PIILO: An open-source system for personally identifiable information labeling and obfuscation

Langdon Holmes[1], Scott Crossley[2], Harshvardhan Sikka[3], and Wesley Morris[4]

[1,2,4] Vanderbilt University

[3] Georgia Institute of Technology

Author Note

Langdon Holmes https://orcid.org/0000-0003-4338-4609

Scott Crossley https://orcid.org/0000-0002-5148-0273

Harshvardhan Sikka https://orcid.org/0000-0002-4006-6659

Wesley Morris https://orcid.org/0000-0001-6316-6479

Individuals can access a demonstration of PIILO hosted by Huggingface Spaces (https://huggingface.co/spaces/langdonholmes/piilo). Software code used to develop the transformer models is hosted by Github (https://github.com/aialoe/deidentification-pipeline-archived). We have no conflicts of interest to disclose.

Correspondence concerning this article should be addressed to the first author.

Abstract

Education is increasingly taking place in technologically mediated settings, making it easier to collect valuable data for learning analytics. However, much of this data is not available to the research community due to concerns about protecting student privacy. Deidentification of student data might address this concern, but deidentification is difficult for unstructured data such as student-generated text. This study reports on an automatic deidentification system for personally identifiable information labeling and obfuscating (PIILO) in student-generated text. The system labels student names using a fine-tuned large language model based on Longformer (Beltagy et al., 2020). The model was developed with a dataset of 5,797 student essays that were human labeled for PII. The model recalled 84% of student names on a held-out testing set. A second model was developed to deidentify other direct identifiers (phone numbers, URLs, e-mail addresses) in PIILO using pattern matching. A combined labeling system automatically detected 75% of direct identifiers in a second dataset of 2,118 classroom discussion board posts that were human labeled for PII. The identifiers in the second dataset were obfuscated using a replacement strategy called hiding-in-plain-sight (HIPS, Carrell et al., 2013, 2019), which replaces labeled identifiers with artificially generated surrogates of the same type, making it difficult to distinguish them from any remaining residual identifiers. In a simulated reidentification attack, experts recovered less than 25% of residual identifiers in HIPS-obfuscated data. The automatic approaches to text deidentification developed in this study and released in PIILO present a low-cost alternative to manual deidentification, making PIILO ideal for situations in which data would not normally be de-identified, such as sharing data between lab members or within an institution.

Keywords: privacy, deidentification, anonymization

PIILO: An open-source system for personally identifiable information labeling and obfuscation

Learning science heavily relies on open-source student data so that hypotheses and interventions can be assessed by teams of researchers through multiple theoretical and methodological lenses. Although learning technologies make it easier to collect and analyze large amounts of student data, there are growing concerns surrounding student privacy and data rights (Pardo & Siemens, 2014; Slade & Prinsloo, 2013). Researchers agree that ethical practice in learning science requires careful attention to student privacy (Rubel & Jones, 2016), but there is active discussion about what it means to protect student privacy, why it should be protected, and how best to maintain privacy.

Student privacy in the context of higher education has been linked to intellectual privacy, which is the idea that students participating in higher education are entitled to explore ideas without fearing the scrutiny of public exposure (Jones et al., 2020). This conceptualization of student privacy is particularly relevant to student writing, where students are likely to express ideas, opinions, and experiences that might make them intellectually vulnerable. It is also likely that anonymization, which is often considered the gold standard for protecting privacy, may not sufficiently protect students’ intellectual privacy. Even if students understand that their writing will be anonymized, they may still experience some fear of scrutiny should their ideas be shared publicly. As a result, utilizing student-generated data will always carry risk, even under ideal circumstances.

Ideal circumstances in this context refer to data that has been collected with informed consent from the student and is fully anonymous. Consent plays an important role in data practices, not only because anonymization alone is insufficient to protect all forms of student privacy but also because perfect anonymization cannot be guaranteed. Recent research has emphasized that consent must be “informed”, which entails providing students with specific guidance about how their data will be collected and analyzed before consent is requested (Sun et al., 2019). Informed consent may also require providing students with choices about how their data may be used: for what purposes and for how long (Jones, 2019). It has also been argued that consent should be revokable or actively reaffirmed by the student on an ongoing basis (Young, 2015). While each of these strategies plainly provides the student with greater authority over the data they produce, not all of them are tenable solutions for every type of data collection, usage, and sharing plan. For example, it would not be possible to allow students to revoke consent once data has been shared publicly because there is no method of un-sharing a publicly released dataset. Public sharing of student data is a consequential scenario because shared datasets present many valuable opportunities for replication research and, in the case of predictive analytics, the ability to benchmark new predictive models. This is a strategy that has been used to great effect by the natural language processing (NLP) scientific community, though it has been met with some criticism (Parra Escartín et al., 2017).

Another critique of the consent-focused model of data collection is that it can result in sampling bias because not all student demographics are equally likely to consent to data collection (Cormack, 2016). Sampling bias in this context could ultimately lead to lower quality analyses and interventions because of lower representation of certain student groups, which would have a negative social impact that could be particularly harmful for those groups who are least likely to provide consent (Li et al., 2022). To alleviate the problem of sampling bias, Cormack suggests that consent might be obtained at the moment of intervention (“May we suggest reading materials based on your reading habits?”) rather than the moment of analysis

("May we analyze your reading habits?"). Cormack provides a framework for learning analytics that relies on safeguards to protect student privacy during analysis, such as anonymizing individual data to the extent that is practical. Cormack points out that the safeguards used must be "appropriate to the level of risks to privacy and other interests" (p. 104). This framework acknowledges that risk is inherent to data collecting and analysis, but also asserts that some level of risk is acceptable when it can be justified by a benefit to the data subject or to the social good. When there is a clear benefit to utilizing individual data, anonymization is a powerful tool for reducing the risk to the data subject.

Anonymization

Anonymization can reduce the risk profile of using individual data by making it infeasible to link the data subject to their data. Anonymization is generally required for the development of large, open-sourced datasets, which can lead to the types of replicable and generalizable studies needed to inform practices. The objective of anonymization is to obfuscate all confidential information while retaining as much non-confidential information as possible. This challenge is referred to as the privacy-utility tradeoff (Hassan et al., 2021) because increasing the protection to participant privacy tends to decrease the utility of the data by making it less authentic or realistic. Fully anonymizing data while retaining its utility for research purposes is an exceedingly challenging task, whether performed manually or by automated methods. Manual anonymization as prescribed by legal frameworks such as the European Union's General Data Protection Regulation (GDPR) is costly and time-consuming (Megyesi et al., 2018). Automated anonymization is also difficult, particularly for unstructured data such as text, audio, and video. Whereas structured data has well-defined fields (such as a database table with a "phone number" column), unstructured data does not contain constrained fields. As a result, it is difficult to identify which parts of an unstructured data object constitute private information.

