De-Identification of Sensitive Personal Data in Datasets Derived from IIT-CDIP

Stefan Larson[1,2*] **Nicole Cornehl Lima**[3] **Santiago Pedroza Díaz**[2,4]
Amogh Manoj Joshi[2,5] **Siddharth Betala**[2,6] **Jamiu Tunde Suleiman**[2,7]
Yash Mathur[2,8] **Kaushal Kumar Prajapati**[2] **Ramla Alakraa**[3]
Junjie Shen[3] **Temi Okotore**[3] **Kevin Leach**[1]

[1]Vanderbilt University [2]ML Collective [3]University of Michigan
[4]Universidad Panamericana [5]Arizona State University [6]IIT Madras
[7]Kyungpook National University [8]Carnegie Mellon University

Abstract

The IIT-CDIP document collection is the source of several widely used and publicly accessible document understanding datasets. In this paper, manual inspection of 5 datasets derived from IIT-CDIP uncovers the presence of thousands of instances of sensitive personal data, including US Social Security Numbers (SSNs), birth places and dates, and home addresses of individuals. The presence of such sensitive personal data in commonly-used and publicly available datasets is startling and has ethical and potentially legal implications; we believe such sensitive data ought to be removed from the internet. Thus, in this paper, we develop a modular data de-identification pipeline that replaces sensitive data with synthetic, but realistic, data. Via experiments, we demonstrate that this de-identification method preserves the utility of the de-identified documents so that they can continue be used in various document understanding applications. We will release redacted versions of these datasets publicly.

1 Introduction

Large volumes of data are becoming increasingly important for training machine learning models for document understanding tasks like classification, information extraction, and visual question answering. One such large volume is IIT-CDIP (Lewis et al., 2006) — containing over 7 million documents (~40 million pages) — which has been used as a source for several smaller, widely-used document understanding datasets like RVL-CDIP (Harley et al. (2015); 400k samples) and DocVQA (Mathew et al. (2021); 12,767 samples). In particular, these datasets have been used by the document understanding research community for benchmarking the performance of modern deep learning architectures, many of which make use of both image and text modalities (e.g., Xu et al. (2020, 2021); Huang et al. (2022); Appalaraju et al. (2021)).

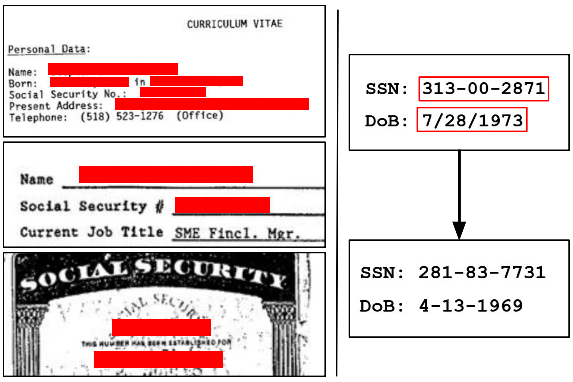


Figure 1: Left: snippets from documents from RVL-CDIP showing sensitive personal information, including US Social Security numbers. Right: example document de-identification with synthetic replacement.

Alarmingly, recent work by Larson et al. (2023) has reported large amounts of sensitive personally identifiable information (PII) in the RVL-CDIP document classification dataset. Such data would violate contemporary guidelines for responsible research in the NLP and machine learning communities, which have established that the presence of sensitive personal data is problematic and such data should be removed or minimized in datasets. For instance, the NeurIPS Code of Ethics states that "datasets should minimize the exposure of any personally identifiable information," and similar guidelines can be found in the ACL Rolling Review's Responsible NLP Research checklist.[1] Such statements are seriously motivated: the presence of one's sensitive personal information — like their Social Security number, birth date, place of birth, etc. — in publicly accessible data heightens their risk to fraud and identity theft (Sweeney, 2006).

*Corresponding author: stefan.larson@vanderbilt.edu.

[1]NeurIPS: neurips.cc/public/EthicsGuidelines and ACL Rolling Review: aclrollingreview.org/responsibleNLPresearch/ (accessed June 2024).

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 21494–21505
November 12-16, 2024 ©2024 Association for Computational Linguistics

Indeed, many organizations now take extra precautions against storing PII like SSNs.[2] Additionally, it has been demonstrated that diffusion models and large language models (LLMs) can leak sensitive data (Carlini et al., 2019, 2021, 2023; Lukas et al., 2023; Nasr et al., 2023), which is a serious concern if these types of models are to be trained on document understanding datasets like RVL-CDIP.

Inspired by the recent work by Larson et al. (2023) — who estimated that roughly 7.7% of documents in RVL-CDIP's `resume` category contain US Social Security numbers (SSNs) — we inspect RVL-CDIP and four other datasets derived from IIT-CDIP for the presence of sensitive PII. In this paper, we thoroughly investigate DocVQA, FUNSD, Tobacco3482, Tobacco800, and RVL-CDIP and determine the relative frequencies of various sensitive PII entities in each dataset. We find sensitive PII in all five of these datasets, including over 2,000 highly sensitive SSNs; Figure 1 displays several examples that contain highly sensitive SSNs. Seeking to re-align these datasets in the spirit of good data stewardship by minimizing the amount of sensitive personal information, we develop and analyze a data de-identification strategy that replaces original sensitive entities with synthetic, but realistic, replacement data (see Figure 1). This strategy aims to preserve the utility, or semantics, of a document while also removing potential privacy-related harms. To help reduce further privacy risks, we release de-identified versions of the datasets.[3]

2 Motivation: Presence of Sensitive PII in Datasets Derived from IIT-CDIP

In this section, we quantify the presence of sensitive personally identifiable information (PII) in 5 datasets derived from IIT-CDIP. We first introduce each of the 5 datasets. We next analyze the feasibility of using automated tools for detecting sensitive PII in documents. Finding these tools to be lacking, we discuss manual annotation. Finally, we quantify the presence of sensitive PII via manual annotation, motivating the need for the development of our document de-identification method, which we will discuss in Section 3.

