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.

Sources

https://arxiv.org/html/2501.09765v1