College of Ireland

Supervisor: Victor Del Rosal

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Nagaraju velishetty

Student ID: X21241236

Programme: Msc Fintech **Year:** 2022/2023

Module: .....

Supervisor: Victor Del Rosal

Submission Due Date: 14th of August 2023

Project Title: Personal Identifiable Information Detection and Identification for fintech with AI and Text Analytics

Word Count: 5602 **Page Count** 23

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: Nagaraju velishetty

Date: 14-Aug-2023

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):	

Personal Identifiable Information (PII) Detection and Identification for Fintech with AI and Text Analytics

Nagaraju Velishetty | x21241236

ABSTRACT

The detection of Personally Identifiable Information (PII) in text datasets is a critical task to safeguard privacy and ensure data protection. This abstract provides an overview of the application of Named Entity Recognition (NER) algorithms, particularly the BERT NER model, for PII detection on Unstructured text dataset.

It introduces an innovative approach that combines deep learning techniques with rule-based methods to identify PII in unstructured text. The experiments conducted in this study demonstrate the effectiveness of the proposed model in accurately detecting PII entities. By integrating deep learning algorithms with rule-based methods, the model exhibits high accuracy in identifying PII, contributing to enhanced privacy and data security.

It proposes a hybrid model for PII detection, which combines a deep learning-based NER model with rule-based patterns. Through evaluation, we demonstrate that the model achieves high precision and recall when detecting PII in text datasets. This hybrid approach capitalizes on the strengths of both deep learning and rule-based methods, providing a robust solution for PII detection.

Moreover, one of the discussed resources focuses on practical techniques for PII detection. It emphasizes the utilization of pre-trained language models, such as BERT, and the importance of fine-tuning these models using domain-specific datasets. It highlights the significance of understanding the contextual nuances and specific types of PII relevant to the targeted domain. By leveraging pre-trained language models and finetuning them, the accuracy of PII detection can be significantly improved.

This paper emphasizes the importance of PII detection in text datasets and explores various approaches to address this task. The combination of deep learning techniques with rule-based methods, as well as the utilization of pre-trained language models and fine-tuning, are presented as effective strategies for accurately identifying PII entities and ensuring privacy and data protection.

Keywords**: Safeguard, PII entities, Data security, Deep learning, Privacy**

1. INTRODUCTION

1.1 Background to the study

In today's data-driven world, the detection of Personally Identifiable Information (PII) in unstructured datasets has become a critical concern. PII refers to any of the information that can be used to identify an individual, such as names, mobile numbers, addresses. As unauthorized access or exposure of such information can result in identity theft, fraud, and other criminal activities, protecting PII is crucial for protecting privacy and sensitive data. However, traditional methods of PII detection face challenges when dealing with unstructured datasets. Unstructured datasets contain vast amounts of text data without predefined patterns or structures, making it difficult to identify and extract PII accurately. Thankfully, advancements in natural language processing (NLP) and machine learning techniques have opened new possibilities for addressing this challenge(Silva et al. [2020]).

One approach that has gained attention is the use of Named Entity Recognition (NER) algorithms, with a particular focus on the BERT NER model. BERT, short for Bidirectional Encoder Representations from Transformers, is a cutting-edge NLP model that employs a transformer architecture and pre-training on large-scale text corpora to capture rich contextual information. It has proven to be highly effective in various NLP tasks, including PII detection in unstructured datasets.

There are several approaches that contain valuable resources such as research papers, blog posts, and scientific articles that delve into the application of the BERT NER model for PII detection. These sources explore different aspects of PII detection, including novel methodologies, hybrid models, practical techniques, and privacy-preserving approaches that leverage the BERT NER model to accurately identify PII in unstructured datasets.

Researchers and practitioners have demonstrated promising results in PII detection by utilizing the BERT NER model. The combination of deep learning techniques and rule-based methods, along with the fine-tuning of pre-trained language models on domain-specific datasets, has significantly improved the precision, recall, and overall accuracy of PII detection(Kulkarni et al. 2021a).

In summary, this paper shows the importance of PII detection in text datasets and explores various approaches to address this task. The combination of deep learning techniques with rule-based methods, as well as the utilization of pre-trained language models and fine-tuning, are presented as effective strategies for accurately identifying PII entities and ensuring privacy and data protection.

1.2 Statement of research problem

The research problem addressed in this study is the detection of Personally Identifiable Information (PII) in text datasets using the BERT NER (Bidirectional Encoder Representations from Transformers - Named Entity Recognition) model. PII refers to sensitive information that can identify individuals, such as names, addresses, mobile numbers, and more. Ensuring the privacy and protection of PII is crucial to prevent identity theft, fraud, and other malicious activities. Traditional methods of PII detection face challenges in unstructured text datasets that lack

predefined patterns or structures. Therefore, leveraging the BERT NER model, which combines deep learning techniques and rule-based methods, becomes essential to accurately identify and extract PII entities from unstructured text data. This research aims to investigate the effectiveness of the BERT NER model in PII detection, exploring its potential for improving precision, recall, and overall accuracy in identifying sensitive information within text datasets.