[2]An example organizational policy from an American university: `mcneese.edu/policy/social-security-number-policy/`.

[3]`https://tinyurl.com/4vt844m9`.

2.1 Datasets Investigated

The datasets that we investigate in this paper are either wholly or in-part subsets of the IIT-CDIP document collection, which is a large collection of document images made publicly available as a result of legal settlements against several US tobacco companies. Documents from IIT-CDIP date from roughly the 1950s to the early 2000s, and are scanned copies of physical documents (i.e., they are *not* born-digital documents). The vast majority of the documents are in the English language.

The datasets investigated in this paper include: **RVL-CDIP** (Harley et al., 2015), consisting of 400,000 document images. This dataset is most often used as a classification dataset, and has 16 document categories, including `resume`. Larson et al. (2023) estimate that 7.7% of documents in the `resume` category contain SSNs. **Tobacco3482** (Kumar and Doermann, 2013), also often used as a document classification dataset; consists of 3,482 document images and 10 categories, including `resume`. **Tobacco800** (Zhu et al., 2007; Zhu and Doermann, 2007) is made up of 1,290 document images, and was originally created to evaluate signature detection algorithms. **FUNSD** (Jaume et al., 2019), an information extraction dataset containing 199 document images. **DocVQA** (Mathew et al., 2021), a visual question answering dataset consisting of 12,767 document images. The documents in DocVQA come from the UCSF Document Industry Library[4] of which IIT-CDIP is a subset.

2.2 Limitations of Automated Detection of Sensitive PII

Based on an initial sample of documents from RVL-CDIP, we developed a core set of sensitive PII entity types, listed in the left side of Table 1. To find and quantify the amount of sensitive PII in the 5 datasets derived from IIT-CDIP, we could ideally use automated tools. However, existing text-based tools like Presidio,[5] an open-source pattern-based tool; Amazon Comprehend;[6] Google DLP;[7] and Microsoft Azure's language service[8] are not equipped to support the detection most of these entity types, as shown in the right side of Table 1.

[4]`https://www.industrydocuments.ucsf.edu/`

[5]`https://github.com/microsoft/presidio`. We use version 2.2.33.

[6]`https://aws.amazon.com/comprehend/`

[7]`https://cloud.google.com/dlp`

[8]`https://learn.microsoft.com/en-us/azure/ai-services/language-service/`

PII Entity Type	Detectors
US Social Security number	Presidio Google Azure Amazon
birth date	Google
birth place	—
age	Google Azure Amazon
marital/parental/spousal status	—
home address	—
home phone number	—
religious affiliation	—
citizenship/nationality	—
sex/gender	Google
health status	—

Table 1: Sensitive personally-identifiable information (PII) entity categories found in the five datasets derived from IIT-CDIP. Not all entity categories are supported by off-the-shelf detectors (Presidio: ; Google: ; Azure: ; Amazon:).

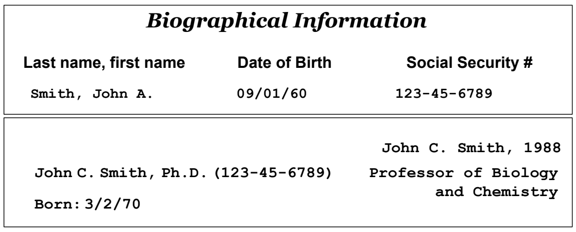
PII Type	Presidio	Azure	Amazon	Google
SSN	0.97	0.70	0.77	0.93
Birth Date	—	—	—	0.82
Age	1.00	1.00	1.00	1.00
Sex/Gender	—	—	—	1.00

Table 2: Automated detection performance measured in document-level recall.

Here, we see that SSN is the only type that all four tools can to detect, and 7 of the 11 PII types are not supported by any of the tools.

We compared the tools' detection performance on subsets of the annotated PII data by measuring document-level recall for each tool on 1,773 documents containing PII. Each tool is text-based, so we used Amazon Textract[9] to extract text from each document. Recall scores are displayed in Table 2. For SSNs, we see that Presidio is the highest performer with 0.97 document-level recall, followed by Google's DLP model at 0.93. The Azure and Amazon models perform considerably worse, with Azure failing to flag roughly 30% of documents that contain valid SSNs. The tools perform well on the Age and Sex/Gender types, perhaps because the contexts in which these entities appear are limited. Example error cases are displayed in Figure 2. Overall, we conclude our analysis of these tools by mentioning that (1) these off-the-shelf tools have limited support of PII (7 out of 11 of the PII types investigated in this paper are *not* supported); (2) tool detection performance is varied, with Azure and Amazon performing poorly on a highly critical PII type. Due to these limitations, we instead use manual annotation to quantify the amount of sensitive PII in the 5 datasets.

[9] https://aws.amazon.com/textract/.


Biographical Information

Last name, first name	Date of Birth	Social Security #
Smith, John A.	09/01/60	123-45-6789

John C. Smith, 1988
John C. Smith, Ph.D. (123-45-6789)
Professor of Biology and Chemistry
Born: 3/2/70

Figure 2: Example (reconstructed and with fake PII) SSN detection failures, apparently due to limited context window (top: Google), and missing context keyword (bottom: Amazon).

