The Cleaned Repository of Annotated Personally Identifiable Information

Langdon Holmes
Vanderbilt University
langdon.holmes@vanderbilt.edu

Scott Crossley
Vanderbilt University
scott.crossley@vanderbilt.edu

Jiahe Wang
Vanderbilt University
jiahe.wang@vanderbilt.edu

Weixuan Zhang
Vanderbilt University
weixuan.zhang@vanderbilt.edu

ABSTRACT

Protecting student privacy is of paramount importance and has historically led to educational datasets not being released to the general community. Instead, many datasets are shared among a small number of researchers working on specific projects. However, these datasets could provide significant value to the educational research community if they were made available and could help ensure replication studies of important educational research. Deidentifying the student data is, in some cases, sufficient to permit data sharing among researchers and even public release. However, most educational datasets are quite large, making deidentification extremely time-consuming and difficult. A solution is automated deidentification, but this is challenging for unstructured text data like that found in educational environments. This paper introduces a new open-source dataset called the Cleaned Repository of Annotated Personally Identifiable Information (CRAPII). CRAPII is designed to test and evaluate automated deidentification methods for educational data. The dataset comprises over 20,000 student essays that have been annotated for personally identifiable information (PII). Within the dataset, all occurrences of PII have been replaced with surrogate identifiers of the same type. The purpose of CRAPII is to promote the development of automated deidentification methods specifically designed for and tested on student writing. To further this goal, we are hosting an open data science competition in which teams of data scientists compete to develop deidentification algorithms using CRAPII.

Keywords

privacy, deidentification, anonymization, educational data, personally identifiable information

1. INTRODUCTION

Student writing samples provide valuable insight into learning that can help researchers understand learning outcomes, learning processes, student motivation, educational development, and user feedback. Student writing also presents valuable resources for developing and evaluating automated content generation systems and approaches to personalized learning including recommendation and feedback systems. Student writing data may be collected from essays, free responses to questions, peer feedback comments, discussion forum posts, and chat messages. Written data has become more valuable for educational research as recent advances in natural language processing (NLP) allow researchers to more easily quantify variables of interest in writing samples, including measures of writing quality, content mastery, and socio-emotional learning [13]. Chat logs produced from student interactions with intelligent tutors and generative AI are also becoming a valuable resource for gathering insights into how students learn and develop [18].

Research into student writing is limited by a paucity of large and open-source collections of samples, despite calls for improved data sharing practices across many fields [33]. One reason that researchers are often reluctant to share data is the presence of private or sensitive information [10, 11]. This problem is especially challenging when datasets comprise student writing, which may contain identifying information such as names, e-mails, and phone numbers. Accidentally sharing such information would compromise students' privacy and could have detrimental effects for researchers including loss of funding and personal fines. Furthermore, educational institutions in the United States who inappropriately handle student data can face penalties under the Family Educational Rights and Privacy Act of 1974 (FERPA) [37].

Written data is distinct from structured (tabular) data [25] in which deidentification can sometimes be as simple as removing the column that contains private information (such as an "email" column). In contrast, deidentifying student writing requires finding the email addresses in the writing samples and then obfuscating them. While this is relatively easy for humans to do, annotator labor is resource intensive. On the other hand, automatic deidentification remains challenging, and while some solutions have reached over 95% recall for medical records [7, 23], these solutions are less effective for student writing genres, such as essays and forum discussions [14]. For example, they would likely fail to distinguish between a student's name (private) and a cited author (not private), resulting in poor precision. Previous work on deidentification of student writing has largely focused on student names, reaching 84% recall in student essays [16] and 95% recall in forum discussion posts [4]. However, these solutions either did not account for or had poor performance labeling other important identifier types in student writing, including usernames and personally identifiable URLs (such as a link to the student's social media profile) [14].

Since no comprehensive automatic deidentification tools exist for student writing, educational researchers that want to share data are forced to deidentify student writing manually [21]. In practice, however, the high costs of manual deidentification means that most data remains private. In the era of big data, this is a major limitation

L. Holmes, S. Crossley, J. Wang, and W. Zhang. The cleaned repository of annotated personally identifiable information. In B. Paaßen and C. D. Epp, editors, *Proceedings of the 17th International Conference on Educational Data Mining*, pages 790–796, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.12729952

on the field of educational text mining. Open datasets promote reproducibility and transparency in research and allow researchers to address research questions without collecting new data, saving time and resources. Open datasets also offer unique opportunities to evaluate the performance of predictive models [22].