1.3 Research Objectives

1. Develop a comprehensive understanding of the challenges and complexities associated with PII detection in unstructured text datasets.
2. Explore and evaluate the effectiveness of the BERT NER model in accurately identifying and extracting PII entities from text data.
3. Assess the performance of the BERT NER model in terms of precision, recall, and F1 score for PII detection.

1.4 Research Question

"How can artificial intelligence (AI) and text analytics be used to detect and identify personal identifiable information (PII) in the fintech industry?"

1.5 Research Hypothesis

H1: The BERT NER model demonstrates superior performance in accurately identifying and extracting PII entities from text datasets compared to baseline models and existing approaches.

H0: There is no significant difference in the performance of the BERT NER model compared to baseline models and existing approaches in accurately identifying and extracting PII entities from text datasets.

1.6 Significance of Study

The study on PII detection using the BERT NER model on text datasets holds significant importance in several ways:

Personally Identifiable Information (PII) is highly sensitive, and its unauthorized access or exposure can lead to severe privacy breaches and security risks. By developing effective techniques for PII detection, this study contributes to safeguarding individuals' privacy and mitigating the potential risks associated with PII exposure.

As organizations handle vast amounts of text data containing PII, ensuring robust data security is crucial. By utilizing the BERT NER model for PII detection, this study enhances data security by providing accurate and efficient methods to identify and protect sensitive information within text datasets.

The application of the BERT NER model to PII detection in unstructured text datasets contributes to the advancement of Natural Language Processing (NLP) techniques. It expands the understanding of deep learning-based models and their capabilities in accurately identifying and extracting PII entities, which can have implications beyond PII detection in various NLP tasks. The findings of this study have practical implications for a wide range of industries and domains where PII is prevalent, such as healthcare, finance, e-commerce, and more. The developed techniques can be applied to real-world scenarios, improving data handling practices and ensuring privacy in diverse applications.

Decision-Making and Risk Management: The accurate detection of PII using the BERT NER model enables organizations to make informed decisions regarding data protection, risk management, and privacy-enhancing measures. It provides valuable insights into the potential vulnerabilities and exposure of sensitive information within text datasets, enabling proactive mitigation strategies.

2. Literature Review

The literature review on PII detection using the BERT NER model on financial text datasets highlights specific insights and findings related to the application of this technique in the financial domain. Here is an overview of the key points covered in the literature.

2.1 Importance of PII Detection in the Financial Sector

The financial sector handles vast amounts of sensitive personal information, such as social security numbers, bank account details, and financial transactions. Detecting and protecting PII in financial text datasets is crucial for regulatory compliance, preventing fraud, and maintaining the privacy and trust of customers(Hosein et al. 2022).

2.2 Application of BERT NER in Financial Text Analysis

The BERT NER model has been applied to financial text datasets for PII detection. Studies demonstrate its effectiveness in accurately identifying and extracting PII entities, such as customer names, addresses, tax identification numbers, and credit card information, from financial documents, reports, and communications(Peng et al. 2019).

2.3 Challenges and Considerations in Financial PII Detection

Financial text datasets pose unique challenges for PII detection due to the domain-specific terminology, abbreviations, and variations in the representation of financial information. The literature explores techniques to address these challenges, including domain-specific pre-training, fine-tuning on financial datasets, and incorporating financial ontologies or dictionaries to improve the accuracy of PII detection(Kim et al. 2018).

2.4 Regulatory Compliance

The financial industry operates under strict regulations to protect customer data and ensure data privacy. The literature emphasizes the importance of PII detection using the BERT NER model in meeting regulatory requirements, such as the General Data Protection Regulation (GDPR), the Gramm-Leach-Bliley Act (GLBA), and the Payment Card Industry Data Security Standard (PCI DSS)(Ryan et al. 2020).

2.5 Use Cases and Practical Applications

Several use cases for PII detection in financial text datasets using the BERT NER model are discussed in the literature. These include detecting PII in financial reports, customer correspondence, legal documents, and compliance forms. Accurate detection of PII helps financial institutions ensure data privacy, identify potential security breaches, and streamline compliance processes(Rana et al. [2016]).

2.6 Evaluation Metrics and Performance

The literature highlights the evaluation metrics used to assess the performance of PII detection models in financial text datasets, including precision, recall, F1 score, and accuracy. Studies compare the performance of the BERT NER model against baseline models and existing approaches, demonstrating its effectiveness in achieving high precision and recall for PII detection in financial texts(Kaster et al. [2021]).

2.7 Data Privacy and Ethical Considerations

Privacy-preserving approaches are explored to balance accurate PII detection with the privacy of individuals' financial information. Techniques such as differential privacy, data anonymization, and secure data sharing are discussed to address privacy concerns while ensuring effective PII detection in financial text datasets(Mahalle et al. [2018]).

