The Learning Agency Lab - PII Data Detection

Authors

Mohammad Atif Quamar[1]* Abdullah Mazhar[2] Zeba Afroz[3]
IIIT Delhi
atif20523@iiitd.ac.in abdullahm@iiitd.ac.in zebaa@iiitd.ac.in

Abstract

This project aims to automate the detection of personally identifiable information (PII) in student essays, with a focus on cost-effective educational data handling. The objective is to facilitate learning science research and tool development by creating a model that can effectively annotate PII within a dataset. We explored various models and found that the DeBERTa ensemble yielded the most promising results. As a result, we achieved a global rank of 791 in a Kaggle competition with an accuracy of 95.24%.

Keywords: PII Detection, Privacy-preserving Information Extraction, spaCy, NER, DistillBERT, RoBERTa

1 Introduction

Modern organizations and government entities leverage online data collection to deepen their understanding of customers and enhance decision-making processes in an era increasingly defined by digitalization. This vast pool of data often includes Personally Identifiable Information (PII), which uniquely identifies individuals and can be classified into sensitive (e.g., Aadhar card numbers) and non-sensitive (e.g., nationality) categories. While sensitive PII necessitates encryption for protection due to its potential harm, disclosure of non-sensitive PII alone may not pose immediate risks. However, combining multiple non-sensitive PII elements could still lead to individual identification.

Accurate identification and secure management of sensitive PII are paramount, as mishandling may lead to data misuse, violating legal and compliance regulations. Concerns over potential data leaks during transfers stem from a need for greater transparency in PII handling and individuals' limited awareness of how their personal information is managed. According to IBM's cost of data breach study, mishandling sensitive data can result in substantial financial and reputational losses, amounting to up to $4.24 million per incident.

Inadequate PII management and security heighten the risk of identity theft, where criminals exploit personal information for fraudulent activities. Amid heightened focus on data security and privacy, there is a discernible trend toward adopting NLP techniques for PII masking. For example, healthcare providers leverage NLP to efficiently identify and mask sensitive patient information, ensuring compliance with privacy regulations while streamlining processes and reducing manual checking costs for PII in documents.

In our project, we conducted experiments across multiple epochs, commencing with RoBERTa achieving 91% accuracy, followed by DistilBERT at 78%, and subsequently DeBERTa v3 Small and DeBERTa v3 Large, attaining 84.1% and 96.6% accuracy, respectively. Ultimately, the DeBERTa ensemble proved most effective, achieving 96.7% accuracy and securing a global rank of 791.

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

2 Literature Review

An exploration of existing tools for detecting Personally Identifiable Information (PII) reveals a range of approaches and technologies employed across different organizations. For example, Microsoft's Presidio Analyzer integrates Named Entity Recognition (NER), regular expressions, rule-based logic, and checksums for identifying sensitive data, but it may encounter false positives due to specific patterns [1]. Nightfall Security, a cloud-native data protection platform, utilizes machine learning (ML) to detect and prioritize sensitive data [2], although its reliance on deep learning concepts can pose challenges in detecting unstructured data, necessitating users with a strong background in artificial intelligence and machine learning.

A data modeling and processing mechanism proposed in [3] emphasizes the extraction of precise features from low-quality data, proving more effective in PII detection than existing approaches, albeit limited to identifying full PII and unable to handle Quasi PII. In a comprehensive evaluation detailed in [4], the importance of dataset selection in NER models is highlighted through a comparison of five popular NER software packages. Another innovative approach, as described in [5], introduces an adaptive PII scanning and consent discovery system compliant with Thailand's Personal Data Protection Act, utilizing regular expressions and a consent validation algorithm across databases, web content, documents, and cookies.