2.3 Manual Inspection

Sensitive Data Types. Prior work by Larson et al. (2023) identified the presence of four sensitive PII entities in RVL-CDIP by reviewing 1,000 samples from RVL-CDIP's `resume` category; these entities are US Social Security Numbers (SSNs), dates of birth, places of birth, and marital statuses. During our inspection of the five datasets derived from IIT-CDIP, we extended this set from four to the list shown in Table 1, which are entities that are considered sensitive by large organizations (e.g., universities) as well as the US government (e.g., DHS data privacy handbook (DHS Privacy Office, 2017)). Home addresses and home phone numbers are included in this list, and it is worth pointing out that we make a distinction between addresses in general and *home* addresses specifically, as well as *home* phone numbers instead of all phone numbers, as we consider the specific case to be sensitive while the more general case can encompass businesses and organizations (like universities), which are much less sensitive (indeed, these are publicly available). It is also worth pointing out that any instances of sensitive data found in the datasets that do not belong to any of the categories listed in Table 1 were categorized as "Other" (e.g., a portrait in a person's drivers license).

Inspecting the Documents. We manually reviewed the document images from DocVQA, FUNSD, Tobacco800, Tobacco3482, and RVL-CDIP. The first four of these datasets are relatively small in comparison to RVL-CDIP, so we inspected these document by document. RVL-CDIP is a much larger dataset, but we were able to break it down into more manageable chunks by using the fact that RVL-CDIP contains many duplicates and near-duplicates (as observed by Larson et al. (2023)). RVL-CDIP also consists of

Dataset	Size	*SSN*	*Birth Date*	*Birth Place*	*Age*	*Home Addr.*	*Home Ph.*	*M/P/S Status*	*Cit./Nat.*	*Sex/Gender*	*Health Status*	*Religion*	*N. Sensitive*
RVL-CDIP	400,000	2,342	12,800	6,125	219	3,908	1,801	4,228	2,647	602	60	43	15,956
Tobacco3482	3,482	9	62	29	—	6	4	18	19	3	—	—	66
Tobacco800	1,290	5	7	7	—	1	—	7	—	—	—	—	12
FUNSD	199	2	—	—	—	1	—	—	—	—	—	—	2
DocVQA	12,767	70	232	123	44	276	143	116	49	88	2	4	360
Total	417,738	2,428	13,101	6,284	263	4,192	1,948	4,369	2,715	670	62	47	16,396

Table 3: Counts of PII types for each dataset. RVL-CDIP is the largest of the datasets derived from IIT-CDIP, and contains the most PII.

16 distinguishable categories, with most of these categories being documents that were meant to be public-facing (e.g., press releases and newspaper clippings in the `news_article` category, newspaper and magazine advertisements in the `advertisement` category, scientific journal articles in `scientific_publication`, etc.) or less likely to contain sensitive personal data (e.g., materials specifications and data sheets in the `specification` category, blank file folders in `file_folder`, etc.). Thus, we were able to (1) avoid inspecting duplicates, and (2) spend less mental effort on public-facing or non-personnel related documents[10] and more effort on documents more likely to contain sensitive data (e.g., resumes). Our annotators were composed of a team of nine people, five of whom are co-authors of this paper. The lead author of this paper served as an "expert annotator" and organized the inspection process. Documents from FUNSD and Tobacco3482 were inspected by at least two annotators (including the expert annotator); DocVQA and Tobacco800 were inspected by the expert annotator; and roughly half of RVL-CDIP was inspected by two annotators, the rest being inspected by the expert only. We estimate inter-annotator agreement with 5 annotators on the Tobacco3482 and FUNSD datasets (judging whether a document contains sensitive PII or not), where we report a Fleiss' Kappa of 0.918. Alternatives to this process include crowdsourcing, but this is cost prohibitive and would defeat the purpose of limiting the exposure of sensitive data to the outside world.

Findings. Counts of each PII entity found in each dataset are summarized in Table 3. US SSNs, which we consider to be the most sensitive of the PII types, appear in all five datasets, with RVL-CDIP containing the most SSNs. Sensitive personal data was found in resumes, invoices, employee forms, and even in scanned images of drivers licenses and Social Security cards. The most common PII types are birth dates, birth places, home addresses, citizenship/nationality statuses, and marital/parental/spousal statuses. We observed that these often appear in resume documents, of which RVL-CDIP, DocVQA, and Tobacco3482 had many. The rarest PII types were health statuses and religious affiliation, which only appeared in RVL-CDIP and DocVQA. In total, we found over 16,000 documents containing sensitive PII, which is roughly 3.9% of all documents analyzed.

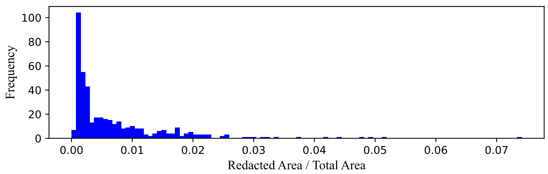


Figure 3: Distribution of redacted region ratios for sampled RVL-CDIP `resume` images.

3 De-Identification of Sensitive PII

The presence of sensitive PII in all five IIT-CDIP-derived datasets is alarming. For the sake of privacy, we will release redacted versions of these datasets along with the publication of this paper. In this section we discuss several ways of redacting sensitive information from document data, including one method that replaces the original sensitive data with synthetic data in order to preserve utility.

3.1 De-Identification Methods

Bounding boxes encircling sensitive textual entities within documents were applied as part of our

[10] A heuristic approach is used here. For instance, if a document is a magazine advertisement, then we do not need to thoroughly inspect it for sensitive personal data.

annotation stage. These bounding boxes encompass sensitive textual entities, so we can measure the amount of sensitive information per document based on the ratio of bounding box area to document page area. These ratios are summarized in Figure 3, where we see that for the majority of documents with sensitive information, only less than 2% of the document page area is occupied by sensitive information. With knowledge of where in 2-dimensional space the sensitive entities occur in documents, we can then apply various redaction strategies to remove and/or replace the sensitive data.

Basic redactions. This first strategy renders a black rectangle that covers sensitive data defined by the bounding boxes. The second strategy is the same except using white pixels instead of black. Examples of these two strategies are shown in the left and middle panes of Figure 4.