Our goal is to release a large-scale corpus of student writing that is annotated for PII to help researchers develop and evaluate automated approaches to the deidentification of student writing. By applying multiple different models to the dataset, progress can be monitored and more effective comparisons between methods can be made, allowing for benchmarks to be set. Shared tasks like PII annotation allow diverse research teams to tackle problems using open-access datasets [17, 32]. Successful models developed on a PII corpus would allow applications across the field of educational text mining and help researchers release open educational datasets in the future, supporting higher quality and replicable research.

1.1 Deidentification

Deidentification is an approach to protecting privacy that entails the obfuscation of a predefined set of identifiers from a dataset [20]. Identifiers can be direct or indirect, where direct identifiers are variables that are unique to a specific person and sufficient to independently reidentify that person while indirect identifiers could only identify a person when combined with other information. Direct identifiers include email address, name, and phone number; indirect identifiers include nationality, employer, and age. In the medical field, *The Health Insurance Portability and Accountability Act* defines 18 identifiers that must be obfuscated before medical patient data can be shared in the United States [26]. The U.S. Department of Education diverges from this usage of the term deidentification, describing deidentified data as having "all personal identifiable information removed" including information that "alone or in combination" could lead to the identification of a student with "reasonable certainty" [36]. In the present paper, we will refer to this level of privacy as "full anonymization."

A limitation of deidentification is that it does not fully eliminate identity disclosure risk. This is because a predefined list of identifiers does not cover all possible types of identifying information. While deidentification can substantially mitigate the risk of revealing a data subject's identity, it may not be sufficient for all contexts and types of data. Data may be considered especially sensitive when the data subject belongs to a vulnerable population, such as a minority, or the data contains sensitive information such as the subject's criminal history or religious beliefs. Additionally, in cases where informed consent to share the deidentified data cannot be obtained, other approaches to protecting privacy may be required. For example, in the context of FERPA, data that is merely deidentified (and not fully anonymized) would require the student's and/or parents' consent before it could be shared publicly by an educational institution in the United States. Full anonymization is time consuming and costly for unstructured data such as student writing because it requires human labor and is likely to involve specialized annotation software, staffing of annotators, training, monitoring progress and accuracy, and the careful development of effective annotation guidelines. Restricted data sharing via data enclaves or restricted use agreements [1, 11] is an alternative option when deidentification is inappropriate and full anonymization would be too costly. However, these approaches have limitations as well. Data enclaves and restricted use agreements require ongoing involvement from the data steward to review requests for data use and maintain the data's availability. Furthermore, data enclaves restrict researchers from using their own computers to analyze the data, which can add friction to the process of analysis and even restrict the types of models that can be developed using the data.

An alternative approach is automatic deidentification, which is a sub-task of named entity recognition (NER). Like NER, automated deidentification is primarily concerned with extracting a predefined set of named entities (people, organizations, dates, etc.) from a text. However, deidentification also requires a decision to be made about the identity disclosure risk associated with each entity.

State of the art approaches to automatic deidentification rely on pretrained, transformer-based language models [7, 23]. Adapting these models for the purpose of deidentification requires a substantial amount of labeled training data. A major barrier to progress in the development of models for deidentifying student writing is the lack of publicly available datasets that can be used for model development and evaluation. Rapid progress in the field of natural language processing (NLP) benefits from shared datasets in which different methods can easily and effectively be compared. Several shared tasks have led to progress in deidentification of other text types, including the i2b2 shared task, in which 18 teams competed to deidentify medical discharge summaries [35]. Several datasets with labeled PII have recently been produced for the purpose of addressing privacy issues in language model pre-training data. The BigCode PII dataset contains 12,000 samples of computer code annotated for PII by crowd-workers [2]. The PII masking dataset from ai4privacy contains 200,000 synthetically-generated samples labeled for PII [28]. These datasets are important resources promoting the development of automatic deidentification systems, but they do not reflect the specific privacy protection requirements of student writing genres.