For unstructured text, anonymization is a two-step process that involves the labeling of personally identifiable information (PII) and its subsequent obfuscation. The first step is to label the PII in the data, which can take the form of direct identifiers or quasi-identifiers (Lison et al., 2021). Direct identifiers are variables that are unique to a specific person and could be independently used to identify that person, such as their name, social media profile, or home address. Quasi-identifiers are information about a person that could not independently be used to identify that person, such as their gender, university, date of birth, or city of residence. However, a combination of quasi-identifiers could be used to reidentify a person, so these identifiers may also need to be labeled.

The second step in data anonymization is to obfuscate the labeled identifiers. The simplest obfuscation strategy is to delete identifiers from the dataset. Another, more common, strategy is to replace identifiers with a label (typically the type of data point, such as “student_name” in place of “John Doe”) or a unique identifier that cannot be traced back to the original individual. A third strategy is known as hiding in plain sight (HIPS) (Carrell et al., 2013). HIPS replaces identifiers with artificially generated data points of the same type. One benefit of HIPS is that obfuscated identifiers have the same form as leaked identifiers, making it difficult to discern between identifiers that were artificially generated and identifiers that were accidentally leaked.

Full anonymization of data requires obfuscating all direct identifiers and enough quasi-identifiers to prevent reidentification of any participant. In practice, full data anonymization poses significant challenges, such that some researchers have questioned whether it is even possible (Weitzenboeck et al., 2022). Deidentification is a more narrowly defined form of anonymization that seeks to obfuscate a predefined set of direct identifiers in a dataset (Lison et al., 2021; Pilán et al., 2022). For example, HIPAA (The Health Insurance Portability and Accountability Act) defines a closed set of 18 direct identifiers (patient names, email addresses, etc.) which must be obfuscated before medical patient data can be shared in the United States (HHS, 2012). While deidentification cannot eliminate identity disclosure risk, due to the presence of quasi-identifiers, it can dramatically reduce this risk to a level that is acceptable in some contexts. Differing legal standards, types of data, and contexts each demand their own strategy for protecting participant privacy.

Given the size of modern datasets, manual deidentification of text is impractical, driving the need for automatic solutions. However, automatic labeling of PII in unstructured text continues to be a challenge. While some direct identifiers follow predictable orthographic patterns, other types of identifiers are more difficult to detect. Assuming standard formatting, email addresses can be extracted using a character-based pattern that looks for a contiguous string of characters before an “@” sign followed by a domain name. On the other hand, names cannot be reliably identified with only character-based patterns. A system that can automatically detect names in unstructured text must recruit linguistic information beyond the character level, so the task requires modeling how words are used in context. Moreover, detecting the names of data subjects requires distinguishing them from other names that may appear in the text, such as public figures and referenced authors. Even though deidentification is difficult, we know that unstructured text can serve as a rich source of information for understanding human behavior. Thus, an automated system for text deidentification would enable the production of open-source datasets that can be used to test, develop, and modify technologies which increasingly rely on data-hungry machine learning algorithms to drive interventions.

Approaches to Text Deidentification

Deidentification of text has historically been carried out by hand, despite the significant costs in terms of infrastructure and human labor (Dorr et al., 2006). The development of automatic text deidentification systems became an active area of research in the early 2000s. For instance, in 2007, 18 teams competed in a shared task to deidentify medical discharge summaries in the i2b2 corpus (Uzuner et al., 2007). While work on medical text has continued, relatively little attention has been devoted to automatic deidentification of educational data. Currently, most educational researchers employ human annotators to label PII before releasing educational datasets (Crossley et al., 2022, in press; Megyesi et al., 2018). However, the success of deidentification systems in the medical domain suggests that it may be possible to develop automated text deidentification systems for other contexts as well.

Automated deidentification systems have been developed using rule-based and deep-learning-based approaches. Rule-based approaches to automatic PII labeling use a combination of rules to identify different types of PII. These approaches have been applied to text deidentification in different domains. Lison et al. (2021), for instance, used a rule-based system to achieve an F_1 of 81% on per-token named entity labeling in a corpus of Wall Street Journal articles. Bosch et al. used a rule-based approach to achieve as high as 95% recall of student names on student-generated text collected from a university classroom's discussion forum (Bosch et al., 2020). These systems have demonstrated strong performance on complex PII-labeling tasks, but a disadvantage is that these systems may not generalize well to text in other datasets because they rely on specific textual features that are not universal to all text types.

More recent approaches to automatic text deidentification have used deep-learning techniques, typically transformer-based language models. A transformer model is a neural network architecture widely used in natural language processing (Vaswani et al., 2017). These models have become state of the art for many language-related tasks, including named entity recognition (NER). NER seeks to extract a defined set of named entities from a text and classify them into types such as locations, organizations, or people. PII labeling can be formulated as a sub-problem of the larger NER task that naturally builds on this work. Previous work has shown success applying transformers to text deidentification in the medical domain, achieving PII recall as high as 99% on some medical datasets (Chambon et al., 2023; Murugadoss et al., 2021). However, an important drawback of these transformer-based approaches is that they require significant amounts of labeled training data, which is typically produced by manual annotation of documents. They also require specialized computational resources, which is especially relevant for private data that may need to be deidentified 'locally' on computers with limited resources. A final consideration is that deep-learning based approaches may exhibit differential performance for different demographic groups (Mansfield et al., 2022). Despite these drawbacks, transformer-based systems are likely to be the most effective approach for labeling complex forms of PII. However, their efficacy has not been tested in the educational domain.

Current Study

We introduce an open-source automatic deidentification system for student text called the Personally Identifiable Information Labeling and Obfuscation (PIILO) system. A web-based demonstration of PIILO and its full code base are available online.[1] PIILO is openly licensed, so researchers may use, review, and modify the code freely. PIILO includes both transformer-based

[1] anonymous.4open.science/r/piilo-E2CE

and rule-based systems for labeling PII in student text. One of the core design principles behind PIILO is that obfuscation is as important as labeling PII in text deidentification systems. PIILO implements obfuscation by way of hiding in plain sight (HIPS). It uses a surrogate name generator that automatically obfuscates student names with realistic and contextually plausible surrogate names. While we do not propose a complete solution for protecting student privacy, we envision that PIILO in its current form can lessen the barriers associated with sharing data when it is used in combination with established practices such as obtaining informed consent and manual anonymization workflows. We hope also to call attention to automated text deidentification as a promising research area in educational data settings. Our argument is that recent advances in natural language processing have made this problem tractable.

This investigation seeks to assess the ability of PIILO's subsystems to protect student privacy in two common forms of student-generated text: written essays and discussion posts. The following research questions will be addressed in three studies, where each study is presented in its own section:

- Research Question 1: How effective is PIILO's finetuned transformer at labeling student names in written essays?
- Research Question 2: How effective is PIILO's deidentification system at labeling PII in forum discussion posts?
- Research Question 3: How effective is PIILO's hiding-in-plain-sight obfuscation strategy at protecting student privacy?

Datasets

We analyzed two datasets to address our three research questions. The first dataset focused on student writing collected from a massive open online course (MOOC). The second dataset comprised forum discussion posts written by university students within a learning management system.