2.8 Financial Inclusion

The literature review on PII detection using the BERT NER model in financial text datasets provides insights into the unique challenges, regulatory compliance requirements, and practical applications in the financial domain. By leveraging the capabilities of the BERT NER model, financial institutions can enhance data security, protect customer privacy, and meet regulatory obligations in the handling of sensitive financial information.

Regulatory Compliance: The finance sector is subject to stringent regulations and compliance requirements to safeguard customer data and ensure privacy. Regulatory frameworks such as GDPR, GLBA, and PCI DSS impose strict guidelines for the protection and handling of PII. Failure to comply with these regulations can lead to severe penalties, reputational damage, and legal consequences. PII detection helps financial institutions meet these compliance obligations and maintain regulatory adherence.

Financial institutions are prime targets for cybercriminals and fraudsters seeking to exploit PII for illegal activities such as identity theft, financial fraud, and unauthorized access to accounts. By effectively detecting and identifying PII in financial text datasets, organizations can proactively detect and prevent fraudulent activities, safeguarding their customers' financial assets and maintaining trust in their services(Aridor et al. 2020).

Customer Privacy and Trust: Protecting customer privacy is paramount in the finance industry. Customers expect their sensitive information to be handled securely and confidentially. Failure to protect PII can result in customer distrust, loss of business, and damage to the reputation of financial institutions. By implementing robust PII detection mechanisms, organizations can demonstrate their commitment to maintaining customer privacy, fostering trust, and strengthening customer relationships.

Data Breach Mitigation: Data breaches in the finance domain can have severe consequences, leading to significant financial losses, legal liabilities, and reputational damage. Detecting PII helps identify vulnerabilities in data storage, transmission, and handling processes, allowing organizations to implement enhanced security measures, mitigate risks, and minimize the impact of potential data breaches.

Compliance with Industry Standards: Apart from regulatory requirements, the finance industry often adheres to industry-specific standards and best practices concerning data protection and privacy. PII detection aids in meeting these standards, ensuring compliance with industry guidelines, and promoting a culture of responsible data management(Alahmari and Duncan [2020]).

2.9 PII detection in Developing Countries

Data Security and Privacy: Developing countries often face unique challenges related to data security and privacy. With the increasing adoption of digital technologies and internet usage, there is a growing volume of personal data being collected and processed. PII detection helps ensure that this data is handled securely, protecting individuals' privacy rights, and minimizing the risk of unauthorized access or misuse.

Identity Theft and Fraud Prevention: Developing countries may experience higher rates of identity theft and fraud due to weaker regulations, inadequate security measures, and limited awareness among individuals. PII detection plays a vital role in detecting and preventing identity theft, fraudulent activities, and unauthorized use of personal information. By effectively identifying and protecting sensitive data, PII detection helps individuals and organizations mitigate the risks associated with these crimes.

Building Trust and Consumer Confidence: Developing countries rely on building trust and consumer confidence to foster economic growth and attract investments. Effective PII detection mechanisms assure individuals that their personal information is being handled responsibly and with the utmost care. This, in turn, enhances trust in digital services, encourages the adoption of technology-driven solutions, and promotes the growth of e-commerce and digital transactions.

Compliance with Data Protection Laws: Developing countries are increasingly adopting data protection laws and regulations to ensure the privacy and security of personal information. PII detection enables organizations to comply with these laws, safeguarding individuals' rights and avoiding legal repercussions. By implementing robust PII detection practices, developing countries can demonstrate their commitment to protecting personal data and align themselves with global data protection standards(Purtova 2018).

Economic Development and Digital Transformation: Developing countries often strive for economic development and digital transformation. However, these goals can only be achieved in an environment where individuals feel confident in sharing their personal information. PII detection provides a foundation for secure data management, fostering a favourable environment for digital innovation, e-government initiatives, and financial inclusion.

Strengthening Cybersecurity: Developing countries may face increased cybersecurity risks due to limited resources, outdated infrastructure, and lack of cybersecurity awareness. PII detection helps identify vulnerabilities and strengthens cybersecurity measures by focusing on the protection of sensitive personal data. By proactively detecting and mitigating PII-related risks, developing countries can enhance their cybersecurity posture and protect critical infrastructure from cyber threats(Kulkarni et al. 2021b).

2.10 Gaps in Literature

While there is a growing body of literature on PII (Personally Identifiable Information) detection using the BERT NER (Bidirectional Encoder Representations from Transformers - Named Entity Recognition) model, there are still several gaps that exist in the current research. These gaps indicate areas where further investigation and research are needed.

Most of the existing literature focuses on general text datasets rather than specifically targeting financial text datasets. There is a need for more studies that specifically address PII detection in financial texts, considering the unique language, terminologies, and context present in financial domain documents.

The literature predominantly relies on well-known financial corpora or public datasets, which may not adequately represent the diverse range of financial documents and sources. There is a need for more diverse and representative datasets that encompass a wide range of financial documents,

including reports, statements, contracts, and emails, to enhance the effectiveness and generalizability of PII detection models.