1.2 Current Work

The current work seeks to promote the development of automated methods for the deidentification of student writing by introducing a large, public dataset of student writing samples that is labeled for PII. The purpose of the dataset is to give educational researchers a sandbox from which to develop and evaluate models for the automatic annotation of PII. The dataset was used in a shared data science task hosted on Kaggle [34]. Development of the dataset is discussed in three sections: data selection, annotation of PII, and obfuscation of PII.

2. THE CRAPII CORPUS

The Cleaned Repository of Annotated Personally Identifiable Information (CRAPII) is a corpus of 22,688 samples of student writing. Of these documents, 31.2% contain at least one instance of personally identifiable information (PII) across 14 distinct PII types. All instances of PII have been replaced with contextually informed and plausible surrogate identifiers to protect the identity of the original authors while also maintaining the utility of the dataset.

2.1 Data Selection

CRAPII was built from student writing samples collected from learners enrolled in a massively open online course. The course was offered by a large university in the United States and focused on 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. At the time of data collection (April 2022), 367,788 students had enrolled, and 39,118 students had completed the course.

The course encompassed lecture videos, a discussion forum, and assessments. To fulfill the requirements of the course, students had

to submit a reflection essay applying the course concepts to a familiar problem. Submissions were required to be in PDF format. At the time of download, there were a total of 221,043 recorded submission events, including multiple entries from some students. Participant consent was obtained by the educational provider who collected the data, and no data was retrieved from participants under the age of 18. The study was evaluated by an Institutional Review Board (IRB) who determined the study did not require review.

To compile CRAPII, we conducted a series of data cleaning procedures summarized in Table 2. Our initial step was to select all submissions that had been graded. Submissions were graded via a peer review process. We selected graded submissions so that the scores could be explored in a separate investigation. In instances where a single student had multiple graded submissions, determining which submission corresponded to the overall course grade proved challenging. Consequently, users with multiple graded submissions were excluded from the study, leaving 38,267 viable submissions.

Table 1 Effect of data selection steps on corpus size

Selection Step	Count Remaining
Submission events	221,043
Graded submissions	44,593
Submissions unique to user	38,267
Valid download URL	32,525
Parsed to English text	29,142
Deduplicated	23,390
Non-anomalous	22,688

Each submission was linked to a download hyperlink. To refine the dataset, we excluded files lacking a valid hyperlink, exceeding 10 megabytes in size, or surpassing 5 pages in length. We eliminated submissions more than 5 pages in length because these were largely irrelevant documents, such as publicly available dissertations, and slide decks, which contained minimal text in paragraph form. Following these criteria, 32,525 assignment submissions remained and were subsequently downloaded.

After downloading, submissions were automatically parsed using the PyMuPDF parsing library [30], which converted them to plain text. If the file was parsed without error, we then ensured that it was written in English using the Chromium language detection algorithm [8]. In addition to filtering out non-English submissions, this step also served to remove submissions that were parsed incorrectly (i.e., the English text was not recovered by the parser). Lastly, we excluded any submissions with less than 50 words (whitespace-delimited tokens) because these would likely not contain enough language for subsequent analysis [9]. After removing documents based on these criteria, the resulting corpus contained 29,152 plain text essays.

To remove near-duplicate documents, we created a document-term matrix (DTM) using the submissions where "terms" are trigrams (contiguous sequences of three tokens). We tokenized the submissions using a simple whitespace tokenizer written as a regular expression. To reduce the number of distinct tokens, we lowercased all characters and removed accents by converting to ASCII encoding. We also removed stopwords using NLTK's English stopwords list [3]. The result was a one-hot encoded vector of trigrams for each document. We then calculated cosine similarity for all pairwise comparisons of documents. We selected cosine similarity over Jaccard similarity [19] or containment score [5] because it can be calculated more efficiently, which was essential given the large number of pairwise comparisons being made. To determine a reasonable threshold value for cosine similarity, we plotted the highest cosine similarity of each document using a cumulative line plot, as shown in Figure 1. While there was a clear inflection point around 0.10, we opted for a more conservative value of 0.9. This means that any document with a cosine similarity over 0.9 to any other document was labeled as a duplicate, and we eliminated both from the dataset.