MOOC Dataset

Our first dataset comprised student writing samples that were collected from learners enrolled in a MOOC offered by a large university in the United States. The topic of the course was critical thinking through design. The course covered thinking strategies intended to help students solve real-world problems, such as storytelling and visualization. Course duration was estimated by the content provider to be 6 hours, and all materials were presented in English. During the data collection period from May 2016 to April 2022, 367,788 students enrolled in the course and 39,118 students completed the course.

The course included lecture videos, a discussion forum, and learning assessments. To complete the course, students were required to submit a written essay in which they reflected on how the course content could be applied to a problem they are familiar with. Submissions were required to be in PDF format, and the files were retained on a third-party hosting platform. In total, 221,043 submission events were recorded, including multiple submissions from the same students.

To assemble a corpus of student writing that contained PII, we carried out a series of data cleaning steps. We first selected all submissions that had been graded. In some cases, the same student had multiple graded submissions, which made it unclear which submission was associated with the overall course grade. As a result, we excluded users who had multiple graded submissions from the study, resulting in 38,267 viable submissions. Each submission was associated with a download hyperlink. We further excluded files that did not have a valid hyperlink, were larger than 10 megabytes, or longer than 5 pages. The latter criterion was established to exclude submissions that were clearly irrelevant documents (such as publicly available dissertations likely not written by the student). The remaining 32,525 assignment submissions were downloaded.

All downloaded submissions were automatically parsed using a PDF parsing library (Artifex, 2022), which converted the PDF file to plain text format. If the file was parsed without errors, the text was passed through the Chromium language detection algorithm (Sites, 2022). This step was carried out to remove non-English submissions and submissions that were corrupted by the parsing library. Lastly, we excluded any submissions with less than 50 words (whitespace-delimited tokens) because these were likely the result of a PDF parsing error and would not contain enough language for analysis. After removing documents based on these criteria, the resulting corpus contained 29,152 plain text documents. Of these, we selected documents from users that also contributed to the discussion boards (N=3,216) to annotate for PII. We randomly selected an additional 3,076 documents for annotation leading to a total of 6,293. As a final step, we removed duplicate documents, which occur when multiple students submit the same essay. After removing duplicates, 5,797 documents remained (see Table 1 for pruning details).

Table 1

Corpus size after applying each processing step from top to bottom.

	Count Remaining
Submission Events	221,043
Graded Submissions	44,593
Graded Submissions Unique to User	38,267
Valid URL	32,525

	Count Remaining
Parsed to English Text	29,142
Labeled for PII	6,293
Deduplicated	5,797

Discussion Post Dataset

Our second dataset comprised forum discussion posts from a learning management system collected from students enrolled in a computer science course at a large university in the United States. The course covered topics in knowledge-based artificial intelligence, and all students who were enrolled in the course during the data collection period were degree-seeking. Students were awarded participation points for posting on the discussion board, but students also had the option to earn participation credit in other ways, so posting was not a course requirement. Students used the discussion board to converse with each other, pose questions to teaching assistants, and share their reactions to course materials. There were daily discussion threads (reflections on Monday, debates on Wednesday, humor on Friday) that elicited a wide range of content and style from student posts.

During the data collection period, there were 227 students enrolled in the course, and 220 participated in the discussion forum. There were 4,328 posts with an average word count of 57, a minimum of 1 word and a maximum of 1,994 words. The forum discussion posts were exported in a plain text format which did not require any additional cleaning steps. Two thousand forum discussion posts were randomly selected for annotation and analysis.

Manual Annotation of PII

All texts from the MOOC dataset were annotated for student names by undergraduate research assistants following annotation guidelines. Annotators were instructed to apply labels liberally if they encountered any names that might be used to identify a student. After documents were labeled, the annotations were reviewed by a third, expert rater to ensure accuracy and consistency. The review process was primarily subtractive and required making judgements with limited information. For example, essays occasionally included named personas to illustrate a user interacting with a product, and it was not always possible to discern with certainty whether the name referred to a real student or a persona. In such cases, labels were retained following the reasoning that it would be better to erroneously label and obfuscate some non-private information than to risk the disclosure of a student's identity through a PII leak. Of the 5,797 submissions in the MOOC dataset, 845 included student names, and there were 1,155 student name annotations in total.

All texts in the Discussion Post dataset were annotated for student names following the same instructions as in the MOOC dataset. However, annotators also labeled additional types of direct identifiers in the posts, including email addresses, phone numbers, URLs, and street addresses. Within the Discussion Post dataset, there were 417 student names, 103 URLs, 3 email addresses, 3 usernames, 1 street address, and 0 phone numbers across 2,118 forum posts.

Study One – Transformer-based System for Labeling Student Names

Our first study assessed the potential for a transformer-based system to label student names in a homogenous collection of student writing samples. We developed a transformer-based student name labeling model (PIILO) to automatically annotate student names, which are a particularly challenging type of identifier to label. Pre-trained transformer-based language models, sometimes called foundation models or simply large language models (LLMs) are neural networks that have been trained on a language modeling task using large quantities of text scraped from the internet. They rely on the transformer, a neural network architecture that has been demonstrated to be effective for a variety of tasks in natural language processing (Vaswani et al., 2017). The weights and biases of the pre-trained neural network can be downloaded and finetuned on a smaller quantity of labeled text data. This workflow, an example of transfer learning, has resulted in state-of-the-art performance for named entity recognition and other natural language tasks.

Method

The labeled MOOC dataset, consisting of student essays, was split into training, development, and testing partitions, which comprised 60%, 20%, and 20% of the data, respectively. The training set was used exclusively to finetune our transformer-based model. The development set was used for validating the model during development. The testing set was used to evaluate the performance of our student-name labeling model. There were 207 student names comprising 470 tokens in the out-of-bag testing set. Of these, 38 were first names by themselves, and 6 names used an initial for the first name, the last name, or both. The remaining 163 names were full names, containing at least a first name and a last name.

We developed our student name labeling model from Longformer, a transformer-based, pre-trained large language model (Beltagy et al., 2020). We selected Longformer for two reasons: it uses an encoder-only architecture, and it has a relatively long maximum sequence length. The original transformer (Vaswani et al., 2017) included two distinct components: the encoder and decoder. Since then, many popular models have been developed that use only one or the other of these components. Encoder-only models such as Bidirectional Encoder

Representation from Transformers (BERT; (Devlin et al., 2019) lend themselves to sequence classification and token classification tasks straightforwardly because they can output token and document embeddings that are informed by the full document context. While decoder-only, “generative,” models such as the Generative Pre-trained Transformer (GPT)-series (Brown et al., 2020) have recently seen a great deal of attention, there have been limited applications of these model types to token classification tasks (but see Yan et al., 2021). Our second reason for selecting Longformer is that it has a maximum sequence length of 4,096 tokens. Most encoder-only models are limited to 512 tokens, which would not be long enough to process longer writing samples like the student essays in the MOOC dataset in one pass.

