Detecting Personal Information in Training Corpora: an Analysis

Nishant Subramani *
AI2, USA
nishant.subramani23@gmail.com

Alexandra Sasha Luccioni *
Hugging Face, Canada
sasha.luccioni@huggingface.co

Jesse Dodge
AI2, USA

Margaret Mitchell
Hugging Face, USA

Abstract

Large language models are trained on increasing quantities of unstructured text, the largest sources of which are scraped from the Web. These Web scrapes are mainly composed of heterogeneous collections of text from multiple domains with minimal documentation. While some work has been done to identify and remove toxic, biased, or sexual language, the topic of *personal information* (PI) in textual data used for training Natural Language Processing (NLP) models is relatively under-explored. In this work, we draw from definitions of PI across multiple countries to define the first PI taxonomy of its kind, categorized by type and risk level. We then conduct a case study on the Colossal Clean Crawled Corpus (C4) and the Pile, to detect some of the highest-risk personal information, such as email addresses and credit card numbers, and examine the differences between automatic and regular expression-based approaches for their detection. We identify shortcomings in modern approaches for PI detection, and propose a reframing of the problem that is informed by global perspectives and the goals in personal information detection.

1 Introduction

The problem of identifying personal information (PI) on the Web is increasingly critical as larger and larger datasets, built by scraping data from the Internet, are made publicly available and used to train machine learning (ML) models (Raffel et al., 2019; Gao et al., 2020; Volske et al., 2017). While the extent to which this information is memorized by Natural Language Processing (NLP) models is largely under-explored, recent work has shown that it is possible to extract specific examples of PI from trained language models such as email addresses, phone numbers, and physical addresses via prompting (Carlini et al., 2019, 2020), while complementary work has shown that it is also possible to steer pretrained models to generate arbitrary sequences without modifying the underlying weights at all via steering vectors (Subramani et al., 2019; Subramani and Suresh, 2020; Subramani et al., 2022) and prompting (Shin et al., 2020; Li and Liang, 2021).

This suggests that it is necessary to better understand the types of PI contained in training corpora and the types of harms that they can cause, and to propose ways for automatically detecting (and, eventually, removing) the most high-risk types of PI from NLP corpora. We endeavor to address both of these directions in the current article: we start with defining different types of PI and propose a novel categorization in Section 2 and discuss the risks of different types of PI. Then, in Section 3, we explore the difficulty in detecting one of the highest-risk and easiest-to-identify types of PI, CHARACTER-BASED identifiers, comparing a model-based PI detection tool, Presidio (Microsoft, 2021) and a simple regular-expression-based approach on the Colossal Clean Crawled Corpus (C4) (Raffel et al., 2019) and the Pile (Gao et al., 2020). We present our results in Section 4 – these suggest that some of the highest risk PI are currently not well-captured in modern tools, opening immense risk to individuals who require anonymity in data. We discuss related efforts and promising research directions in Section 5, and conclude with a discussion of our results and propose ways forward to improve the extent to which our field takes PI into account in Section 6.

2 Types of Personal Information

2.1 Classes of Personal Information

The very definition of what constitutes personal information varies, with vague and often conflicting definitions proposed depending on regions and contexts, ranging from Personally Identifying Information, or *PII*, defined in the United States (Ex-

*Both authors contributed equally to this research

Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023), pages 208–220

ecutive Office of the President, 2006; United States Department of Defense, 2007; Office of the Secretary of Defense, 2007), to *personal data*, defined by the U.N. (UN High-Level Committee on Management (HLCM), 2018), the U.K. (Dtaa Protection Act, 2018), the E.U. (Summary, 2020), and Brazil (of Brazil, 2020); as *personal information* in China (Creemers, Rogier and Webster, Graham, 2021), Australia (Commonwealth Consolidated Acts, 1988) and South Africa (South African Parliament, 2013). It is therefore important to formally define these categories of personal information, in order to better understand their levels of risk and how they can enable unique identification.

In our proposed categorization of personal information, we distinguish:

Birth-centered characteristics true of a person at birth, most of which are difficult or impossible to change, such as nationality, gender, caste, etc.

Society-centered include characteristics that commonly develop throughout a person's life and are defined in many countries as a specially-designated "status", such as immunization status.

Social-based categories are categories that follow from the definitions outlined, but are rarely given as examples. These categories are discussed in Social Identity Theory (Hogg, 2020), and Self-categorization theory (SCT) (Turner, 2010), corresponding to social groups such as teams or affiliations – e.g. member of the women's softball team, student of Carnegie Mellon University.

