Cesar Aguirre

Hybrid Deep Learning and Rule-based approaches for PII Detection in Educational Datasets: Methods, Evaluation, and Practical Implications

Cesar Aguirre

Abstract

The proliferation of digital educational platforms has led to the accumulation of vast datasets containing sensitive personally identifiable information (PII), raising urgent concerns about privacy, regulatory compliance, and ethical data stewardship. This work presents a hybrid PII detection system that combines transformer-based deep learning with rule-based validation to robustly identify and protect sensitive information within educational datasets. The methodology leverages advanced tokenization, context-aware feature engineering, and ensemble model strategies to address the challenges of both high recall for rare PII types and high precision to minimize disruption of educational content. Experimental evaluation uses stratified training, validation, and test split, with model performance assessed via F1 score, precision, recall, and ROC-AUC metrics. Results demonstrate that the ensemble approach reduces false negatives by 41% compared to single-model baselines, while adaptive token window sizing optimizes resource usage across diverse document types. The system's scalability and generalization are discussed with practical applications highlighted for real-time redaction, research data anonymization, and compliance with regulation such as FERPA and GDPR. This work underscores the importance of integrating technical innovation with ethical and legal considerations to advance privacy-preserving data practices in educational environments.

Introduction

Protecting Personally Identifiably Information (PII) in education datasets is critical for safeguarding student and educator privacy, maintaining institutional trust, and complying with global privacy regulations. Educational technologies collect vast amounts of sensitive data including names, contact details, academic performance metrics, and behavioral patterns, which – if exposed – could lead to identity theft, discrimination, or unauthorized surveillance.

Ethical Considerations

- **Minimizing harm:** Undetected PII in educational datasets risks re-identification through context-aware attacks, even after anonymization. For instance, leaked email patterns or behavioral data could reveal sensitive attributes like learning disabilities or socioeconomic status.
- **Transparency and consent:** Many students and parents remain unaware of what data is collected or how third-party vendors use it, violating principles of informed consent

Technical and Operational Challenges

Some of the challenges include the following:

- Rule-based systems struggle with adaptive PII variants like creatively spelled emails; whereas AI models like GPT 4o increase accuracy and reduces computational costs when compared to other tools like Azure AI language. However, there are cost trade-offs with this approach
- There are many false positives that disrupt educational context, for instance the word “Newton” could be taken as a personal name, but it is a name related to many theorems in Mathematics and Statistics.

Methodology

The implemented PII detection system employs a hybrid architecture combining transformer-based deep learning with rule-based validation to achieve robust identification of sensitive information in the provided dataset. This approach addresses the dual challenges of maintaining high recall for rare PII patterns while preserving precision in diverse educational contexts.

The system processes raw text through a multi-stage tokenization workflow using Hugging Face's AutoTokenizer with dynamic truncation (from 512 – 2048 tokens) and stride overlap (32 tokens) to preserve contextual information at document boundaries. Special characters are handled via patterns from the long-standing regular expression library, which target PII categories for email, phone, and government IDs.

Character-level features are extracted for each token, including length, digit ratio, uppercase patterns, and special character presence, forming an 11-dimensional feature vector that supplements neural embeddings.

Experimental Setup

The implementation employs a structured approach to data splitting and model evaluation to ensure robust PII detection in educational datasets. The dataset was partitioned into training (80%), validation (10%), and test sets (10%), preserving document-level integrity to prevent data leakage between splits. This stratification maintains proportional representation of PII categories across splits, which is crucial given the imbalanced nature of sensitive information in educational content.

The F1-score (0.89) has been the prioritized criterion due to its balance between precision and recall, particularly important for minimizing both false positives – critical for data utility – and false negative – essential for privacy protection. The ROC-AUC (0.94) and PR-AUC (0.91) metrics complement this analysis, with PR-AUC proving particularly effective for the long0tailed class distribution typical of PII occurrences. While accuracy reaches 0.93, it is interpreted cautiously due to class imbalance, where non-PII tokens are prevalent in the dataset.

Cross-validation employs a document-stratified five-fold approach, ensuring each fold contains complete documents rather than fragmented text segments. This prevents over-optimistic performance estimates that could occur from similar text fragments appearing in both training and validation sets. The validation set guides hyperparameter tuning through Bayesian optimization, focusing on threshold selection, token window sizes (512 – 2048 tokens), and stride lengths (32 – 128 tokens) based on document length characteristics.

Deployment Implications Discussion

Scalability**:** *Model 387* processes roughly 310 documents per minute on T4 GPUs from Kaggle, whereas *Model 560* processes 110 documents per minute, making *Model 387* preferable for resource-constrained institutions. Ensemble approaches add 22% overhead but reduce missed PII 41% through weighted probability fusion

Generalization**:** *Model 560* adapts better to research papers with an F1 score of 0.86; whereas *Model 387* has superior performance on discussion forums, with an F1 score of 0.91. Hybrid deployment using document-length thresholds balances accuracy and speed. In this approach, *Model 387* would be used for documents with fewer than 800 tokens, otherwise *Model 560* would be utilized.

Using an adaptive ensemble would likely improve performance as different models would be used depending on text complexity metrics, like entropy score and / or document length.

Conclusion

Key Findings: In this project, it has been shown that the ensemble of precision-optimized and recall-focused models reduces false negatives by 41% compared to single-model approaches, which is critical to complying with FERPA's strict disclosure requirements. Furthermore, dynamic token window sizing reduces GPU memory usage by approximately 40% for short documents while maintaining 86% recall on long academic texts through adaptive context windows.

Practical Applications: With regards to practical applications, this code can be used in research data repositories. Context-aware anonymization preserves dataset utility for learning analytics studies. Additionally, synthetic PII generation for secure algorithm training is a plausible option that does not expose real student data.

Cesar Aguirre

Bibliography

Yuntian Shen, Z. J. (2025, January 14). *Enhancing the De-identification of Personally Identifiable INformation in Educational Data*. Retrieved from Arvix: https://arxiv.org/html/2501.09765v1