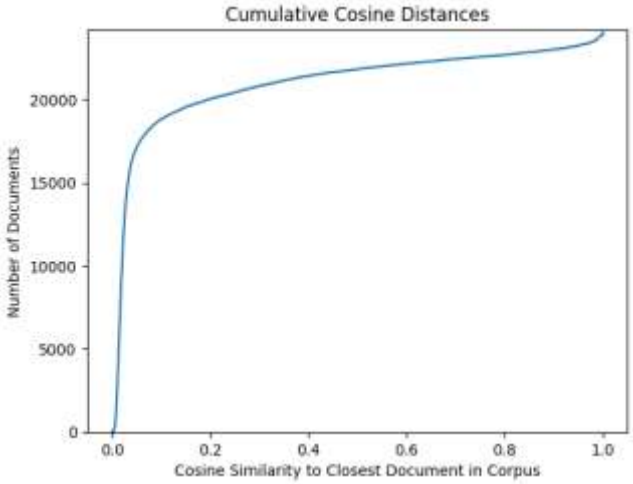

Figure 1 Cumulative density plot of cosine similarity

At this stage, we observed that some documents contained errors that likely resulted from the automatic PDF parsing process. For example, multiple documents had the "io" digraph (in words like "creation") replaced with a "3". However, these errors were too diverse and numerous to build special rules to detect all of them reliably. As a result, we utilized an anomaly detection algorithm known as an elliptic envelope to discover documents with anomalous distributions of characters. For this approach, we utilized a DTM where "terms" are individual characters. We again removed accented characters by converting to ascii because we thought the approach would be more effective for characters with consistent, normal densities across document (for example, the density of the letter 'e' or a space ' ' across documents is roughly normally distributed). We still applied ASCII encoding to remove accents but preserved capitalization and stopwords. This resulted in a matrix with 96 feature columns, each corresponding to an ASCII character. We then fit an elliptic envelope on the DTM. The elliptic envelope is an unsupervised machine learning method that we used to detects texts with atypical character distributions. The approach represents data as a high-dimensional Gaussian distribution and allows for covariance among features. The algorithm aims to draw an ellipse that covers the majority of data occurrences. Data that lies outside the ellipse is considered anomalous. We utilized the sckit-learn implementation [27] and set the contamination parameter to 0.3, which encourages the model to find an envelope that excludes 30% of the data. This value was arrived at by manually reviewing small samples of predicted outliers under different configurations. The final model found 702 outliers. Manual inspection of a sample of these documents revealed that a majority exhibited formatting problems consistent with PDF parsing errors.

The resulting corpus of student writing is substantially more usable, with nearly all essays representing a legitimate effort to complete the assignment. However, the techniques employed to select writing samples are imperfect, and it is likely that many problematic

samples remain in the dataset. These include character encoding or linespacing errors from PDF parsing, as well as partially duplicated documents (likely plagiarism). We hope that any messiness remaining in the dataset will lead to the development of deidentification techniques that generalize well to diverse forms of student writing.

2.2 Annotation of PII

We developed an annotation scheme that included seven direct identifier types and six indirect identifier types. This annotation scheme was based on previous work in text anonymization [29], including our own previous efforts [14, 15]. The direct identifiers were student names, instructor names, email addresses, usernames, IDs, phone numbers, personal URLs, and street addresses. In all cases except for instructor names, only labels that could be used to identify a student were considered PII. Indirect identifiers were ages, dates, locations (including nationality), educational affiliations, employment affiliations, and "other," which served as a catch-all category for any PII that did not fit into another category.