The Longformer model was finetuned using SpaCy. Like most LLMs, Longformer uses a tokenizer that can divide words into so-called sub-word tokens. In order to generate predictions at the level of the SpaCy token, we used a reduction approach called mean pooling. If a SpaCy token consists of multiple Longformer tokens, the vector representations for these sub-word tokens are averaged before a label is predicted. In this way, our labels and predictions are handled using the SpaCy tokenization scheme, which roughly corresponds to whole words or whitespace-delimited tokens. To train the model, we used the Adam optimizer with a learning rate of 0.00005 and 250 warmup steps. Model performance was evaluated using an F_β measure, where recall is treated as β times more important than precision. We set β to 9 to emphasize recall over precision during model validation. Model validation occurred after processing every 200 documents during the training phase. We ended finetuning when 10 of these evaluation steps had gone by (2,000 documents) without any improvement in performance. After finetuning, the version of the model that performed best on the validation set was selected as the fully developed model.

We assess the performance of PIILO on the testing partition of the written assignment dataset. All results are calculated on a per-token basis. For example, if the span “Carlos de Campos” was labeled as a student name, and our system labeled only “de Campos”, that would be counted as one false negative and two true positives on a per-token basis. This metric is straightforward to calculate and allows the model to get “partial credit” for correctly labeling part of an identifier (Wang et al., 2021). Recall is the proportion of true student name tokens that were successfully identified. Precision is the proportion of predicted student name labels that were correct. F_1 is the harmonic mean of these two values, which measures overall performance.

To contextualize the performance of PIILO model, we report results alongside two other models: SpaCy’s large English NER model (Explosion AI, 2022) and the Stanford Deidentifier (Chambon et al., 2023). SpaCy’s large model was trained on the OntoNotes5 corpus (Weischedel et al., 2013) to detect all person names, rather than just students’ names. Since all student names are also person names, true positives and false positives can be interpreted straightforwardly. False negatives may be person names even though they are not students’ names. The Stanford Deidentifier was trained on a corpus of radiology reports. It predicts two labels that correspond to person names, “healthcare worker” and “patient.” Similar to the SpaCy model, we map both of these labels to the student name label.

Results

We evaluate the performance of PIILO’s labeling mechanisms in terms of per-token classification accuracy. Results are reported on the held-out testing set. We also report performance using SpaCy and the Stanford Deidentifier (see Table 2 for results). On the in-domain testing set, PIILO labeled 84% of student name tokens with a precision of 68%, resulting in an F_1 of .75. This is an improvement over SpaCy’s general-purpose NER model, which reported an F_1 of .35 on this dataset. This is also stronger performance than the Stanford Deidentifier, which reported an F_1 of .50 on this dataset.

Table 2

Student name detection performance on the MOOC dataset.

	True Positive	False Negative	False Positive	Precision	Recall	F_1
SpaCy Large	325	145	1,080	.23	.69	.35
Stanford Deidentifier	329	141	524	.39	.70	.50
PIILO	393	77	184	.68	.84	.75

Discussion

PIILO recalled 84% of student name tokens on the held-out testing portion of the MOOC dataset. This recall was substantially higher than both the Stanford Deidentifer and SpaCy model on this dataset. This is true even though the Stanford Deidentifier and SpaCy model were trained on much larger datasets. However, 84% is still lower than the Stanford Deidentifier's performance on radiology reports, where it achieved 96% recall of patient names. These results highlight the need for deidentification models and systems that are specifically designed for the educational domain.

False negatives resulting from PIILO, or leaked identifiers, were primarily first names. These should be seen as less concerning than full names, because first names are generally not unique to an individual. As a result, they can only be used to reidentify a student when combined with other information, such as the individual's employer. PIILO failed to label six out of 163 full names (13 tokens), which results in an accuracy of 96%. The discrepancy between these metrics and the token-level metrics reported for all models is partially explained by multi-token names. One leaked name contained a space between each character of the name (as in "S a m u e l"), which was likely an artifact of the PDF parsing process. This was a full name that PIILO failed to label, but it was counted as 10 tokens. Overall, PIILO’s student name labeler outperformed other systems on the MOOC dataset, and its performance was better for full names than for partial names, but recall was not perfect.

Study Two – Forum Discussion Posts

Our second study assessed PIILO’s performance at labeling PII in forum discussion data, which is another important type of student-generated text. In this case, we test PIILO’s ability to label student names as well as other direct identifiers: URLs, usernames, email addresses, street addresses, and phone numbers. We integrated a rule-based PII-labeling technique into PIILO to complement the transformer model developed and tested in Study One.

Method

All direct identifiers except for student names were labeled by a rule-based analyzer, which we implemented using the Presidio anonymization library for the Python programming language (*Presidio - Data Protection and Anonymization API*, 2018/2022). The pattern matching systems are regular expressions that look for combinations of characters in the text. For example, three digits and four digits separated by a hyphen (“258-9713”) would match a phone number (and similar patterns exist for other common phone number formats). These basic patterns are effective for identifier types that have a predictable and distinctive structure. The full set of identifiers include country-specific national identifiers (including United States Social Security numbers) and is available in the Presidio documentation.[2] Since no models were trained using this data, results are reported on the full dataset.

[2] microsoft.github.io/presidio/supported_entities/

Results

We evaluate PIILO on all 2,118 forum posts, again comparing the performance of its student name labeling model to the SpaCy large model and the Stanford Deidentifier. We then report the labeling performance of PIILO on other direct identifiers including URLs, usernames, email addresses, street addresses, and phone numbers. No comparison was possible with SpaCy and the Stanford Deidentifier because they were not trained to label these types of identifiers.

Student names

Student name labeling performance of the three models in the Discussion Post dataset are reported in Table 3. Both the SpaCy and Stanford Deidentifier models achieved higher recall of student names on the Discussion Post dataset than the MOOC dataset. In terms of recall, they also outperformed PIILO, but they underperformed in terms of precision. In terms of F_1, PIILO reports .71, which is an improvement over the Stanford Deidentifier's F_1 of .52 and the SpaCy model's F_1 of .50.

Table 3

Student name detection performance on the Discussion Post dataset.

	True Positive	False Negative	False Positive	Precision	Recall	F_1
SpaCy Large	520	98	933	.36	.84	.50
Stanford Deidentifier	523	95	857	.38	.85	.52
PIILO	433	185	156	.74	.70	.71

Other direct identifiers

We further evaluate PIILO's ability to label other direct identifiers (URLs, usernames, email addresses, street addresses, and phone numbers). Table 4 shows the number of true positives, false negatives, and false positives for each identifier type. PIILO recalled 89% of personal URLs but also recalled many URLs that were not labeled as PII, resulting in a precision of 24% and an F_1 of .38. PIILO recalled all 3 emails with perfect precision. The current version of PIILO does not include any systems for labeling street addresses or usernames, so there are no false positives for these types. Two street addresses and two usernames were counted as false negatives. An additional two usernames were labeled as student names and are counted as true positives.

Table 4

PII detection performance of PIILO on the Discussion Post dataset.