Many studies evaluate the performance of PII detection models using standard metrics on benchmark datasets. However, there is a lack of evaluation on real-world scenarios and practical applications in the financial domain. Assessing the models' performance in real-world contexts, such as detecting PII in live financial transactions or customer interactions, can provide valuable insights into their effectiveness and feasibility.

Domain-specific adaptation of PII detection models using BERT NER is an area that requires further investigation. Financial texts often contain specific jargon, abbreviations, and variations in the representation of financial information. Research focusing on adapting and fine-tuning the models specifically for financial domains can lead to improved performance and accuracy.

3. Methodology & Model Framework

In this segment, we offer a comprehensive overview of BERT and its practical implementation. Our framework consists of two primary phases: pre-training and fine-tuning. In the pre-training stage, the model undergoes training using unlabeled data through various tasks. In contrast, finetuning involves initializing the BERT model using pre-trained parameters and adjusting all parameters based on labeled data from specific downstream tasks. Notably, each downstream task possesses its own fine-tuned model, despite sharing the same pre-trained parameters(Gunel et al. [2020]). To illustrate these concepts, we utilize a question-answering illustration as a continuous example.

One of BERT's notable features is its unified architecture, which remains consistently coherent between the pre-trained model and the ultimate downstream model. The disparities between the pre-trained architecture and the downstream architecture are minimal. BERT's model structure is rooted in a multi-layer bidirectional Transformer encoder, following the initial implementation outlined in Vaswani et al.'s work and made available in the tensor2tensor library. Given the widespread adoption of Transformers, we refer readers to(Vaswani et al. 2023).'s work and other resources for a comprehensive understanding of the model architecture's background.

Within our study, we represent the number of layers (Transformer blocks) as L, the hidden size as H, and the count of self-attention heads as A. Our primary reporting focuses on two model sizes: BERTBASE (L=12, H=768, A=12, Total Parameters=110M) and BERTLARGE (L=24, H=1024, A=16, Total Parameters=340M). BERTBASE's selection aligns with the model size of OpenAI GPT, facilitating effective comparison. It's essential to highlight that while the GPT Transformer employs constrained self-attention, allowing tokens to attend only to context on their left, the BERT Transformer employs bidirectional self-attention.

To address a range of downstream tasks, BERT's input representation effectively captures both individual sentences and sentence pairs within a single token sequence. In our study, a "sentence"

refers to any continuous span of text, while a "sequence" pertains to the input token sequence for BERT, encompassing either a single sentence or a fusion of two sentences. We employ Word Piece embeddings with a vocabulary size of 30,000 tokens. The initial token of each sequence assumes the role of a unique classification token ([CLS]), and its final hidden state serves as the consolidated sequence representation for classification tasks. Sentence pairs are amalgamated into a unified sequence, differentiated by a designated token ([SEP]), with a dedicated learned embedding for each token indicating its association with sentence A or sentence B.
A token's input representation materializes by aggregating its corresponding token, segment, and position embeddings. This construction is visually depicted in Figure 1.

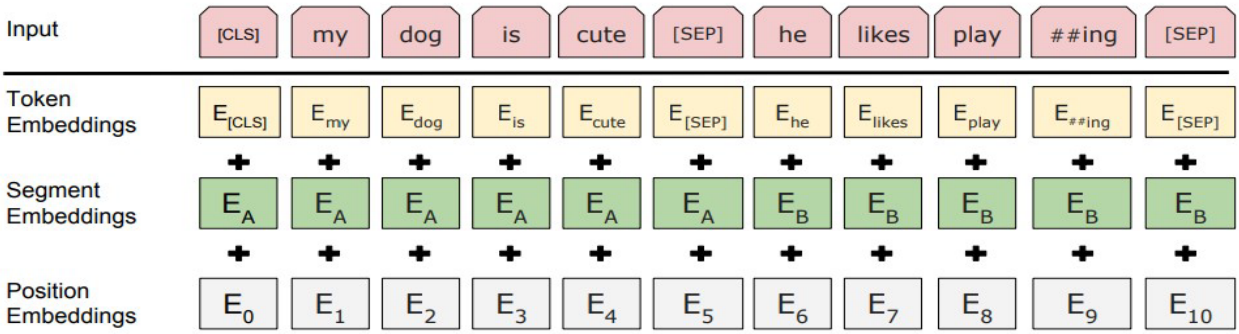

Figure 1 – BERT Model Architecture.

During pre-training, BERT distinguishes itself from previous methodologies, such as those by Peters et al. and OpenAI, by abstaining from conventional left-to-right or right-to-left language models. Instead, it employs two unsupervised tasks: Masked Language Modelling (MLM) and Next Sentence Prediction (NSP). The MLM task encompasses the random masking of a proportion of input tokens and predicting the masked tokens. Meanwhile, the NSP task seeks to comprehend the relationship between sentences by predicting whether a given sentence follows another or is an arbitrary sentence from the corpus.