Privacy-preserving personal information extraction is addressed in [6], utilizing Bidirectional Encoder Representations from Transformers-based NER models trained on WikiPII. This research underscores a human-in-the-loop model, aiding annotators in entity localization and tag recommendation to reduce manual data extraction costs. In [7], a Deep Transfer Learning (DTL) framework with Graph Convolutional Networks (GCNs) is developed for PII extraction from social media, specifically Twitter, although requiring manual annotation and a significant amount of labeled data for training.

Finally, [7] explores pseudonymization techniques across multiple datasets for text classification and summarization. While effective, careful selection of a pseudonymization system is essential to replace sensitive information with minimal impact on text quality.

3 Objective

The objective is to develop a model that automatically detects personally identifiable information (PII) in student writing. The submissions are evaluated based on a classification metric (F-beta) 1 that values recall and precision, emphasizing recall, by setting the beta value to 5.

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}} \tag{1}$$

In terms of evaluation metrics, the main objective is to create models that have a good balance between precision and recall but with a stronger focus on achieving high recall and effectively detecting personally identifiable information (PII) in student writing, emphasizing minimizing false negatives.

- **NAME_STUDENT:** Encompassing full or partial names of students, excluding instructors, authors, or other persons' names.
- **MAIL:** Represents a student's email address.
- **USERNAME:** Signifies a student's username on any platform.
- **ID_NUM:** Denotes a numerical or character sequence, potentially serving as a student identifier (e.g., student ID or social security number).
- **PHONE_NUM:** Corresponds to a phone number associated with a student.
- **URL_PERSONAL:** Represents a URL that may be linked to student identification.
- **STREET_ADDRESS:** Encompasses complete or partial street addresses related to students, including their home addresses.

4 Dataset

The dataset comprises 22,000 essay samples from students participating in a massively open online course (MOOC). These essays respond to a singular assignment prompt, prompting students to apply course concepts to real-world issues. To safeguard student privacy, original personally identifiable information (PII) has been substituted with surrogate identifiers of similar types. The test set retains 70% of the data. Notably, the dataset includes seven distinct PII labels as outlined below:

The dataset is in JSON format, containing key fields such as the document identifier, the complete essay text, a token list, details about whitespace, and token annotations. Token labels adhere to the BIO format, where the PII type is prefixed with "B-" for the beginning of an entity and "I-" for token continuations within an entity. Tokens devoid of PII are labeled as "O."

5 Methodology

The approach to identifying personally identifiable information (PII) in student writing entails a series of crucial steps. The dataset, consisting of 22,000 essay samples from a massively open online course (MOOC).

Our comprehensive methodology encompasses several stages: data preprocessing, model selection,tokenization, input representation, model training, and evaluation metrics. The overarching goal is to develop robust models capable of accurately identifying PII in student writing, emphasizing model performance and computational efficiency

5.1 Data Preprocessing

We implemented a series of data preprocessing steps tailored to support BIO sequence labeling. The preprocessing pipeline involved developing and utilizing two essential functions: first, to tokenize and second, to find span.

The tokenize function serves as a fundamental component of our preprocessing workflow. It accepts an example dictionary containing tokens and trailing whitespace lists alongside a tokenizer object. This function iterates through the tokens and their corresponding whitespace indicators, reconstructing the original text while constructing a token map linking each character to its respective token index. The reconstructed text is then tokenized using the specified tokenizer, incorporating options for truncation and a maximum length constraint inference max length. The output of the tokenize function comprises a dictionary containing the tokenized text and the associated token map, which preserves the alignment between tokens and their positions within the original input text.

Additionally, the find span function was instrumental in our data preprocessing pipeline for identifying sequences of tokens (referred to as target) within a more extensive list of tokens (document). This utility function initializes an empty span list and systematically traverses the document, seeking matches with the initial token of the target sequence. Upon encountering matching tokens, the function appends its indices to the span list. If the entire target sequence is found consecutively within the document, the resulting span list representing this sequence is added to the span list. This process iterates until all occurrences of the target sequence are detected within the document, with each identified span encapsulated as a list of corresponding token indices.