	True Positive	False Negative	False Positive	Precision	Recall	F_1
Personal URL	108	13	350[a]	.24	.89	.38
Username	2[b]	2	-	-	.50	-
Email	3	0	0	1.0	1.0	1.0
Street Address	0	2	-	-	0.0	-
Phone Number	0	0	0	-	-	-
Total	113	17	350	.27	.87	.41

Note: Precision and F_1 are only defined when false positives can be calculated.

[a] False positive URLs include both entities that are not URLs (e.g., *main.py*), and URLs that could not be used to identify a student (e.g., *en.wikipedia.org/wiki/Python*).

[b] Two username tokens were labeled as student names and are counted as true positives.

Discussion

PIILO recalled 74% of student name tokens in the Discussion Post dataset, and 87% of direct identifier tokens that are not student names. Both SpaCy and the Stanford Deidentifier outperformed PIILO in terms of recall of student names; however, these models were less precise because they do not look specifically for student names, and the reported F_1 was much higher for PIILO as compared to SpaCy and the Stanford Deidentifier. Student-generated writing contains many names which are not personally identifiable, such as cited authors, invented characters, and public figures. This is generally not the case with medical documents. Thus, a major challenge of labeling student names is distinguishing them from other person names in the data that are not PII, as this impacts the utility of the deidentified data.

This is also a limitation of rule-based labelers, which cannot easily distinguish between private and non-private information. It is relatively easy to label URLs with a pattern-based rule, but significantly harder to detect which URLs are PII (Facebook pages, LinkedIn profiles, etc.). We saw that PIILO's rule-based URL labeler resulted in low precision for personal URL labeling because all URLs were treated as private, even though the Discussion Post Dataset included many URLs that could not be used to reidentify a student.

Similar to the results of the first study, PIILO performed differently for full names and first names by themselves. For full names, PIILO labeled 129 of 141 (91% accuracy), leaving 12 leaked full names in the deidentified data (9% inaccuracy). For first names, PIILO labeled 149 of 276 (54% accuracy), leaving 127 leaked first names in the deidentified data (46% inaccuracy). Review of the false negatives (leaked names) revealed that leaked names were likely to be used in the second or third person (e.g., "Hi Tiziana," "Stefan would like this."). PIILO's student name labeling model may have learned to look for person names in the first person ("I am…", "My name is…") and names outside of sentences ("Learning Reflection – Samuel Johnson") while ignoring names used in the third-person, which are much more likely to be cited authors or lecturers in the MOOC dataset. The second-person perspective is also unlikely to appear in the MOOC dataset on which the model was trained, because it is made of essays. In the Discussion Post dataset, this learned bias was harmful to model performance, as students frequently referred to each other by name. The results from Study Two indicate that a general-purpose text deidentification model specific to educational genres will require training on more diverse student writing data.

Study Three – Evaluating the Hiding in Plain Sight Obfuscation Strategy

Studies One and Two demonstrated that student names are difficult to label because they do not follow regular patterns or appear in consistent contexts and are difficult to distinguish from other person names (such as authors and public figures). Knowing that machine learning models cannot reach 100% accuracy in PII labeling, Study Three assessed the practicality of implementing a hiding in plain sight (HIPS) obfuscation strategy to protect student privacy, even when a student name is not correctly labeled by PIILO.

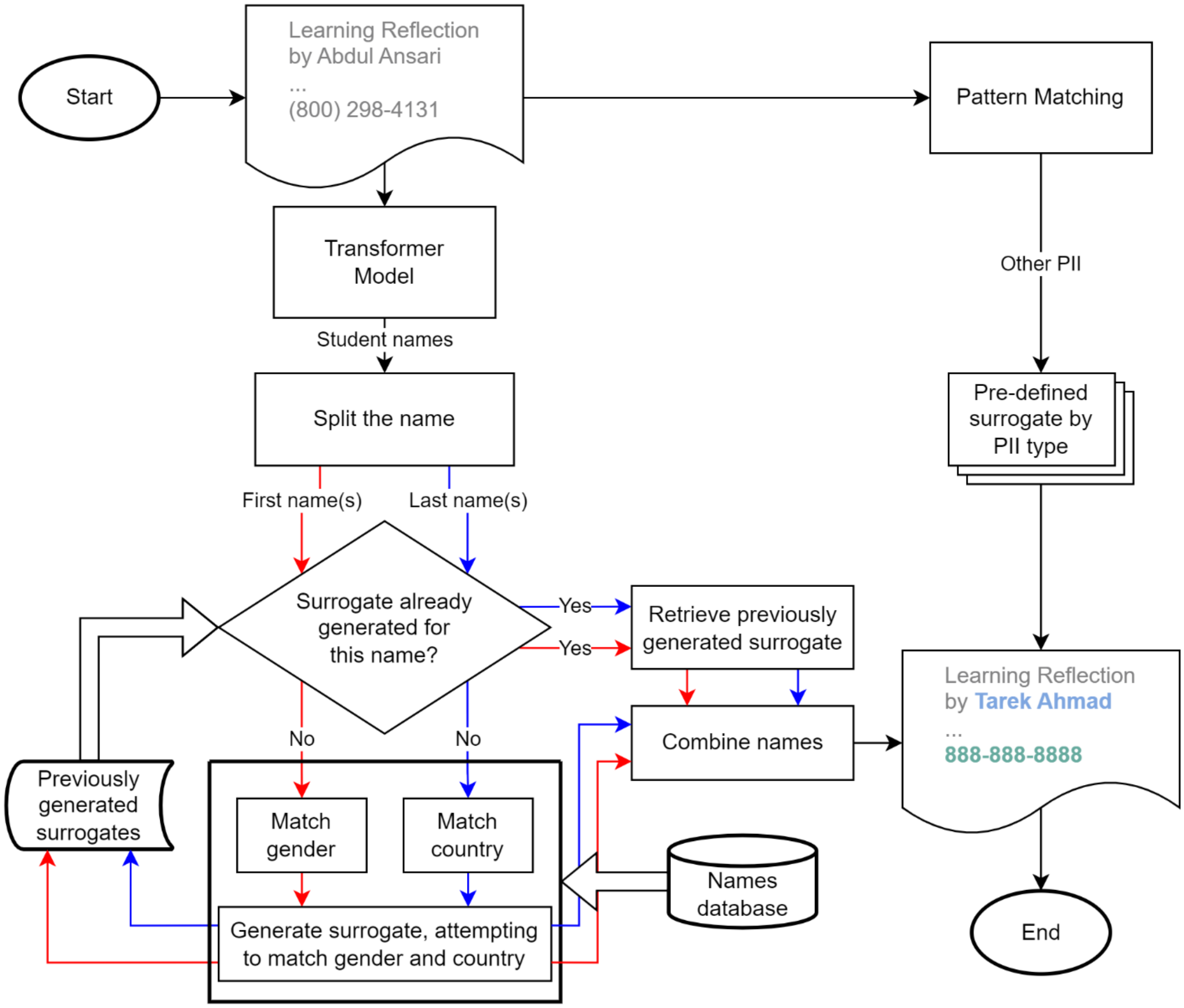
Method

Once the PII in each document was automatically labeled, the identifiers were obfuscated depending on the PII type. PIILO replaced student names with a suitable surrogate that matches the gender and country of the original name (as shown in Figure 1). PIILO attempts to match the gender and country of the original name primarily because texts may contain information about a student's gender or nationality, and a mismatch between this information and a surrogate name could reveal the surrogate name as fake. For instance, consider the sentences, "My name is Jessica (or Jess). I am the youngest of 3 girls." These sentences contain two instances of a student name. If only the first instance was correctly labeled, the sentence could be obfuscated as either "My name is Mike" or "My name is Jess" followed by "I am the youngest of 3 girls." In the case of a gender mismatch (i.e., "Mike" matched with "girl"), it would be fairly clear that "Mike" was obfuscated.