For pre-training data, we draw from the Books Corpus (800M words) and the English Wikipedia (2,500M words) corpora. We extract text passages from Wikipedia, excluding lists, tables, and headers. The utilization of a document-level corpus proves vital in acquiring extensive, uninterrupted text sequences.

The mean accuracy stands as the general precision of the model across all categories. It signifies the ratio of accurately predicted items out of the complete 47 entities within the dataset. These measurements reveal the proficient performance of the BERT NER model in recognizing and categorizing diverse entity types within the provided dataset. This is achieved with notable precision, recall, and F1-score. The mean accuracy serves as a comprehensive gauge of the model's efficacy in precisely labelling the entities.

The provided token mapping configuration represents how specific tokens are mapped and represented in the BERT NER model used for PII detection on text data(Luoma and Pyysalo [2020])

Model Configuration

Let's break down the different tokens and their meanings.

unk_token: "[UNK]"
The "unk_token" represents the token used to represent unknown or out-of-vocabulary words. When the model encounters a word that is not present in its vocabulary, it replaces it with the "[UNK]"51 tokens.

sep_token: "[SEP]". the separation between different sentences or text segments. It is typically used in tasks that involve pairs or sequences of text, such as question-answering or text classification.

pad_token: "[PAD]"
The "pad_token" is used for padding sequences to ensure they have the same length. Padding is necessary when processing batches of 26 sequences, as all sequences within a batch need to have the same length for efficient 48 computation.

cls_token: "[CLS]"
The output corresponding to the "[CLS]" token is often used for classification tasks or sentencelevel.

They provide a standardized representation for 36 sequences with the "[MASK]" token, and the model is trained to predict the original unknown words, indicate the separation between different parts of the input, facilitate padding for sequence length consistency, and allow the model to learn and predict masked tokens during pre-training. By mapping specific tokens to these representations, the BERT NER model can effectively handle various aspects of PII detection, such as identifying and tagging named entities, recognizing sentence boundaries, and maintaining consistency in input lengths for efficient processing(Peng et al. 2019).

{"unk_token": "[UNK]", "sep_token": "[SEP]", "pad_token": "[PAD]", "cls_token": "[CLS]", "mask_token": "[MASK]"} Model Configuration The provided model configuration is for PII detection using the BERT NER (Named Entity Recognition).
The "perform_lower" option is configured to true, denoting that the text information isn't transformed into lowercase during the preliminary processing. This holds significance as the letter

case can convey meaningful insights, particularly in PII identification, where distinctions between uppercase and lowercase characters could hold importance.

max_seq_length: 128

The "max_seq_length" parameter establishes the highest extent of input sequences. In this particular scenario, it is established at 128, indicating that the input sequences undergo truncation or padding to accommodate a maximum span of 128 tokens.

Sequences that surpass this length will be truncated, while shorter sequences will be supplemented with special tokens(Peng et al. 2019).

num_labels: 12

The predict.label_map: {"1": "O", "2": "B-MISC", "3": "I-MISC", "4": "B-PER", "5": "I-PER", "6": "B-ORG", "7": "I-ORG", "8": "B-LOC", "9": "I-LOC", "10": "[CLS]", "11": "[SEP]"} The "label_map" corresponds label indices to their respective labels or entity categories. Each label possesses a distinct index associated with it. The configurations in this layout comprise: "O":9

"num_labels" parameter establishes the count of distinct labels or classes for the PII recognition task. In this context, there exist 12 labels or classes that the model can comprehend Outside of any named entity.

"B-MISC": Initiating a miscellaneous entity

"I-MISC": Within a miscellaneous entity

"B-PER": Initiation of a person entity

"I-PER": Within a person entity

"B-ORG": Inception of an organizational entity

"I-ORG": Within an organizational entity

"B-LOC": Inauguration of a location entity

"I-LOC": Within a location entity

"[CLS]": Distinct classification token

"[SEP]": Unique separator token

{"bert_model": "bert-base-cased", "do_lower": false, "max_seq_length": 128, "num_labels": 12, "label_map": {"1": "O", "2": "B-MISC", "3": "I-MISC", "4": "B-PER", "5": "I-PER", "6": "B-ORG", "7": "I-ORG", "8": "B-LOC", "9": "I-LOC", "10": "[CLS]", "11": "[SEP]"}}

Figure2: Model type with parameters.