De-Identification with Replacement Data. Instead of simply redacting sensitive personal data by "covering" it with black or white (or inpainted) boxes in an image and removing it from OCR transcriptions, an alternative approach is data de-identification with synthetic replacement data. Data de-identification aims to replace original sensitive data with fake, but realistic, data.[11] As the sensitive PII data types that we seek to de-identify are relatively basic (e.g., birth dates, phone numbers, etc.), we can use data generation tools to synthesize replacement data. In this paper, we thus use two straightforward approaches to generate fake replacement data: (1) sampling from gazetteer lists (e.g., lists of nationalities, religions, months) and (2) the Faker[12] Python library, which is useful for generating data like phone number and SSN patterns, as well as home addresses.

For each document containing sensitive data, we first mask (i.e., redact) the original sensitive data by masking it with white pixels (or inpainting the sensitive regions, in the case of DocVQA), then we randomly generate fake replacement data using the two aforementioned approaches. Next, the fake data is rendered to images using the Pillow[13] Python image processing library, with which we use several font types. Since IIT-CDIP documents often contain noisy text as exhibited by Figure 5 — that is, text that appears visually degraded from noisy printing or scanning processes — we apply augmentations to the rendered fake data in order to mimic common noise types seen in the original documents. We use Augraphy[14] (Groleau et al., 2023) (a document-centric image augmentation library) and Albumentations[15] (Buslaev et al., 2020) (a general purpose image augmentation library) to achieve noise-like effects. In particular, we found Augraphy's `InkMottling` and `LowInkRandomLines` and Albumentation's `Rotate` augmentations to be useful. Examples of augmented data are shown in Figure 6. Put together, the redacted documents with fake, augmented data look like the examples shown in Figure 7.

The main benefit to this synthetic data replacement approach is to preserve the overall meaning of text in a document. That is, replacing one entity with a synthetic replacement still maintains the underlying semantic meaning, even if a different entity is used.

4 Experiments: Impact of Redactions

Having redacted sensitive PII from document images, we next determine whether such redactions impact downstream modeling performance. To help answer this question, we conduct several experiments to intrinsically and extrinsically compare the various redaction approaches with the original un-redacted documents.

4.1 Document Similarity

In our first experiment, we compute document similarity and distance scores between un-redacted and redacted versions of a random set of 445 `resume` documents form RVL-CDIP. Each of these documents has a corresponding un-redacted, black (i.e., black pixel redactions), white, and synthetic-replacement version. We compute embedding representations of each of these documents using CLIP with the ViT-32 model backbone (Radford et al., 2021). Then, we compute the cosine similarity and

[11]The term "de-identification" encompasses simple redaction of sensitive data, but in some sources the term also includes replacing the data with fake data (Yogarajan et al., 2018). In others, "pseudonymization" includes replacing original data with fake data (Volodina et al., 2023; Yermilov et al., 2023). Still others use pseudonymization to imply that there is a reversible mapping between real and fake data, so that the original data can be recovered from the fake data (Johner, 2019). In our case, we do not have a need for recovering the real data.

[12]`https://github.com/joke2k/faker`. We use version 23.1.0.

[13]`https://python-pillow.org/`. We use version 9.3.0.

[14]`https://github.com/sparkfish/augraphy`. We use version 8.2.6.

[15]`https://albumentations.ai/`. We use version 1.3.1.

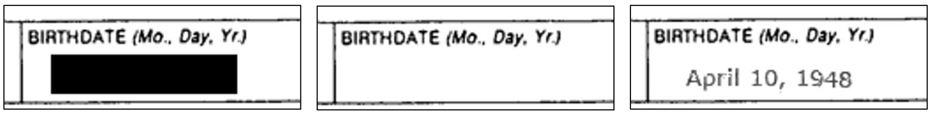


Figure 4: The three redaction approaches investigated: black (left), white (center), and pseudonymization with fake data (right).

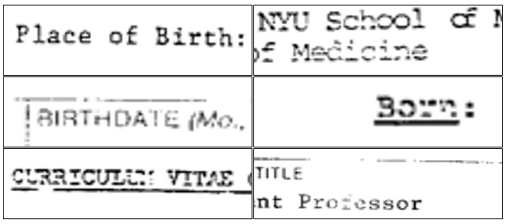


Figure 5: Examples of noise seen in documents from RVL-CDIP. (Best if viewed digitally.)

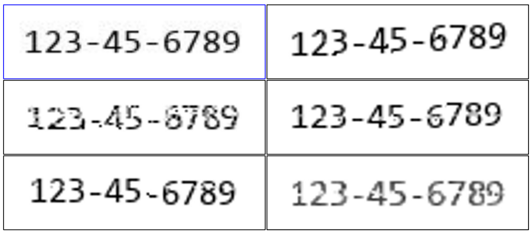


Figure 6: Various augmentations of un-augmented text (upper left). We use augmentations for the pseudonymized data. (Best if viewed digitally.)

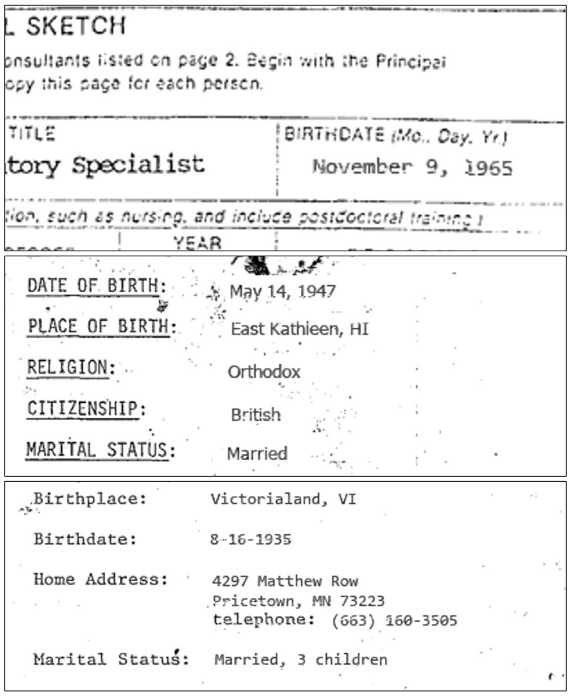


Figure 7: Examples of documents from RVL-CDIP pseudonymized by us. Our document pseudonymization method replaces real sensitive data with fake, augmented data. (Best if viewed digitally.)

euclidean distance between each redacted version and the un-redacted document.