In order to match the gender and country of labeled names, we rely on the Python package names-dataset (Remy, 2021), which includes a large database of names coded for gender and country. We note that this is a crude approach to emulating any gender or ethnolinguistic information that a name can convey, but we will show that it is reasonably effective for the purpose of generating plausible surrogate names. PIILO first splits names into first names and last names using a rule-based name parser (Gulbranson, 2023). The name parser splits names on whitespace and attempts to parse them into titles, first names, nicknames, middle names, last names, and suffixes. For example, the name “Sra. Maria (mary) Teresa de Arroyo II” would be parsed into the title “Sra.”, the first name “Maria”, the nickname “mary”, the middle name “Teresa”, a last name with two tokens “de Arroyo”, and the suffix “II”. The parser works well considering its simplicity, but it can improperly parse names in some situations. For example, it would parse “Dean” as a first name even though it may be used as a title, and it would parse “John” as a last name if it is preceded by a title and not followed by another name, as in “Mr. John”, even though “John” is likely to be a first name. After parsing a name, PIILO attempts to match gender based on the first names and country based on the last names. If a match for gender and/or country is found, a surrogate name is sampled at random from the names database using a weighted sampling strategy. If neither country nor gender can be matched based on the original name, PIILO will choose a name at random from the full dataset (also weighted by frequency of occurrence). In either case, student names that are labeled by PIILO are replaced by a randomly sampled surrogate name. Original and surrogate names are stored in memory while PIILO runs, so student names that are repeated in a document or across the dataset will be replaced with the same surrogate. All direct identifiers except names are replaced with a constant surrogate identifier, such as *janedoe@aol.com* for email addresses and *(888)-888-8888* for phone numbers. PIILO uses a straightforward substitution strategy for these identifiers because they can be labeled with high accuracy.

Figure 1

A flowchart describing the full PIILO system.



To evaluate the efficacy of the HIPS obfuscation strategy, we selected 460 documents from the discussion post dataset that were known to contain PII. We selected only documents that contain PII because the review process was time-consuming. These documents were first processed with PIILO to automatically label and obfuscate any PII. In order to assess privacy disclosure risk, we designed a procedure (inspired by Carrell et al., 2013) in which two reviewers served as attackers, attempting to reidentify students from the obfuscated texts. This configuration was inspired by the methodology of Carrel et al. (2013), which used a similar reidentification attack to evaluate the HIPS strategy on medical texts. Both attackers examined all 460 documents, labeling any text they believed to be a direct identifier of a student.

A third reviewer served as an evaluator. The evaluator reviewed the original document alongside both attackers' annotations. The evaluator then marked each reidentification attempt as successful or unsuccessful. These results are reported in terms of true positives (how many students were successfully reidentified) and false negatives (how many students' identities were protected). Each student identity is counted only once per document, as distinct from previous analyses which were calculated on a per-token basis. Successfully obfuscating a student's name in one location has little benefit if the student can be reidentified using other available information. We also include false positives (how many reidentification attempts were unsuccessful) because HIPS may help to reduce the attackers' confidence about a student's identity, even if they successfully recover it.

Results

Attacker One made 46 reidentification attempts. Of these, 30 were leaked identifiers and 16 were surrogate names generated by PIILO. Full names were involved in 19 out of the 46 reidentification attempts, and 12 of these were leaked identifiers. Attacker One used textual clues to reidentify students, noting that names repeated across multiple documents were suspicious. This strategy was moderately successful because PIILO leaked some names consistently. For example, one student who was highly active in the forum discussions was mentioned 36 times and leaked (first name only) 20 times. PIILO generated the same surrogate name ("Tiziana") for this student across all 16 obfuscated instances. Attacker Two made 31 reidentification attempts. Of these, 18 were leaked identifiers, and 13 were surrogate names generated by PIILO. Full names were involved in 22 out of the 31 reidentification attempts, and 12 of these were leaked identifiers. Attacker Two noted that unusual or non-American names were suspicious (likely to be leaked). This strategy worked against them because PIILO uses a diverse database of names to generate surrogates. The reidentification performance of the attackers is summarized in Table 5. Both attackers performed below chance in reidentification.

Table 5

Human reidentification performance on the Discussion Post dataset.

	True Positive	False Negative	False Positive	Precision	Recall	F_1
Attacker One	30	108	16	.65	.22	.33
Attacker Two	17	121	13	.57	.12	.20

In terms of agreement, both attackers made the same reidentification attempt only 15 times, and of these, they were correct 14 times. Both attackers made the same incorrect reidentification attempt on one occasion. Twelve of the shared, correct reidentification attempts were made in reference to a single post that included 12 unique leaked first and last names. These 12 were the only full names recovered by either attacker. While both attackers noted that they had low confidence about this reidentification attempt, this constitutes the most concerning leak and an important qualification on the system's performance.

Discussion

The results of Study Three indicate that the HIPS strategy makes it more difficult to distinguish leaked identifiers, thereby partially compensating for labeling systems with less than 100% recall. This is consistent with other studies that have shown HIPS to reduce the privacy risk of leaked identifiers in the medical domain (Carrell et al., 2013, 2019). This investigation also revealed some critical areas for future development. The biggest issue is that 12 names were leaked in a single post, and both attackers identified these names (albeit with low confidence in both cases).

While Study 2 highlighted some important areas for further improvement of PIILO’s labeling systems, the HIPS obfuscation strategy worked as intended. Artificially generated surrogate names were plausible enough to thwart reidentification attempts in a majority of cases. Attackers had reduced confidence about their reidentification attempts even when they were correct, and their accuracy was below chance. Further developments to PIILO’s labeling systems would increase the risk reduction profile of the HIPS obfuscation strategy.

Conclusion

In this paper we have evaluated the performance of PIILO, an open-source system for automatically de-identifying unstructured text in the educational domain. The evaluation was carried out in three studies. Study One demonstrated that PIILO recalled 84% of student name tokens on in-domain testing data, but many of the false negatives were partial names that represent a lower risk to student privacy. Considering only full names, the accuracy was 96%. Study Two evaluated PIILO’s ability to detect student names and other identifiers on out-of-domain student-generated text. In a dataset of classroom discussion forum posts, PIILO recalled 70% of student names and 87% of other direct identifiers. Again, many of the leaked names were first names. Considering only full names, the accuracy was 91%. These results are promising, but also highlight the need for more diverse training data and continuous evaluation of deidentification performance. In Study Three, we evaluated the capacity for HIPS to protect student privacy. We found that it was difficult to discern between leaked identifiers and artificially generated surrogate names, which reduced the identity disclosure risk associated with PII that may be leaked.