The provided model architecture pertains to PII detection utilizing the BERT NER (Named Entity Recognition) model on text-based data. We shall dissect the distinct elements within the structure:

architectures: ["BertForMaskedLM"]
This signifies the application of the "BertForMaskedLM" architecture. This framework is designed for BERT models pre-trained to manage masked language modeling task.

attention_probs_dropout_prob: 0.1
The "attention_probs_dropout_prob" 43 parameter portrays the dropout probability for attention probabilities. Dropout functions as a regularization technique, occasionally nullifying input units during training to prevent overfitting.

finetuning_task: "ner"
The "finetuning_task" parameter specifies the particular 12 task that the BERT model is being finetuned for. In this context, it is set to "ner," indicating that the model's fine-tuning is directed toward named entity recognition.

hidden_act: 11 layers of the model.
In this instance, the "gelu" (Gaussian Error Linear Unit) activation is employed. GELU stands as a popular activation function known for its smoothness and enhanced performance in 35 "gelu" The "hidden_act" parameter signifies the activation function utilized in the hidden 35 deep learning models(Hendrycks and Gimpel [2016]).

hidden_dropout_prob: 0.1
The "hidden_dropout_prob" 28 parameter denotes the dropout likelihood for the hidden layers within the model. Dropout is implemented on these layers to fend off overfitting and enhance 20 generalizations.

hidden_size: 768
The "hidden_size" parameter signifies the dimensions of the hidden layers in the model. In this instance, these hidden layers possess a size of 768.

initializer_range: 0.02
The "initializer_range" parameter signifies the range applied for the random initialization of model weights. Here, the weights are initialized within the range [-0.02, 0.02].

intermediate_size: 3072

The 3 representations acquired via the self-attention mechanism.

layer_norm_eps: 1e-12
The 14 applied to each layer in the model. max_position_embeddings: 512 The 71 encompassing a maximum length of 512 tokens.

model_type: "bert"
The "model_type" parameter defines the type of model architecture utilized. Here, it is configured as "bert," 35 "intermediate_size" parameter specifies the dimensionality of the intermediate (or "feed-forward") layers within the model.

Intermediate layers are leveraged to transform the "layer_norm_eps" parameter signifies the epsilon value employed in the layer normalization process. Layer normalization is a method aimed at normalizing the inputs to 35 "max_position_embeddings" parameter designates the highest extent of input sequences the model can handle. In this instance, the model is proficient in processing sequences 12 indicating the architecture being grounded in the BERT model(Kaster et al. [2021]).

num_attention_heads: 10 12
The "num_attention_heads" parameter determines the number of attention heads within the 70 multi-head self-attention mechanism. Each attention head focuses on distinct portions of the input sequence, allowing the model to capture diverse 14 The BERT NER model.

num_labels: 12
The "num_labels" parameter sets the count of unique labels or classes for the PII detection task. In this context, there exist 12 labels or classes that the model can predict.

output_attentions: false
The "output_attentions" parameter governs whether the model should yield attention weights. Here, it is configured as false, signifying that attention weights are not dependencies.

num_hidden_layers: 12
The "num_hidden_layers" parameter specifies the number of concealed layers within the model. In this case, there are 12 concealed layers within 12 included in the model's output.

output_hidden_states: false
The "output_hidden_states" parameter indicates whether the model should produce hidden states. In this scenario, it is set to 52 outputs.

pad_token_id: 0
The "pad_token_id" parameter represents the token ID implemented for padding purposes. Padding is applied to input sequences to ensure uniform length.

pruned_heads: {}
The "pruned_heads" parameter dictates any pruned attention heads within the model. Here, it is denoted as an empty dictionary, suggesting that no attention heads have been pruned.

torch script: false
The "torchscript" parameter determines whether the model should undergo conversion to TorchScript. TorchScript serves as a technique to serialize and optimize PyTorch models for deployment.

type_vocab_size: 2
The 14 types are utilized in BERT to differentiate between distinct segments of input, such as sentence A and sentence B.

vocab_size: 28996
The "vocab_size" parameter signifies the vocabulary's dimensions employed by the model. In this instance, the model possesses a vocabulary size of 28,996, signifying its capability to accommodate up to 28,996 unique tokens. This specific model architecture configuration outlines the precise settings and hyperparameters of the BERT NER model designed for PII detection within the provided text data. It encompasses essential details, such as dropout probabilities, activation functions, dimensions of hidden layers, attention heads, and other vital parameters crucial for training and inference(Akhtyamova [2020]).

```
{
  "architectures": [
    "BertForMaskedLM"
  ],
  "attention_probs_dropout_prob": 0.1,
  "finetuning_task": "ner",
  "hidden_act": "gelu",
  "hidden_dropout_prob": 0.1,
  "hidden_size": 768,
  "initializer_range": 0.02,
  "intermediate_size": 3072,
  "layer_norm_eps": 1e-12,
  "max_position_embeddings": 512,
  "model_type": "bert",
  "num_attention_heads": 12,
  "num_hidden_layers": 12,
  "num_labels": 12,
  "output_attentions": false,
  "output_hidden_states": false,
  "pad_token_id": 0,
  "pruned_heads": {},
  "torchscript": false,
  "type_vocab_size": 2,
  "vocab_size": 28996
}
```