Character-based categories are sequences of letters and numbers that can often uniquely identify a person or a small group of people; they change relatively infrequently and can therefore persist as sources of identification for years or decades – e.g. a credit card number, IBAN, or e-mail address.

Records-based information typically consists of a persistent document or electronic analog that is not generally-available, but can allow for the (reasonable) identification of an individual – e.g. financial or health records.

Situation-based is basic information that can be used to pinpoint a specific situation, or be combined with other categories to uniquely identify an individual, but that is restricted to a given context or point in time – e.g. date, time, GPS location.

2.2 Risks of Personal Information

When PI of the types described above are widely disseminated, it can open the door to a series of harms, ranging from identity theft (Irshad and Soomro, 2018) to discrimination based on sensitive characteristics (Kang et al., 2016; Bertrand and Mullainathan, 2004). Individuals may also desire to keep their PI private to escape harmful situations or to block psychologically traumatic interactions; people with stalkers, victims of domestic abuse, and other situations where a person is a direct target of another person to inflict emotional or psychological harm need to be able to remove trails for contacting them. The dissemination of different types of PI therefore exposes individuals to different *risk levels*, which we introduce below:

Low Risk Only applies to a large group of people without uniquely identifying an individual or small group.

Medium Risk Applies to a small group of people without providing sensitive information and does not uniquely identify an individual.

High Risk Uniquely identifies an individual[1] or applies to a small group of people with exposed sensitive information.

Extreme Risk Uniquely identifies an individual and provides sensitive information about them.

Based on the classes of personal information described in section 2.1, the "CHARACTER-BASED" class is one of the most critical classes in terms of risk exposure. This class includes information such as credit card numbers, international bank account numbers (IBAN), and U.S. social security numbers, which have a high risk for harm if not appropriately obfuscated, such as being used for identity theft, scamming, or loss of wealth (see Section 2.2). Similarly, they have high exposure levels, uniquely identifying a single person, or in some cases just a few people (such as when a phone number or email address is shared). However, most of the personal information in this class consists of alphanumeric sequences that follow predefined conventions, making them difficult but not impossible to identify in text [2]. This is why we focused on these CHARACTER-BASED forms of PI for our case study, aiming to identify the PI that puts individuals most at risk but that can be identified programmatically. We describe our approach in the section below.

[1] Unique identifiers as used here may also be identifiers that can also apply to multiple people, such as when a couple shares a personal email address.

[2] Given that most phone numbers have between 8 and 10 digits, there are roughly between 10^8 to 10^{10} possible combinations of numbers.

3 Case Study: Personal Information in the Common Crawl and the Pile

To estimate the quantity of high-risk CHARACTER-BASED personal information in two popular corpora, we run both an out-of-the-box personal information detection tool and a regular-expression based approach on them. We present the methodology that we adopt, our evaluation approach, and our results in the current section.

3.1 Types of Character-based Personal Information

We choose the following subset of the Character-based personal information types for detection, based on their potential for risk and identification:
NAME: a series of one or several names that uniquely identify an individual.
PHONE NUMBER: a series of digits that may include: a country or region code, a three-digit area code, a three-digit central office code, and four digits for the line number.
EMAIL ADDRESS: which are typically composed of 4 parts: the prefix, the @ sign, the domain provider, and the suffix (e.g., *johndoe* + @ + *yahoo* + *.com*).
U.S. SOCIAL SECURITY NUMBER (SSN): SSNs are used in the US as centralized numbers, both for taxation and identification purposes. They are composed of nine digits, divided into three parts (area, group, and serial number) and are necessary for activities such as opening bank accounts.
CREDIT CARD NUMBER: credit cards such as Visa and MasterCard are composed of 8 to 19 digits, with a part of the number identifying the industry, the issuer, and the account itself. The final digit of credit card is calculated using the Luhn algorithm, which is a checksum formula used to validate identification numbers (Wikipedia contributors, 2021).
INTERNATIONAL BANK ACCOUNT NUMBER (IBAN): an international system for identifying bank accounts made of a sequence of up to 34 numbers, constituted of a country code, two check digits, the account number and routing information, with check digits calculated using MOD-97-10.
U.S. BANK ACCOUNT NUMBER: composed of 8 to 17 digits, and used internally by US financial institutions to transfer funds between accounts.
INTERNET PROTOCOL ADDRESS (IP ADDRESS): a numerical label used to identify a device that is connected to a computer network that uses the Internet Protocol for communication.