All essays were annotated independently by two raters and any disagreements were adjudicated by a third rater. The first 500 annotated essays from each annotator were reviewed by the first author, and a one-on-one meeting was held to correct any misinterpretations of the annotation guidelines and to discuss edge cases. Annotation and adjudication took place on the UBIAI annotation platform. Inter-annotator agreement (before adjudication) was 83%, as reported by UBIAI. Throughout the process, annotators were reminded to apply labels liberally. After all essays had been annotated by two raters, a single adjudicator reviewed all disagreements across the dataset. Disagreements occurred when one annotator applied a label to a sequence of tokens that differed from the other annotator's label. This includes cases where one annotator labeled a sequence of tokens and the other did not, when one annotator used a different label for the same sequence of tokens, or when both annotators applied the same label to an overlapping sequence of tokens (perhaps one annotator included a piece of punctuation or an article while the other did not). The adjudicator resolved any disagreements by selecting one of the two annotations or by making an entirely new decision. After adjudication, the essays were exported from the web-based annotation software. The approximate counts of each PII type are shown in Figure 2.

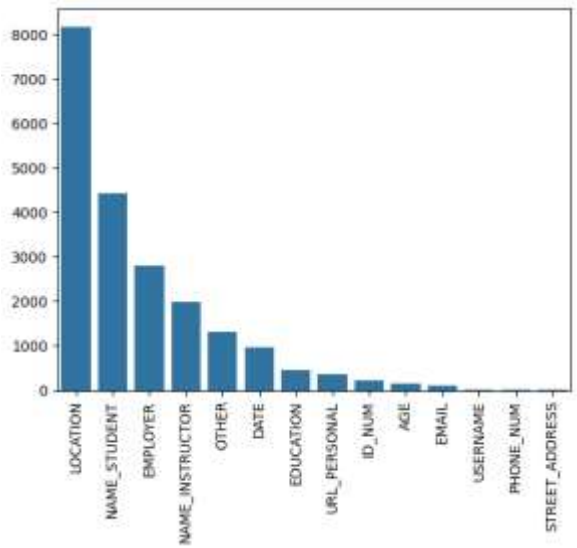

Figure 2 Count of PII Types

2.3 Obfuscation of PII

All instances of PII within the annotated essays were obfuscated using a strategy known as hiding-in-plain-sight or HIPS [6]. We used HIPS because no text deidentification method can guarantee 100% accuracy at scale, not even human annotation. HIPS is a strategy for reducing the identity disclosure risk of leaked identifiers. Under normal circumstances, text deidentification makes it clear what was removed – and what was not – by leaving markers such as "[REDACTED]" or "<name_a>" where the PII has been replaced. As a consequence, any PII that has been leaked is easily distinguished. The HIPS strategy obfuscates labeled PII by replacing that PII with a contextually appropriate surrogate identifier. For instance, a string of characters identified as a student's name (e.g., Juan) would be replaced with the string of characters "Samuel". For the construction of the CRAPII dataset, we adopt this same approach.

2.3.1 Manual Obfuscation

Identifiers labeled as "other" were manually obfuscated by the first author. Using a custom-built graphical interface, each of these identifiers was reviewed in context, and a suitable surrogate was manually generated.

2.3.2 Shuffling of Indirect Identifiers

Indirect identifiers cannot uniquely identify a student on their own, so they are no longer private when taken out of context. As a result, we adopted a simple shuffling strategy for some indirect identifier types. This works by replacing each instance of an identifier with a randomly sampled identifier of the same type from another document in the dataset. This means that the surrogate identifier will often not be contextually appropriate. This approach was deemed acceptable since the focus of the dataset and the competition is on direct identifiers.

The shuffling strategy was adopted for locations (including nationalities), employers, and educators. The sampling was completely random, except that all repeated mentions of the same identifier within an essay were replaced with the same surrogate.

2.3.3 Randomization of Direct Identifiers

For some direct identifiers, it was possible to develop a randomization system that mutated the original value while keeping its orthographic shape (e.g., replacing numbers with different numbers). This strategy allowed us to preserve the form of some identifiers while nullifying their identity disclosure risk.

For student IDs, we replaced each alphabetic character with a random character of the same type (maintaining case), numbers with random numbers, and retained punctuation (such as spaces and hyphens).