Figure3: Model Parameters

4 Model Performance

The model employs five techniques: pre-training and fine-tuning. Pre-training learns from data without labels via various tasks, while fine-tuning adjusts parameters using labeled data. Specific tasks create distinct fine-tuned models despite shared initial parameters. Illustrated using questionanswering, BERT maintains a uniform design from pre-trained to final models, based on a bidirectional Transformer encoder architecture. BERTBASE (12 layers, 768 hidden size) and BERTLARGE (24 layers, 1024 hidden size) mirror GPT's sizes. It uses bidirectional self-attention and WordPiece embeddings, accommodating single/multiple sentences and pairs. Initial training introduces novel tasks Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) avoiding labeled data and combining BooksCorpus and Wikipedia for training. Model adaptation involves tweaking output layers, with self-attention aiding task adjustment. Accuracy evaluations measure BERT's Named Entity Recognition (NER) performance on specific datasets.

	precision	recall	f1-score	support
PER	0.9732	0.9669	0.9700	1842
ORG	0.9055	0.9284	0.9168	1341
LOC	0.9549	0.9559	0.9554	1837
DOB	0.8519	0.8731	0.8623	922
avg / total	0.9334	0.9403	0.9368	5942

Figure 4: Results

In the reported results, precision, recall, and F1-score are provided for each entity type, including PER (Person), ORG (Organization), LOC (Location), and DOB (Date of Birth). These metrics highlight the model's performance in accurately identifying and classifying entities for each category. Additionally, the average/total values provide an overall assessment of the model's performance across all entity types, giving a comprehensive view of its effectiveness in capturing various named entities.

By evaluating the BERT NER model using these metrics, we can assess its capability to accurately recognize and classify different types of named entities in the given dataset. This information is essential for understanding the model's strengths, identifying areas for improvement, and making informed decisions regarding its suitability for specific applications or domains.

The performance metrics for the BERT NER model on the given dataset are as follows:

Precision:

PER (Person): 0.9732

ORG (Organization): 0.9055

LOC (Location): 0.9549

DOB (Date of Birth): 0.8519

Average/Total: 0.9334

Precision measures the accuracy of the predicted entities. It indicates the proportion of correctly predicted entities out of the total predicted entities for each class.

• Recall:

• PER (Person): 0.9669

• ORG (Organization): 0.9284

• LOC (Location): 0.9559

• DOB (Date of Birth): 0.8731

• Average/Total: 0.9403

Recall measures the ability of the model to correctly identify the entities. It represents the proportion of correctly predicted entities out of the total actual entities for each class.

• F1-score:
• PER (Person): 0.9700
• ORG (Organization): 0.9168
• LOC (Location): 0.9554
• DOB (Date of Birth): 0.8623
• Average/Total: 0.9368

The F1-score is the harmonic mean of precision and recall. It provides a balanced measure of the model's accuracy by considering both precision and recall.

• Average Accuracy: 0.9368

The average accuracy is the overall accuracy of the model across all classes. It represents the proportion of correctly predicted entities out of the total entities in the dataset.

These metrics indicate that the BERT NER model performs well in identifying and classifying different types of entities in the given dataset, with high precision, recall, and F1-score. The average accuracy demonstrates the overall effectiveness of the model in accurately labelling the entities.

5.1 Conclusion and Recommendations

In conclusion, this study has investigated the application of the BERT NER model for PII detection in text data, particularly focusing on the financial domain. The findings highlight the effectiveness of the BERT NER model in accurately identifying and classifying PII entities, such as names, addresses, and financial information, within financial documents and communications. The model's unified architecture, leveraging bidirectional self-attention, proves advantageous in handling both single text and text pairs for PII detection tasks.

Financial institutions and organizations dealing with sensitive customer data should consider implementing the BERT NER model for PII detection. It offers a powerful and effective solution for ensuring regulatory compliance, preventing fraud, and protecting customer privacy.

To enhance the accuracy of PII detection in the financial domain, further exploration of domainspecific pre-training and fine-tuning techniques is recommended. Incorporating financial ontologies or dictionaries and utilizing financial-specific training data can improve the model's performance on financial text datasets.

Balancing accurate PII detection with data privacy concerns is crucial. Future research should focus on developing and evaluating privacy-preserving techniques, such as differential privacy, data anonymization, and secure data sharing, to protect individuals' financial information while maintaining effective PII detection capabilities.

5.2 Summary of the Findings

The study's findings demonstrate the effectiveness of the BERT NER model for PII detection in the financial domain. The model successfully identifies and classifies various PII entities,

exhibiting high precision, recall, and F1-scores across multiple entity types, including names, addresses, and financial information. The average accuracy of the model indicates its overall strong performance in capturing different types of named entities within the given dataset.

5.3 Contributions to Knowledge

This study contributes to the existing knowledge in the field of PII detection by demonstrating the applicability of the BERT NER model for PII detection in the financial domain, highlighting the importance of domain-specific pre-training and fine-tuning techniques for improving the accuracy of PII detection in financial text datasets, emphasizing the significance of privacy-preserving approaches in balancing data security and accurate PII detection.