Figure 8 shows a document with the three types of redactions applied, along with similarity and distance scores between each and the original un-redacted version. Figures 9 and 10, which plot histograms of similarity and distance scores between redacted and un-redacted documents, clearly indicate that the black redaction approach produces documents that are more dissimilar from their original un-redacted counterparts, at least in the CLIP embedding space. The distributions of scores of white and synthetic-replacement (called "pseudo" in the figures) redaction approaches are quite similar, with the synthetic-replacement documents being slightly more similar (and closer) to their un-redacted counterparts. Overall, the synthetic-replacement documents were more similar than the white redacted documents to the un-redacted versions roughly 61.6% of the time, while the opposite was true roughly 28.3% of the time; the similarity scores were equal roughly 10.1% of the time, and for no documents were the black redactions more similar to the un-redacted versions than the other two approaches. For the euclidean distance scores, these numbers were 61.8%, 28.3%, and 9.9%, respectively. Lastly, Figure 11 charts the relationship between redacted region area in a document image and similarity to its un-redacted original. Here, the relationship is quite clear: as the area occupied by redactions increases in a document image, the less that redacted image resembles the original.

4.2 Downstream Model Confidence

In our second experiment, we compare the predictions of a document classifier on original un-redacted documents and their redacted counterparts. Our goal here is to examine the impact of redactions (quantified, as before, by the ratio between redacted area to total image area) on modeling performance on the task of document classification. Model label predictions are critical for the task of classification, and thus it would be concerning if these predictions were to change on the redacted

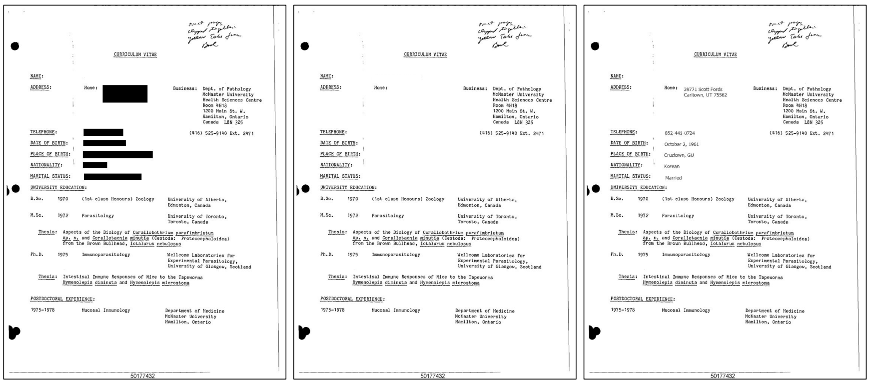


Figure 8: Comparison of different redaction methods. Redacting sensitive personal data using black redactions (left) typically causes the redacted image to be more dis-similar to the original document image (not shown) in terms of the image embedding space than white redactions (center) and synthetic-replacement redactions (right). Here, Sim is cosine similarity and Dist is euclidean distance in CLIP (ViT-32) embedding space between the redacted image and the original un-redacted version.

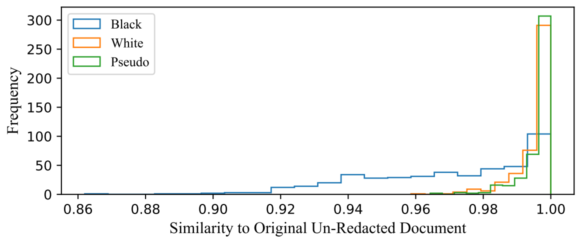


Figure 9: Similarity score distributions for three redaction types. Each similarity score is between a redacted document image and its un-redacted original counterpart.

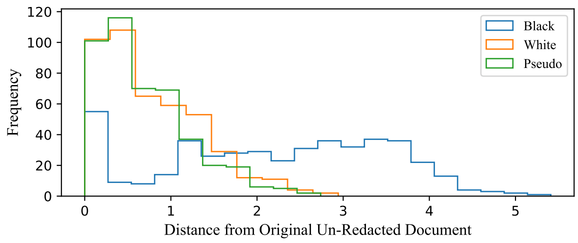


Figure 10: Distance score distributions for three redaction types. Each distance score is between a redacted document image and its un-redacted original counterpart.

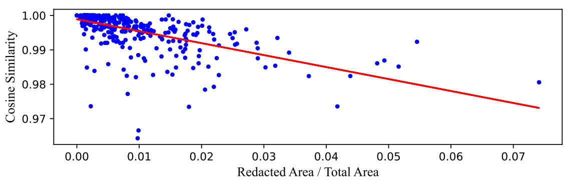


Figure 11: The relationship between redacted region area and cosine similarity between original and redacted (synthetic-replacement) document pairs in CLIP embedding space.

versions of the documents. Similarly, model confidence scores are important, as they can be used as measures of certainty in the presence of out-of-domain or out-of-distribution inputs (Larson et al., 2022). We use the DiT-base document classification model (Li et al., 2022), using weights that have been fine-tuned on the RVL-CDIP training set.[16] We then sample 445 documents from the test split of RVL-CDIP's `resume` category that contain PII, and compute model predictions and softmax confidence scores on un-redacted and redacted versions of each sample (repeated for each redaction method: black, white, and synthetic-replacement).