PIILO includes a transformer-based model that allows it to detect student names in educational text. This model performed well on in-domain data. However, on out-of-domain data, the model did not perform as well. There were many leaked names, particularly first names that appeared in the second and third person, which underscores the need for more diverse training data and continuous evaluation of model performance when applied to new educational datasets. We hope to address this limitation by developing new transformer models trained on more diverse data.

PIILO uses rule-based systems to label other types of direct identifiers. These systems achieved high recall, but there were some limitations that indicate that additional rule-based systems will need to be developed for street addresses and usernames. Overall, results provide evidence that pattern matching is an effective approach for labeling some types of PII, but substantial work remains to develop these systems and evaluate their performance on diverse types of student-generated text.

PIILO uses HIPS to obfuscate labeled identifiers. In Study Three, we evaluated the potential for HIPS to reduce the identity disclosure risk associated with leaked identifiers. This is a crucial component of PIILO because even as we improve our labeling systems, we do not expect to be able to guarantee 100% recall of identifiers. As a result, we assessed the risk of a student being reidentified using a leaked identifier. We found that HIPS dramatically reduces but does not eliminate the risk of identity disclosure. We concluded that HIPS should be implemented in deidentification systems broadly.

We argue that PIILO is a powerful but currently imperfect tool for text deidentification. A difficult question data stewards face when deciding to use any deidentification strategy is whether that strategy performs well enough to protect student privacy. This question is challenging because no agreed upon performance thresholds exist that are quantifiable and achievable. One temptation is to treat any solution with less than 100% recall of identifiers as insufficient, but there are several issues with this perspective. First, even trained human annotators will not reach 100% agreement on a complex labeling task such as PII annotation. While some PII is straightforward to annotate, many other instances of PII will be less clear. Second, this perspective ignores or assigns little value to the potential social good that can come from collecting, sharing, and analyzing student data. There is a difficult tradeoff between the risk of disclosing a student's identity and the benefit of sharing data, which can ultimately be used to improve student learning outcomes. Our perspective is that a good first step for automated text deidentification systems is to reach the performance of human annotators. However, labeled PII data is needed from at least two raters for each instance to calculate inter-rater reliability, which is a future direction for follow-up studies. This criterion will not directly help data stewards make the difficult decision of whether they can share data, but it may help guide the development and evaluation of automated text deidentification systems. In practical scenarios, we would argue that student-generated text can never be treated as fully anonymized; rather anonymization efforts serve to reduce the risk to the student. Researchers and policy makers must weigh this risk against the benefits of data collection, analysis, and data sharing.

Limitations of the Present Study

The primary limitation of the studies presented here is the diversity and quantity of the data used. The MOOC dataset contains 5,797 essays that were all submitted in response to the same, open-ended prompt. The Discussion Post dataset contains 2,118 labeled posts that were all written in the same discussion forum at one university. It is likely that more labeled data from a greater variety of contexts would allow for training a superior transformer-based model. More diverse data would also improve the system’s ability to generalize to additional contexts. Despite these limitations, we see significant potential for deidentification systems like PIILO to improve research in the learning sciences by reducing the labor required to deidentify and share student writing data. Any reduction in the labor requirements of deidentification and the privacy risk of sharing data will make it easier for researchers to share data, enabling more open and reproducible research. Furthermore, as more data becomes available, researchers may carry out studies without collecting data themselves. This economization of research will benefit the field and learners by lowering barriers to scientific inquiry.

PIILO, in its current form, is not sufficient to fully protect student privacy in all situations. Rather, our perspective is that recent advancements in natural language processing have made deidentification of student writing a much more tractable problem. This perspective is informed by deidentification systems that have achieved greater than 95% recall of identifiers in medical texts (Chambon et al., 2023; Murugadoss et al., 2021) and the success of PIILO reaching 91% recall of full names in discussion post data and 96% recall of full names in student essays. Furthermore, we believe there is potential to compensate for less-than-perfect recall by using an obfuscation strategy like hiding-in-plain-sight, which would further reduce the risk of breaching student privacy. Finally, as norms surrounding data privacy and data sharing are rapidly changing, informed consent will continue to play an important role in the creation of shared datasets. We see PIILO as a springboard for future work that explores how deidentification technology can be combined with consent practices to make collecting and sharing student writing data more practical within an ethical framework.

References

Artifex. (2022). *PyMuPDF* [Computer software]. https://pymupdf.readthedocs.io/en/latest/intro.html#license-and-copyright

Beltagy, I., Peters, M. E., & Cohan, A. (2020). *Longformer: The Long-Document Transformer* (arXiv:2004.05150). arXiv. https://doi.org/10.48550/arXiv.2004.05150

Bosch, N., Crues, R. W., & Shaik, N. (2020). "Hello, [REDACTED]": Protecting student privacy in analyses of online discussion forums. *Proceedings of The 13th International Conference on Educational Data Mining*, 11.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. *ArXiv:2005.14165 [Cs]*. http://arxiv.org/abs/2005.14165

Carrell, D., Cronkite, D. J., Li, M. (Rachel), Nyemba, S., Malin, B. A., Aberdeen, J. S., & Hirschman, L. (2019). The machine giveth and the machine taketh away: A parrot attack on clinical text deidentified with hiding in plain sight. *Journal of the American Medical Informatics Association*, *26*(12), 1536–1544. https://doi.org/10.1093/jamia/ocz114

Carrell, D., Malin, B., Aberdeen, J., Bayer, S., Clark, C., Wellner, B., & Hirschman, L. (2013). Hiding in plain sight: Use of realistic surrogates to reduce exposure of protected health information in clinical text. *Journal of the American Medical Informatics Association, 20*(2), 342–348. https://doi.org/10.1136/amiajnl-2012-001034

Chambon, P. J., Wu, C., Steinkamp, J. M., Adleberg, J., Cook, T. S., & Langlotz, C. P. (2023). Automated deidentification of radiology reports combining transformer and "hide in plain sight” rule-based methods. *Journal of the American Medical Informatics Association*, *30*(2), 318–328. https://doi.org/10.1093/jamia/ocac219

Cormack, A. N. (2016). A data protection framework for learning analytics. *Journal of Learning Analytics*, *3*(1), Article 1. https://doi.org/10.18608/jla.2016.31.6

Crossley, S. A., Baffour, P., Tian, Y., Picou, A., Benner, M., & Boser, U. (2022). The persuasive essays for rating, selecting, and understanding argumentative and discourse elements (PERSUADE) corpus 1.0. *Assessing Writing*, *54*, 100667. https://doi.org/10.1016/j.asw.2022.100667