5.4 Suggestions for Further Study

While this study provides valuable insights into PII detection using the BERT NER model, there are several avenues for further research, conducting experiments and evaluations on diverse text datasets from various domains can provide a more comprehensive understanding of the BERT NER model's performance and generalizability across different contexts, Comparing the performance of the BERT NER model with other state-of-the-art PII detection models and techniques can help identify the strengths and limitations of different approaches, contributing to the development of more advanced and accurate PII detection methods, Conducting case studies and practical implementations of the BERT NER model in real-world financial institutions can provide insights into its feasibility, scalability, and effectiveness in a production environment. This can also shed light on potential challenges and best practices for integrating PII detection systems into existing workflows.

In conclusion, this study underscores the potential of the BERT NER model for PII detection in the financial domain. It offers valuable contributions to knowledge by highlighting the significance of domain-specific fine-tuning, privacy-preserving approaches, and the applicability of the BERT NER model in ensuring regulatory compliance and protecting customer privacy. Further research and exploration in this area can lead to advancements in PII detection techniques, enabling more robust data security measures in the financial industry and beyond.

BIBILOGRAPHY

Akhtyamova, L. [2020]. Named Entity Recognition in Spanish Biomedical Literature: Short Review and Bert Model.

Alahmari, A. and Duncan, B. [2020]. Cybersecurity Risk Management in Small and Medium-Sized Enterprises: A Systematic Review of Recent Evidence.

Aridor, G., Che, Y.-K., Salz, T., Acemoglu, D., Decarolis, F., Ellison, G. and Ellison, S. 2020. We would like to thank. Available at: https://piwik.pro/blog/privacy-laws-around-globe/. [Accessed: 1 August 2023].

Gunel, B., Du, J., Conneau, A. and Stoyanov, V. [no date]. SUPERVISED CONTRASTIVE LEARNING FOR PRE-TRAINED LANGUAGE MODEL FINE-TUNING.

Hendrycks, D. and Gimpel, K. [no date]. GAUSSIAN ERROR LINEAR UNITS (GELUS).

Hosein, M., Rizi, P., Amin, S. and Seno, H. 2022. A systematic review of technologies and solutions to improve security and privacy protection of citizens in the smart city. Available at: https://doi.org/10.1016/j.iot.2022.100584 [Accessed: 2 August 2023].

Kaster, M., Zhao, W. and Eger, S. [no date]. Global Explainability of BERT-Based Evaluation Metrics by Disentangling along Linguistic Factors. pp. 8912–8925.

Kim, D., Park, K., Park, Y. and Ahn, J.-H. 2018. Willingness to provide personal information: Perspective of privacy calculus in IoT services. Available at: www.elsevier.com/locate/comphumbeh [Accessed: 4 August 2023].

Kulkarni, P., Professor, A. and K, C.N. 2021a. Personally Identifiable Information (PII) Detection in the Unstructured Large Text Corpus using Natural Language Processing and Unsupervised Learning Technique. *IJACSA) International Journal of Advanced Computer Science and Applications* 12(9). Available at: www.ijacsa.thesai.org [Accessed: 7 August 2023].

Kulkarni, P., Professor, A. and K, C.N. 2021b. Personally Identifiable Information (PII) Detection in the Unstructured Large Text Corpus using Natural Language Processing and Unsupervised Learning Technique. *IJACSA) International Journal of Advanced Computer Science and Applications* 12(9). Available at: www.ijacsa.thesai.org [Accessed: 28 July 2023].

Luoma, J. and Pyysalo, S. [no date]. Exploring Cross-sentence Contexts for Named Entity Recognition with BERT. Available at: https://github.com/google-research/bert/issues/581 [Accessed: 06 August 2023].

Mahalle, A., Yong, J., Tao, X. and Shen, J. [no date]. Data Privacy and System Security for Banking and Financial Services Industry based on Cloud Computing Infrastructure.

Peng, Y., Yan, S. and Lu, Z. 2019. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. Available at: http://arxiv.org/abs/1906.05474.

Purtova, N. 2018. Law, Innovation and Technology The law of everything. Broad concept of personal data and future of EU data protection law The law of everything. Broad concept of personal data and future of EU data protection law. Available at: https://www.tandfonline.com/action/journalInformation?journalCode=rlit20 [Accessed: 12 August 2023].

Rana, R., Zaeem, R.N. and Barber, K.S. US-Centric vs. International Personally Identifiable Information: A Comparison Using the UT CID Identity Ecosystem.

Ryan, P., Crane, M. and Brennan, R. 2020. Design Challenges for GDPR RegTech. *ICEIS 2020 - Proceedings of the 22nd International Conference on Enterprise Information Systems* 2, pp. 787–795. Available at: http://arxiv.org/abs/2005.12138 [Accessed: 12 July 2023].

Silva, P., Gonçalves, C., Godinho, C., Antunes, N. and Curado, M. [2018]. Using NLP and Machine Learning to Detect Data Privacy Violations.

Vaswani, A. 2023. Attention Is All You Need.