None of the sampled images saw DiT model predictions change (e.g., flipped from `resume` to `invoice`) for all three redaction types. However, we do observe a slightly increasing trend in the relationship between the relative cumulative areas of redacted bounding box region in an image against the absolute model confidence score difference between un-redacted and redacted versions of an image. This relationship is also captured in Figure 12 for the synthetic-replacement data, where we removed 11 outliers with confidence score differences of above 0.004 for visual clarity. These outliers are interesting: in the most extreme case, we observed a datapoint with a relative area of redacted PII of about 3% that yielded a confidence score decrease of roughly 0.30 from un-redacted to redacted. However, these cases are rare, and the

[16]Specifically, we use the model from https://huggingface.co/microsoft/dit-base-finetuned-rvlcdip.

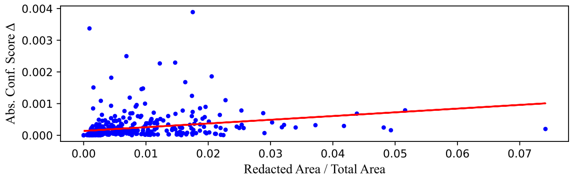


Figure 12: The relationship between redacted region relative area and confidence score difference.

median confidence score difference among these 11 outliers is 0.045.

The relationship between redacted area ratio and classifier confidence score difference for the synthetic-replacement redaction type is similar to that of the white type. The trend for the black type redacted data is similar as well, but the confidence score differences for the black type tend to be larger than the other types. (The mean differences for black, white, and synthetic-replacement data are 0.0036, 0.0022, and 0.0024, respectively.) Overall, though, the impact of redactions on model labeling ability and confidence score is very minimal.

5 Related Work

With the growing importance of data in training and evaluating machine learning models, recent work has started to examine datasets and corpora with the goal of evaluating data quality. This notion of quality encompasses notions of correctness (e.g., uncovering and quantifying label annotation errors (Chen et al., 2016; Radenović et al., 2018; Niu and Penn, 2019; Northcutt et al., 2021; Ying and Thomas, 2022)), and diversity and difficulty (e.g., measuring the similarity or overlap between test and train splits (Croft et al., 2023; Elangovan et al., 2021; Finegan-Dollak et al., 2018; Allamanis, 2019; Barz and Denzler, 2020; Laatiri et al., 2023; Lewis et al., 2021; Larson et al., 2023; Wen et al., 2022)), as well as whether datasets are free of potentially harmful contents like undesirable social biases (e.g., Yang et al. (2020); Hirota et al. (2022); Sahoo et al. (2022); Smith et al. (2023)), harmful language or concepts (e.g., hate speech and sexually explicit content (Birhane and Prabhu, 2021; Luccioni and Viviano, 2021)), and sensitive personally identifiable data (Murgia, 2019; Yang et al., 2020; Prabhu and Birhane, 2020; Harvey and LaPlace, 2021; Yang et al., 2022). In particular, Subramani et al. (2023) uncovered the presence of PII in two massive web-scale corpora using automated methods, including Presidio. Recent work by Larson et al. (2023) observed the presence of sensitive PII in RVL-CDIP, but did not investigate this in-depth or for other IIT-CDIP-derived corpora. Our work is among the first to examine the presence of sensitive personal data in datasets derived from IIT-CDIP in depth.

De-identification and pseudonymization have been investigated in computer vision in the context of de-identifying and pseudonymizing faces (e.g., Gross et al. (2009); Brkic et al. (2017); Li and Lin (2019); Gafni et al. (2019); Cai et al. (2024); Yang et al. (2022)) and other potentially sensitive image entities in images (e.g., Orekondy et al. (2018)). De-identification and pseudonymization also plays an important role in healthcare applications of natural language processing (e.g., Friedrich et al. (2019); Lothritz et al. (2023); Vakili and Dalianis (2022); Sánchez et al. (2014); Murugadoss et al. (2021)), where healthcare records — often legally — must be anonymized before they can be used as data for training or evaluating models. Despite de-identification and pseudonymization being active areas of research in computer vision and healthcare text processing applications, we found relatively little work on the topic in prior work on document image processing, with exceptions including work on applying black redactions over sensitive regions of documents (Liu et al., 2019; Pagel et al., 2024). In particular, Pagel et al. (2024) investigated the impact of applying black redaction boxes to documents from RVL-CDIP, finding that this type of redaction strategy tends to negatively impact classifier performance. However, they did not investigate any other redaction or de-identification strategies.

We therefore argue that sensitive data de-identification and pseudonymization should become active areas of future work for the document image processing research community; for instance, while we found simple data generation and augmentation strategies using Augraphy and Faker to be effective in imitating real data, perhaps more recent font style transfer methods (e.g., Atarsaikhan et al. (2017); Gomez et al. (2019); Wu et al. (2019)) could be investigated for this task in the future.

6 Conclusion

In this paper, we uncover large amounts of sensitive personally identifiable information (PII) within five datasets derived from IIT-CDIP. We measure the coverage and performance of several off-the-shelf PII detection tools, finding that performance varies widely across tools at the task of detecting SSNs, and that as a whole, coverage of the various PII types found in IIT-CDIP is limited. We analyze the impact of redactions on the data to various modeling tasks, and we will make redacted versions of these dataset publicly available.

Limitations

The primary goal of this work is to minimize the immediate risk of there being sensitive personal data in several well-known, publicly available datasets. In order to take immediate action to remedy this risk, we introduce a method to replace this sensitive data with synthetic, but realistic, replacement data. This work is therefore less concerned with rigorous privacy notions (e.g., differential privacy).

Due to the scale of the original datasets, there is a small chance that our manual annotation process may have been imperfect. However, we believe that all (or almost all) of the highly sensitive PII entities (e.g., US Social Security Numbers) have been de-identified. Regardless, since we have redacted so much data with data replacement, any original sensitive entities are now "hidden in plain sight", and potential malevolent actors will have difficulty finding real sensitive entities given the plethora of realistic-looking synthetic entities.