By integrating these preprocessing techniques into our research workflow, we effectively prepared and structured the textual data required for subsequent analysis and model development within the context of student essays and PII detection.

Algorithm 1: Tokenization Algorithm

```
Input   : example: dictionary, tokenizer: object
Output : dictionary
text ← empty list;
token_map ← empty list;
idx ← 0;
for t, ws in zip(example["tokens"], example["trailing$_w hitespace$”]) do
    text.append(t);
    token_map.extend([idx]*len(t));
    if ws then
        text.append(" ");
        token_map.append(-1);
    end
    idx += 1;
end
tokenized ← tokenizer("".join(text), return_offsets_mapping=True, truncation=True,
  max_length=INFERENCE_MAX_LENGTH);
return dictionary containing tokenized data and token_map
```

Algorithm 2: Span Finding Algorithm

```
Input   : target: list of strings, document: list of strings
Output : list of lists of integers (spans)
idx ← 0;
spans ← empty list;
span ← empty list;
for i, token in enumerate(document) do
    if token $\neq$ target[idx] then
        idx ← 0;
        span ← empty list;
        continue;
    end
    span.append(i);
    idx += 1;
    if idx == len(target) then
        spans.append(span);
        span ← empty list;
        idx ← 0;
        continue;
    end
end
return spans
```

5.2 Model selection

We have chosen four BERT-based models for BIO tagging and implemented an ensemble technique incorporating various DeBERTa large models—specifically deberta-models-cuerpo-de-piiranha, deberta-models-cola-del-piinguuino, deberta-models-cabeza-del-piinguuino, and deberta-models-piidd-org-sakura. RoBERTa is well-suited for this task due to its foundation on the BERT architecture, utilizing a bidirectional Transformer encoder to effectively capture contextual information crucial for Named Entity Recognition (NER). Pre-trained on a large text corpus with masked language modeling (MLM) objectives, RoBERTa learns word representations and their contextual relationships, proving beneficial for identifying named entities. DistilBERT is also suitable, being a distilled version of BERT designed for increased efficiency while maintaining performance. Despite its smaller size, DistilBERT effectively captures contextual dependencies essential for BIO tagging. DeBERTa-v3-small, a variant of DeBERTa, is adept at handling long-range dependencies and contextual nuances, essential for NER tasks, albeit smaller in size compared to DeBERTa-v3-large, which boasts increased model capacity and parameter count. This larger variant excels in capturing intricate contextual patterns crucial for accurate entity recognition. Lastly, the DeBERTa ensemble combines multiple DeBERTa model variants, leveraging their respective strengths to enhance overall performance, making it a practical choice for BIO tagging tasks by mitigating individual model weaknesses and capturing a broader spectrum of patterns and features for improved accuracy and robustness in entity recognition.

5.3 Fine-Tuning with Transformer Models

To enhance the precision of personally identifiable information (PII) detection, we conducted fine-tuning various transformer models. Utilizing the Hugging Face Transformers library, we tokenized datasets to ensure proper alignment of labels with tokenized inputs. The training process involved specifying training parameters, creating a trainer instance, and training the model. Key metrics such as recall, precision, and F1 score were computed during the evaluation.

5.4 Model Training

Our strategy for detecting personally identifiable information (PII) in student essays involves deploying a diverse array of state-of-the-art transformer-based models optimized for specific aspects of natural language processing (NLP) tasks. RoBERTa, an advanced variant of BERT (Bidirectional Encoder Representations from Transformers), employs a bidirectional Transformer encoder to effectively capture contextual nuances crucial for Named Entity Recognition (NER) tasks. By undergoing pre-training on extensive text corpora using masked language modeling (MLM) objectives, RoBERTa acquires deep insights into word representations and their contextual dependencies, achieving an impressive accuracy of 91% after ten epochs of rigorous training.