We also randomized dates. Dates can appear in a wide variety of formats even without factoring in typographic errors. There are also restrictions on what constitutes a valid date and what dates are contextually appropriate. We first broke dates into day, month, and year components using regular expressions and then randomly sampled an appropriate value for each component. Any sequence of one or two digits with a value of 12 or less was replaced with a random value of 12 or less. Values greater than 12 were replaced with a value from 1-30. This allows us to randomize day and month components without distinguishing between them, and only results in very few impossible dates (e.g., February 30). For years, we sampled from a distribution centered near the original year and skewed towards the past to reduce the risk of generating a contextually inappropriate date (replacing a past date with a future date). Historical

dates and citations are generally not PII, so these dates remain unperturbed.

While it would not be feasible to obfuscate all URLs through randomization, a majority of personal URLs annotated in the dataset belonged to a small set of social media sites. For these URLs, it was possible to deconstruct the string and randomize the identifying component of the URL following the format of major social media site URLs. This approach was preferable to replacing the full URL, because it maintains the original shape and meaning of the URL.

2.3.4 Procedural Generation

We obfuscated usernames, email addresses, phone numbers, and any remaining URLs (URLs not referring to a major social media website) using procedural generation provided by the Faker Python package [12]. Procedural generation with Faker uses a combination of pre-defined patterns for composite entities and lists for fine-grained entity types. For example, URLs can include a protocol (https://), a domain name ("[www.]william-jackson"), a top-level domain (".com"), and a page location ("/blog/posts/1"). Faker randomly selects a pattern string comprising some or all of these elements ("[protocol][domain][top-level domain]"). It then randomly selects fine-grained entities from corresponding lists (or random letters and digits, where appropriate) to populate each component of the pattern.

Most personal URLs not belonging to major social media sites were either a student's personal webpage or blog, and Faker generated plausible surrogate URLs for these cases. Usernames, email addresses, and phone numbers were also generated by Faker using the same strategy of randomly selecting a pattern and populating it with random elements.

2.3.5 Contextual Procedural Generation

Student names are the most frequent type of direct identifier in the dataset by a wide margin. As a result, we adopted a more complex strategy for student names.

Using a large, international dataset of names with associated gender and nationality codes [31], we sampled contextually appropriate names to use as replacements for the originals. The first step in this process was to break the original name into first and last components. We achieved this by using a rule-based name parser [24]. Then, we matched gender using the first name component and nationality using the last name component. Using these values, we randomly sample a full name from the dataset, matching on gender and nationality. While this approach is somewhat crude, it was effective in practice, generating diverse and contextually plausible surrogate names. If a gender or nationality could not be determined, then no filter was applied for this value during sampling. If the combination of gender and nationality resulted in a filtered list of names that was less than 50, a name was sampled from the full dataset at random.

Each of these surrogate generation strategies was implemented with the help of the Presidio anonymization library, which simplified the process of aligning the PII labels to the newly generated PII [38]. The result is a first-of-its-kind, large-scale, publicly released dataset of PII in student writing. The purpose of the corpus is to develop and evaluate deidentification systems, but the corpus itself has been fully anonymized by obfuscating all identifiers present in the writing samples. The corpus is openly available through Kaggle [34]. Code artifacts are available in a GitHub repository (https://github.com/langdonholmes/CRAPII).

3. DISCUSSION

In this paper, we introduced the Cleaned Repository of Annotated Personally Identifiable Information (CRAPII). The corpus provides researchers within the educational data mining community with a resource from which to develop and test deidentification strategies for student writing.

The dataset was developed using the records of an online, open-access course. As a result, significant cleaning steps were required to convert a subset of the data into a usable, plain text form. We attempted to exclude documents written in other languages, that were plagiarized, or that suffered from character encoding issues. To address the latter problem, we implemented a novel approach for identifying documents with anomalous character distributions in a large dataset.

All PII in the dataset was labeled by human annotators, and any disagreements were resolved by an adjudicator. Before adjudication, inter-annotator agreement was measured at 83%, which we consider to be a good value for a complex sequence labeling task. To obfuscate the labeled direct identifiers in the corpus, we adopted several strategies for generating surrogate identifiers that were inspired by the hiding-in-plain-sight approach to deidentification. Replacing the original PII with surrogate identifiers allows the dataset to be released publicly. We hope that the CRAPII corpus will serve as a useful resource for developing and evaluating deidentification methods.