We only apply our de-identification method to datasets derived from IIT-CDIP, which are scanned images of printed documents. However, we believe our de-identification method could be applied to born-digital documents as well. Indeed, such documents may be easier to deal with since they contain less noise.

While we will make our de-identified datasets publicly available, the original un-redacted versions are still currently hosted on the web. Prior to the publication of this paper we will be contacting the hosts of these datasets to share our findings in the hopes that these un-redacted datasets will be taken down. We have already had some success on this front with Hugging Face, where we convinced this site to remove a public preview of the RVL-CDIP dataset.[17]

Acknowledgements

We thank the anonymous reviewers for their helpful feedback on this paper. This work is supported by Vanderbilt University's Lacy-Fischer grant as well as gift funds from Amazon. We thank the University of Michigan's UROP program for supporting the work of NCL, RA, JS, and TO.

[17]Hugging Face's "Dataset Viewer" is disabled for the RVL-CDIP (aharley/rvl_cdip) dataset. See https://huggingface.co/datasets/aharley/rvl_cdip/discussions/4.

References

Miltiadis Allamanis. 2019. The adverse effects of code duplication in machine learning models of code. In *Proceedings of the 2019 ACM SIGPLAN International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software*.

Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R. Manmatha. 2021. DocFormer: End-to-end transformer for document understanding. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*.

Gantugs Atarsaikhan, Brian Kenji Iwana, Atsushi Narusawa, Keiji Yanai, and Seiichi Uchida. 2017. Neural font style transfer. In *Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*.

Björn Barz and Joachim Denzler. 2020. Do we train on test data? Purging CIFAR of near-duplicates. *Journal of Imaging*, 6(6).

Abeba Birhane and Vinay Uday Prabhu. 2021. Large image datasets: A pyrrhic win for computer vision? In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*.

Karla Brkic, Ivan Sikiric, Tomislav Hrkac, and Zoran Kalafatic. 2017. I know that person: Generative full body and face de-identification of people in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. 2020. Albumentations: Fast and flexible image augmentations. *Information*, 11(2).

Zikui Cai, Zhongpai Gao, Benjamin Planche, Meng Zheng, Terrence Chen, M. Salman Asif, and Ziyan Wu. 2024. Disguise without disruption: Utility-preserving face de-identification. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI)*.

Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. 2023. Extracting training data from diffusion models. In *Proceedings of the 32nd USENIX Conference on Security Symposium*.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *Proceedings of the 28th USENIX Conference on Security Symposium*.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *Proceedings of the 30th USENIX Conference on Security Symposium*.

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A thorough examination of the CNN/Daily Mail reading comprehension task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Roland Croft, M. Ali Babar, and M. Mehdi Kholoosi. 2023. Data quality for software vulnerability datasets. In *Proceedings of the 45th International Conference on Software Engineering (ICSE)*.

DHS Privacy Office. 2017. Handbook for safeguarding sensitive PII: Privacy policy directive 047-01-007, revision 3. *DHS Privacy Office*.

Aparna Elangovan, Jiayuan He, and Karin Verspoor. 2021. Memorization vs. generalization: Quantifying data leakage in NLP performance evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Catherine Finegan-Dollak, Jonathan K. Kummerfeld, Li Zhang, Karthik Ramanathan, Sesh Sadasivam, Rui Zhang, and Dragomir Radev. 2018. Improving text-to-SQL evaluation methodology. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Max Friedrich, Arne Köhn, Gregor Wiedemann, and Chris Biemann. 2019. Adversarial learning of privacy-preserving text representations for de-identification of medical records. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Oran Gafni, Lior Wolf, and Yaniv Taigman. 2019. Live face de-identification in video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Raul Gomez, Ali Furkan Biten, Lluis Gomez, Jaume Gibert, MarRusiñol, and Dimosthenis Karatzas. 2019. Selective style transfer for text. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*.

Alexander Groleau, Kok Wei Chee, Stefan Larson, Samay Maini, and Jonathan Boarman. 2023. Augraphy: A data augmentation library for document images. In *Proceedings of the 17th International Conference on Document Analysis and Recognition (ICDAR)*.

Ralph Gross, Latanya Sweeney, Jeffrey Cohn, Fernando de la Torre, and Simon Baker. 2009. Face de-identification. In Andrew Senior, editor, *Protecting Privacy in Video Surveillance*. Springer London, London.

Adam W. Harley, Alex Ufkes, and Konstantinos G. Derpanis. 2015. Evaluation of deep convolutional nets for document image classification and retrieval. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*.

Adam Harvey and Jules LaPlace. 2021. Exposing.ai.

Yusuke Hirota, Yuta Nakashima, and Noa Garcia. 2022. Gender and racial bias in visual question answering datasets. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. LayoutLMv3: Pre-training for document AI with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*.

Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. FUNSD: A dataset for form understanding in noisy scanned documents. In *Proceedings of the 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*.

Christian Johner. 2019. Anonymization and pseudonymization. *Johner Institute*.

Jayant Kumar and David Doermann. 2013. Unsupervised classification of structurally similar document images. In *Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR)*.

Seif Laatiri, Pirashanth Ratnamogan, Joël Tang, Laurent Lam, William Vanhuffel, and Fabien Caspani. 2023. Information redundancy and biases in public document information extraction benchmarks. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*.

Stefan Larson, Gordon Lim, Yutong Ai, David Kuang, and Kevin Leach. 2022. Evaluating out-of-distribution performance on document image classifiers. In *Proceedings of the 36th Conference on Neural Information Processing Systems Track on Datasets and Benchmarks*.

Stefan Larson, Gordon Lim, and Kevin Leach. 2023. On evaluation of document classification with RVL-CDIP. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

D. Lewis, G. Agam, S. Argamon, O. Frieder, D. Grossman, and J. Heard. 2006. Building a test collection for complex document information processing. In *Proceedings of the 29th Annual International ACM SIGIR Conference*.

Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2021. Question and answer test-train overlap in open-domain question answering datasets. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Junlong Li, Yiheng Xu, Tengchao Lv, Lei Cui, Cha Zhang, and Furu Wei. 2022. DiT: Self-supervised pre-training for document image transformer. In *Proceedings of the 30th ACM International Conference on Multimedia*.

Tao Li and Lei Lin. 2019. AnonymousNet: Natural face de-identification with measurable privacy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

Chuanyi Liu, Peiyi Han, Yingfei Dong, Hezhong Pan, Shaoming Duan, and Binxing Fang. 2019. Cloud-DLP: Transparent and automatic data sanitization for browser-based cloud storage. In *Proceedings of the 28th International Conference on Computer Communication and Networks (ICCCN)*.

Cedric Lothritz, Bertrand Lebichot, Kevin Allix, Saad Ezzini, Tegawendé Bissyandé, Jacques Klein, Andrey Boytsov, Clément Lefebvre, and Anne Goujon. 2023. Evaluating the impact of text de-identification on downstream NLP tasks. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*.

Alexandra Luccioni and Joseph Viviano. 2021. What's in the box? an analysis of undesirable content in the Common Crawl corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*.

Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. 2023. Analyzing leakage of personally identifiable information in language models. In *Proceedings of the 2023 IEEE Symposium on Security and Privacy (SP)*.

Minesh Mathew, Dimosthenis Karatzas, and C.V. Jawahar. 2021. DocVQA: A dataset for vqa on document images. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*.

Madhumita Murgia. 2019. Who's using your face? the ugly truth about facial recognition. *Financial Times Magazine*.

Karthik Murugadoss, Ajit Rajasekharan, Bradley Malin, Vineet Agarwal, Sairam Bade, Jeff R. Anderson, Jason L. Ross, William A. Faubion, John D. Halamka, Venky Soundararajan, and Sankar Ardhanari. 2021. Building a best-in-class automated de-identification tool for electronic health records through ensemble learning. *Patterns*, 2(6):100255.

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. 2023. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*.

Jingcheng Niu and Gerald Penn. 2019. Rationally reappraising ATIS-based dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. 2021. Pervasive label errors in test sets destabilize machine learning benchmarks. In *Proceedings of the 35th Conference on Neural Information Processing Systems Track on Datasets and Benchmarks*.

Tribhuvanesh Orekondy, Mario Fritz, and Bernt Schiele. 2018. Connecting pixels to privacy and utility: Automatic redaction of private information in images. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Johannes Pagel, Stefanie Vogl, and Laura Israel. 2024. Analyzing the impact of redaction on document classification performance of deep CNN models. *OSF Preprints*.

Vinay Uday Prabhu and Abeba Birhane. 2020. Large datasets: A pyrrhic win for computer vision? *arXiv preprint arXiv:2006.16923*.

Filip Radenović, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. 2018. Revisiting Oxford and Paris: Large-scale image retrieval benchmarking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*.

Nihar Sahoo, Himanshu Gupta, and Pushpak Bhattacharyya. 2022. Detecting unintended social bias in toxic language datasets. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*.

David Sánchez, Montserrat Batet, and Alexandre Viejo. 2014. Utility-preserving privacy protection of textual healthcare documents. *Journal of Biomedical Informatics*, 52:189–198. Special Section: Methods in Clinical Research Informatics.

Brandon Smith, Miguel Farinha, Siobhan Mackenzie Hall, Hannah Rose Kirk, Aleksandar Shtedritski, and Max Bain. 2023. Balancing the picture: Debiasing vision-language datasets with synthetic contrast sets. *arXiv preprint arXiv:2305.15407*.

Nishant Subramani, Sasha Luccioni, Jesse Dodge, and Margaret Mitchell. 2023. Detecting personal information in training corpora: an analysis. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP)*.

Latanya Sweeney. 2006. Protecting job seekers from identity theft. *IEEE Internet Computing*, 10(2):74–78.

Thomas Vakili and Hercules Dalianis. 2022. Utility preservation of clinical text after de-identification. In *Proceedings of the 21st Workshop on Biomedical Language Processing*.

Elena Volodina, Simon Dobnik, Therese Lindström Tiedemann, and Xuan-Son Vu. 2023. Grandma karl is 27 years old - research agenda for pseudonymization of research data. In *Proceedings of the Ninth International Conference on Big Data Computing Service and Applications (BigDataService)*.

Yuqiao Wen, Guoqing Luo, and Lili Mou. 2022. An empirical study on the overlapping problem of open-domain dialogue datasets. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*.

Liang Wu, Chengquan Zhang, Jiaming Liu, Junyu Han, Jingtuo Liu, Errui Ding, and Xiang Bai. 2019. Editing text in the wild. In *Proceedings of the 27th ACM International Conference on Multimedia*.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 2021. LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. LayoutLM: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*.

Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. 2020. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT)*.

Kaiyu Yang, Jacqueline Yau, Li Fei-Fei, Jia Deng, and Olga Russakovsky. 2022. A study of face obfuscation in ImageNet. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*.

Oleksandr Yermilov, Vipul Raheja, and Artem Chernodub. 2023. Privacy- and utility-preserving NLP with anonymized data: A case study of pseudonymization. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*.

Cecilia Ying and Stephen Thomas. 2022. Label errors in BANKING77. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*.

Vithya Yogarajan, Michael Mayo, and Bernhard Pfahringer. 2018. A survey of automatic de-identification of longitudinal clinical narratives. *arXiv preprint arXiv:1810.06765*.

Guangyu Zhu and David Doermann. 2007. Automatic document logo detection. In *Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR)*.

Guangyu Zhu, Yefeng Zheng, David Doermann, and Stefan Jaeger. 2007. Multi-scale structural saliency for signature detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.