Moving forward, we leveraged DistilBERT, a compact iteration of BERT designed for enhanced speed and efficiency without compromising on performance integrity. DistilBERT achieves this feat by streamlining certain architectural components and reducing parameter counts, making it adept at capturing essential contextual dependencies vital for BIO tagging tasks, albeit achieving a slightly lower accuracy of 78% compared to more complex models like RoBERTa.

Transitioning to the DeBERTa models, our experiments focused on DeBERTa-v3-small and DeBERTa-v3-large, attaining 84.1% and 96.6% accuracy, respectively. DeBERTa, built upon BERT's foundation but specifically enhanced for managing long-range dependencies and contextual intricacies, showcases superior performance. The disparity in accuracy between DeBERTa-v3-small and DeBERTa-v3-large can be attributed to differences in model capacity; while smaller variants like v3-small may encounter limitations in capturing intricate data patterns, larger models such as v3-large excel in comprehending complex relationships within the data.

In our pursuit of optimizing accuracy and robustness, we culminated our efforts with a DeBERTa ensemble, incorporating various DeBERTa large models—namely deberta-models-cuerpo-de-piiranha, deberta-models-cola-del-piinguuino, deberta-models-cabeza-del-piinguuino, and deberta-models-piidd-org-sakura resulting in an impressive accuracy of 96.7%. This ensemble methodology effectively harnesses the collective strengths of multiple DeBERTa variants to elevate performance levels and bolster robustness in PII detection tasks. Through meticulous model selection and iterative refinement processes, we successfully optimized accuracy and secured a commendable global rank of 791, underscoring the remarkable efficacy of transformer-based models for advanced NLP tasks, particularly in the domain of PII detection within student writing.

5.5 Evaluation Metric(s)

The primary goal is to develop models for automatically detecting personally identifiable information (PII) in student writing, focusing on achieving a well-balanced performance between precision and recall. The chosen evaluation metric for this task is the F-beta score, where the beta value is set to 5, refer to eq 1, emphasizing a strong inclination towards recall over precision.

Regarding Type I and Type II errors, the evaluation strategy prioritizes minimizing false negatives (Type II errors) 2. This approach underscores the significance of correctly identifying personally identifiable information, such as names, emails, usernames, and other sensitive data, within the context of student submissions.

$$F_\beta = (1+\beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}} \tag{2}$$

The aim is to strike a balance between precision and recall, with an increased emphasis on recall to ensure effective detection while minimizing the risk of overlooking potential instances.

$$\text{Efficiency} = \frac{F_{\beta_5}(\text{Submission}) - F_{\beta_5}(\text{Benchmark})}{\max F_{\beta_5}} + \frac{\text{RuntimeSeconds}}{32400} \tag{3}$$

Additionally, the efficiency score for each submission serves as a comparative metric 3, assessing the submission's performance relative to a benchmark score, the maximum score attained by all submissions on the Private Leaderboard, and the time required for evaluation, denoted as RuntimeSeconds. The aim is to minimize the efficiency score, highlighting submissions that outperform the benchmark and peers, all while undergoing swift assessment. This emphasizes the significance of not solely prioritizing model accuracy but also considering the computational efficiency of the detection process.

6 Results

Table 1 presents a summary of the accuracy scores achieved by various BERT-based models that were evaluated in our BIO tagging task. The table encapsulates the performance metrics for each model, providing a comparative analysis of their effectiveness in accurately identifying and tagging named entities within the text data. This comprehensive evaluation enables us to assess and understand the relative strengths and weaknesses of the different BERT variants considered during our experimentation.

Sr. no.	**Model**	**Accuracy**
1	Deberta-ensemble-v1	0.967
2	Deberta-ensemble-v2	0.966
3	Deberta-v3-large	0.954
4	RoBERTa	0.916
5	Deberta-v3-small	0.841
6	DistillBERT	0.784

Table 1: Accuracies of different models

7 Analysis

Our analysis revealed that the ensemble approach yielded the highest performance, showcasing the effectiveness of leveraging multiple models. DistilBERT, with an accuracy of 78%, represents a streamlined version of BERT optimized for efficiency, albeit at the expense of some performance due to reduced model complexity. This reduction in capacity may limit DistilBERT's ability to capture intricate data patterns and dependencies compared to more comprehensive models like RoBERTa and more significant DeBERTa variants (such as v3-small and v3-large).