4. FUTURE DIRECTIONS

One of the major challenges of deidentification is that a successful tool must not only categorize named entities; it must also assess the identity disclosure risk of the named entity. For this reason, we expect successful models of PII detection to rely on pre-trained, transformer-based large language models, which tend to perform well for semantically informed tasks.

The purpose of the shared data science task making use of the CRAPII corpus [34] is to develop such models. These models can then be utilized in learning analytics pipelines and by researchers wishing to share their data in a deidentified form. A secondary goal of the shared task is to call attention to the twin challenges of protecting student privacy and practicing science openly. As we have argued, deidentification is not a good fit for all scenarios, but when paired with ethical and thoughtful consideration on the part of data stewards, we believe that cost-effective methods for the deidentification of student writing will benefit educational research by making data sharing more practical.

5. ACKNOWLEDGMENTS

The authors would like to thank The Learning Agency Lab for their support.

This material is based upon work supported by the National Science Foundation under Grant 2112532 Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

6. REFERENCES

[1] Alter, G.C. and Vardigan, M. 2015. Addressing Global Data Sharing Challenges. *Journal of Empirical Research on Human Research Ethics*. 10, 3 (Jul. 2015), 317–323. DOI:https://doi.org/10.1177/1556264615591561.

[2] BigCode PII dataset: 2023. *https://huggingface.co/datasets/bigcode/bigcode-pii-dataset*. Accessed: 2024-01-29.

[3] Bird, S., Klein, E. and Loper, E. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc.

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

[5] Broder, A.Z. 1998. On the resemblance and containment of documents. *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)* (Salerno, Italy, 1998), 21–29.

[6] Carrell, D., Malin, B., Aberdeen, J., Bayer, S., Clark, C., Wellner, B. and 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 (Mar. 2013), 342–348. DOI:https://doi.org/10.1136/amiajnl-2012-001034.

[7] Chambon, P.J., Wu, C., Steinkamp, J.M., Adleberg, J., Cook, T.S. and 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 (Jan. 2023), 318–328. DOI:https://doi.org/10.1093/jamia/ocac219.

[8] Compact language detector 2: 2022. *https://github.com/CLD2Owners/cld2*. Accessed: 2022-10-04.

[9] Crossley, S.A. 2018. How Many Words Needed? Using Natural Language Processing Tools in Educational Data Mining. *Proceedings of the 10th International Conference on Educational Data Mining (EDM)* (2018), 630–633.

[10] Darby, R., Lambert, S., Matthews, B., Wilson, M., Gitmans, K., Dallmeier-Tiessen, S., Mele, S. and Suhonen, J. 2012. Enabling scientific data sharing and re-use. *2012 IEEE 8th International Conference on E-Science* (Oct. 2012), 1–8.

[11] Donaldson, D.R. and Koepke, J.W. 2022. A focus groups study on data sharing and research data management. *Scientific Data*. 9, 1 (Jun. 2022), 345. DOI:https://doi.org/10.1038/s41597-022-01428-w.

[12] Faker: *https://github.com/joke2k/faker*. Accessed: 2023-11-01.

[13] Ferreira-Mello, R., André, M., Pinheiro, A., Costa, E. and Romero, C. 2019. Text mining in education. *WIREs Data Mining and Knowledge Discovery*. 9, 6 (2019), e1332. DOI:https://doi.org/10.1002/widm.1332.

[14] Holmes, L., Crossley, S., Sikka, H. and Morris, W. 2023. PIILO: an open-source system for personally identifiable information labeling and obfuscation. *Information and Learning Sciences*. 124, 9/10 (Jan. 2023), 266–284. DOI:https://doi.org/10.1108/ILS-04-2023-0032.

[15] Holmes, L., Crossley, S.A., Haynes, R., Kuehl, D., Trumbore, A. and Gutu, G. in press. Deidentification of student writing in technologically mediated educational settings. *Proceedings of the 7th conference on Smart Learning Ecosystems and Regional Development (SLERD)* (Bucharest, Romania, in press).

[16] Holmes, L., Crossley, S.A., Morris, W., Sikka, H. and Trumbore, A. 2023. Deidentifying Student Writing with Rules and Transformers. *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky* (Cham, 2023), 708–713.