Crossley, S. A., Tian, Y., Baffour, P., Franklin, A., Kim, Y., Morris, W., Benner, M., Picou, A., & Boser, U. (in press). Measuring second language proficiency using the English Language Learner Insight, Proficiency and Skills Evaluation (ELLIPSE) corpus. *International Journal of Learner Corpus Research*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv:1810.04805 [Cs]*. http://arxiv.org/abs/1810.04805

Dorr, D. A., Phillips, W. F., Phansalkar, S., Sims, S. A., & Hurdle, J. F. (2006). Assessing the difficulty and time cost of de-identification in clinical narratives. *Methods of Information in Medicine*, *45*(3), 246–252. https://doi.org/10.1055/s-0038-1634080

Explosion AI. (2022). *English SpaCy models documentation*. SpaCy. https://spacy.io/models/en

Gulbranson, D. (2023). *Name Parser* (1.12) [Python]. https://github.com/derek73/python-nameparser

Hassan, F., Sanchez, D., & Domingo-Ferrer, J. (2021). Utility-preserving privacy protection of textual documents via word embeddings. *IEEE Transactions on Knowledge and Data Engineering*, 1–1. https://doi.org/10.1109/TKDE.2021.3076632

Jones, K. M. L. (2019). Learning analytics and higher education: A proposed model for establishing informed consent mechanisms to promote student privacy and autonomy. *International Journal of Educational Technology in Higher Education*, *16*(1), 24. https://doi.org/10.1186/s41239-019-0155-0

Jones, K. M. L., Asher, A., Goben, A., Perry, M. R., Salo, D., Briney, K. A., & Robertshaw, M. B. (2020). “We’re being tracked at all times”: Student perspectives of their privacy in relation to learning analytics in higher education. *Journal of the Association for Information Science and Technology*, *71*(9), 1044–1059. https://doi.org/10.1002/asi.24358

Li, W., Sun, K., Schaub, F., & Brooks, C. (2022). Disparities in students’ propensity to consent to learning analytics. *International Journal of Artificial Intelligence in Education*, *32*(3), 564–608. https://doi.org/10.1007/s40593-021-00254-2

Lison, P., Pilán, I., Sanchez, D., Batet, M., & Øvrelid, L. (2021). Anonymisation models for text data: State of the art, challenges and future directions. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 4188–4203. https://doi.org/10.18653/v1/2021.acl-long.323

Mansfield, C., Paullada, A., & Howell, K. (2022). *Behind the mask: Demographic bias in name detection for PII masking* (arXiv:2205.04505). arXiv. https://doi.org/10.48550/arXiv.2205.04505

Megyesi, B., Granstedt, L., Johansson, S., Prentice, J., Rosén, D., Schenström, C.-J., Sundberg, G., Wirén, M., & Volodina, E. (2018). Learner corpus anonymization in the age of GDPR: Insights from the creation of a learner corpus of Swedish. *Proceedings of the 7th Workshop on NLP for Computer Assisted Language Learning*, 47–56. https://aclanthology.org/W18-7106

Murugadoss, K., Rajasekharan, A., Malin, B., Agarwal, V., Bade, S., Anderson, J. R., Ross, J. L., Faubion, W. A., Halamka, J. D., Soundararajan, V., & Ardhanari, S. (2021). Building a best-in-class automated de-identification tool for electronic health records through ensemble learning. *Patterns*, *2*(6), Article 6. https://doi.org/10.1016/j.patter.2021.100255

Pardo, A., & Siemens, G. (2014). Ethical and privacy principles for learning analytics. *British Journal of Educational Technology*, *45*(3), 438–450. https://doi.org/10.1111/bjet.12152

Parra Escartín, C., Reijers, W., Lynn, T., Moorkens, J., Way, A., & Liu, C.-H. (2017, April 4). *Ethical considerations in NLP shared tasks*. First Workshop on Ethics in Natural Language Processing, Valencia, Spain. https://doi.org/10.18653/v1/W17-1608

Pilán, I., Lison, P., Øvrelid, L., Papadopoulou, A., Sánchez, D., & Batet, M. (2022). *The text anonymization benchmark (TAB): A dedicated corpus and evaluation framework for text anonymization* (arXiv:2202.00443). arXiv. https://doi.org/10.48550/arXiv.2202.00443

Presidio—Data protection and anonymization API. (2022). [Python]. Microsoft. https://github.com/microsoft/presidio (Original work published 2018)

Remy, P. (2021). *Name dataset* [Computer software]. GitHub. https://github.com/philipperemy/name-dataset

Rubel, A., & Jones, K. M. L. (2016). Student privacy in learning analytics: An information ethics perspective. *The Information Society*, *32*(2), 143–159. https://doi.org/10.1080/01972243.2016.1130502

Sites, D. (2022). *Compact language detector 2* [C++]. https://github.com/CLD2Owners/cld2

Slade, S., & Prinsloo, P. (2013). Learning analytics: Ethical issues and dilemmas. *American Behavioral Scientist*, *57*(10), 1510–1529. https://doi.org/10.1177/0002764213479366

Sun, K., Mhaidli, A. H., Watel, S., Brooks, C. A., & Schaub, F. (2019). It's my data! Tensions among stakeholders of a learning analytics dashboard. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3290605.3300824

Uzuner, Ö., Luo, Y., & Szolovits, P. (2007). Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association*, *14*(5), 550–563. https://doi.org/10.1197/jamia.M2444

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *ArXiv:1706.03762 [Cs]*. http://arxiv.org/abs/1706.03762

Wang, K., Stevens, R., Alachram, H., Li, Y., Soldatova, L., King, R., Ananiadou, S., Schoene, A. M., Li, M., Christopoulou, F., Ambite, J. L., Matthew, J., Garg, S., Hermjakob, U., Marcu, D., Sheng, E., Beißbarth, T., Wingender, E., Galstyan, A., … Rzhetsky, A. (2021). NERO: A biomedical named-entity (recognition) ontology with a large, annotated corpus reveals meaningful associations through text embedding. *Systems Biology and Applications*, *7*(1), Article 1. https://doi.org/10.1038/s41540-021-00200-x

Weischedel, R., Palmer, M., Marcus, M., Hovy, E., Pradhan, S., Ramshaw, L., Xue, N., Taylor, A., Kaufman, J., Franchini, M., El-Bachouti, M., Belvin, R., & Houston, A. (2013). *OntoNotes release 5.0* [dataset]. Linguistic Data Consortium. https://doi.org/10.35111/XMHB-2B84

Weitzenboeck, E. M., Lison, P., Cyndecka, M., & Langford, M. (2022). The GDPR and unstructured data: Is anonymization possible? *International Data Privacy Law*, *12*(3), 184–206. https://doi.org/10.1093/idpl/ipac008

Yan, H., Gui, T., Dai, J., Guo, Q., Zhang, Z., & Qiu, X. (2021). *A unified generative framework for various NER subtasks* (arXiv:2106.01223). arXiv. http://arxiv.org/abs/2106.01223

Young, E. M. (2015). Educational privacy in the online classroom: FERPA, MOOCS, and the big data conundrum. *Harvard Journal of Law & Technology*, *28*(2).