DeBERTa-v3-small, achieving an accuracy of 84.1%, is a compact version of the DeBERTa model tailored to handle long-range dependencies and contextual nuances. However, its slightly lower performance compared to DeBERTa-v3-large (95.4%) underscores the impact of model size and capacity on accuracy. More extensive models like v3-large possess a greater capacity to comprehend complex data structures, leading to enhanced performance in tasks like PII detection.

The success of DeBERTa-v3-large, with its significantly higher accuracy, emphasizes the critical role of increased model capacity in improving performance. Larger models can better learn and represent underlying data structures, making predictions more accurate.

The DeBERTa ensembles versions, boasting an impressive accuracy of 96.7% and 96.6%, respectively, amalgamate multiple models, potentially including variants like v3-small and v3-large. Ensemble methods capitalize on the diverse strengths of individual models to mitigate weaknesses and enhance overall performance. The superior accuracy of the DeBERTa ensemble compared to individual models like v3-large underscores the effectiveness of combining multiple models to capture a broader range of data patterns and features, ultimately resulting in more robust and accurate predictions.

8 Conclusion

In conclusion, our analysis underscores the significance of leveraging diverse models and ensemble techniques to achieve optimal performance in natural language processing tasks like PII detection. While streamlined versions such as DistilBERT offer efficiency, they may compromise on performance due to reduced complexity. Compact variants like DeBERTa-v3-small excel in handling specific nuances but are outperformed by larger counterparts like DeBERTa-v3-large, emphasizing the pivotal role of model size and capacity in enhancing accuracy. The remarkable success of the DeBERTa ensemble, with its amalgamation of multiple models, highlights the efficacy of combining strengths to mitigate individual weaknesses, resulting in robust and accurate predictions. Ultimately, larger and ensemble models prove instrumental in capturing intricate data patterns and features, leading to superior performance in challenging BIO tagging tasks.

9 Distribution of work among group members

- Mohammad Atif Quamar - Worked on DeBERTa-ensemble-v1 and DeBERTa-v3-small.
- Abdullah Mazhar - Worked on RoBERTa and DeBERTa-ensemble-v2.
- Zeba Afroz - Worked on Distillbert and DeBERTa-v3-large.

References

[1] Presidio. Presidio: Data protection and de-identification sdk, microsoft presidio. [Online]. Available: `https://microsoft.github.io/presidio/`

[2] Nightfall Security. [Online]. Available: `reference_link_for_nightfall`

[3] P. Kulkarni and N. Cauvery, "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*, vol. 12, no. 9, 2021.

[4] X. Schmitt, S. Kubler, J. Robert, M. Papadakis, and Y. LeTraon, "A replicable comparison study of ner software: Stanfordnlp, nltk, opennlp, spacy, gate," in *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*. IEEE, 2019, pp. 338–343.

[5] R. Hathurusinghe, I. Nejadgholi, and M. Bolic, "A privacy-preserving approach to extraction of personal information through automatic annotation and federated learning," *arXiv preprint arXiv:2105.09198*, 2021.

[6] Y. Liu, F. Y. Lin, M. Ebrahimi, W. Li, and H. Chen, "Automated pii extraction from social media for raising privacy awareness: A deep transfer learning approach," in *2021 IEEE International Conference on Intelligence and Security Informatics (ISI)*, 2021, pp. 1–6.

[7] O. Yermilov, V. Raheja, and A. Chernodub, "Privacy-and utility-preserving nlp with anonymized data: A case study of pseudonymization," *arXiv preprint arXiv:2306.05561*, 2023.