[17] Howard, A., bskim90, Lee, C., Shin, D., Jeon, H.P., Baek, J., Chang, K., kiyoonkay, NHeffernan, seonwooko and Dane, S. 2020. Riiid Answer Correctness Prediction. Kaggle.

[18] Kochmar, E., Vu, D.D., Belfer, R., Gupta, V., Serban, I.V. and Pineau, J. 2020. Automated Personalized Feedback Improves Learning Gains in An Intelligent Tutoring System. *Artificial Intelligence in Education* (Cham, 2020), 140–146.

[19] Leskovec, J., Rajaraman, A. and Ullman, J.D. 2020. Finding Similar Items. *Mining of massive datasets*. Cambridge University Press. 72–130.

[20] Lison, P., Pilán, I., Sanchez, D., Batet, M. and 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)* (Online, Aug. 2021), 4188–4203.

[21] Megyesi, B., Granstedt, L., Johansson, S., Prentice, J., Rosén, D., Schenström, C.-J., Sundberg, G., Wirén, M. and 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* (Stockholm, Sweden, Nov. 2018), 47–56.

[22] Mihaescu, M.C. and Popescu, P.S. 2021. Review on publicly available datasets for educational data mining. *WIREs Data Mining and Knowledge Discovery*. 11, 3 (2021), e1403. DOI:https://doi.org/10.1002/widm.1403.

[23] Murugadoss, K., Rajasekharan, A., Malin, B., Agarwal, V., Bade, S., Anderson, J.R., Ross, J.L., Faubion, W.A., Halamka, J.D., Soundararajan, V. and Ardhanari, S. 2021. Building a best-in-class automated de-identification tool for electronic health records through ensemble learning. *Patterns*. 2, 6 (Jun. 2021), 100255. DOI:https://doi.org/10.1016/j.patter.2021.100255.

[24] Name Parser: 2023. *https://github.com/derek73/python-nameparser*. Accessed: 2023-08-11.

[25] Neild, R.C., Robinson, D. and Agufa, J. 2022. *Sharing Study Data: A Guide for Education Researchers*. U.S. Department of Education, Institute of Education Sciences.

[26] Office for Civil Rights 2012. *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule*. United States Department of Health & Human Services.

[27] Pedregosa, F. et al. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 12, (2011), 2825–2830.

[28] PII Masking: 2024. *https://huggingface.co/datasets/ai4privacy/pii-masking-200k*. Accessed: 2024-01-29.

[29] Pilán, I., Lison, P., Øvrelid, L., Papadopoulou, A., Sánchez, D. and Batet, M. 2022. The Text Anonymization Benchmark (TAB): A Dedicated Corpus and Evaluation Framework for

Text Anonymization. *Computational Linguistics*. 48, 4 (Dec. 2022), 1053–1101. DOI:https://doi.org/10.1162/coli_a_00458.

[30] PyMuPDF: 2022. *https://pymupdf.readthedocs.io/en/latest/intro.html#license-and-copyright*. Accessed: 2022-10-04.

[31] Remy, P. 2021. Name dataset. GitHub.

[32] Settles, B. 2019. Data for the 2018 Duolingo Shared Task on Second Language Acquisition Modeling (SLAM). Harvard Dataverse.

[33] Tedersoo, L., Küngas, R., Oras, E., Köster, K., Eenmaa, H., Leijen, Ä., Pedaste, M., Raju, M., Astapova, A., Lukner, H., Kogermann, K. and Sepp, T. 2021. Data sharing practices and data availability upon request differ across scientific disciplines. *Scientific Data*. 8, 1 (Jul. 2021), 192. DOI:https://doi.org/10.1038/s41597-021-00981-0.

[34] The Learning Agency Lab - PII Data Detection: *https://kaggle.com/competitions/pii-detection-removal-from-educational-data*. Accessed: 2024-04-19.

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

[36] What constitutes de-identified records and information? | Protecting Student Privacy: *https://studentprivacy.ed.gov/faq/what-constitutes-de-identified-records-and-information*. Accessed: 2024-02-12.

[37] 1974. *Family Education Rights and Privacy Act*.

[38] 2022. Presidio - data protection and anonymization API. Microsoft.