3.2 Datasets Analyzed in our Study

In this work we analyze two corpora created from a scrape of the Internet: the Colossal Clean Crawled Corpus (Raffel et al., 2019) and the Pile (Gao et al., 2020). We first describe them below:

The Colossal Clean Crawled Corpus (C4) C4 is one of the largest language datasets, consisting of over 365 million documents with a total of 173 billion tokens (using the GPT-2 tokenizer (Black et al., 2022)) originally collected from the Internet by Raffel et al. (Raffel et al., 2019), and subsequently used to train models like T5 and the Switch Transformer (Fedus et al., 2021). This corpus consists of text taken from Common Crawl then passed through a number of filters with the intention of retaining high-quality English text The `C4-en` validation set of the C4 dataset that we analyzed was created by taking the April 2019 snapshot of Common Crawl corpus and applying a number of filters, such as discarding documents that have obscene words, those that contain placeholder text, or those that are less than five sentences long.

The Pile The Pile (Gao et al., 2020) is a composite English dataset that consists of 22 smaller datasets — such as PubMed, OpenWebText2, OpenSubtitles, and YoutubeSubtitles — that were combined during its creation, resulting in text from a variety of genres including science, law, research papers, mathematics, books, subtitles, patents, and philosophy. Certain portions of the dataset were filtered including some deduplication and language-based filtering to keep only English text. It contains 383 billion tokens (based on the GPT-2 tokenizer (Black et al., 2022)) and was explicitly designed to aid in the training of large-scale LMs and has been used for this purpose since its creation.

3.3 Personal Information Detection Methods

Many existing ML-based techniques for detecting PI are Named Entity Recognition (NER) inspired, relying heavily on regular expressions, which can be hand-crafted to correspond to kinds of information and achieve fair accuracy on specific types of PII (Aura et al., 2006). There are also several language-specific tools for detecting PI in written text, such as PIICatcher and Poverty Action's PII Detection tool, which rely on approaches ranging from pattern-matching to statistical models to detect different types of PI. However, these tools frequently only work on structured sources of data

PI type	Presidio Count-C4	Reg Ex Count-C4	Presidio Count- Pile	Reg Ex Count- Pile
PHONE NUMBER	19,592,273	22,349,098	23,191,595	74,421,644
EMAIL ADDRESS	9,056,833	8,707,343	13,336,793	13,827,399
US BANK NUMBER	7,139,838	N/A	69,763,678	N/A
US SSN	2,352,339	5,344,044	12,541,022	60,976,242
IP ADDRESS	1,890,090	1,425,070	14,975,663	9,334,985
CREDIT CARD	61,405	344,771	741,815	19,092,364
IBAN CODE	4,777	53,806	7,601	1,637,235
NAME	1,444,683,066	N/A	3,273,163,949	N/A
TOTAL	**1,484,780,621**	**38,224,132**	**3,407,722,116**	**179,722,808**

Table 1: Types of PI and their counts in C4 and Pile, as detected by Presidio and Regular Expressions.

such as tables and dataframes.

Of the existing tools that can detect different types of PI in textual data, **Presidio** is the only tool that is able to identify entities in unstructured text using both pattern-based matching as well as ML models trained on labeled data. Most importantly, Presidio is able to detect CHARACTER-BASED types of personal information such as credit cards and phone numbers, which we have identified as the types of PI that have the highest risk. As a baseline comparison, we also adopted a **regular-expression** (regex)-based approach for detecting the same types of character-based entities– we define the regexes we used in Section 6 [3].

4 Results

We first ran Presidio and our set of regular expressions to detect the different kinds of personal information listed in the previous section on the entirety of C4 and the Pile. In order to validate these results, we then manually verified the top 100 documents with the most detected PI, as well as a random sample of 2800 entities detected by the two approaches. We present our results in the sections below.

4.1 Detected Personal Information Counts

Running both Presidio and the set of regular expressions on all of the 364,868,892 documents of `C4-EN` and 210,607,728 documents of `Pile`, we detected millions of instances of personal information, which we present in Table 1. While we cannot meaningfully compare the total number of PI the two approaches detected, we can compare them per-type: both approaches detected a comparable amount of email addresses in both datasets. However, regular expressions systematically captured more instances of PI than Presidio for phone numbers, US SSNs, credit cards, and IBAN codes with between 1.2 and 1000 times more detections. For IP addresses, Presidio detected about 1.5 times as many instances as regular expressions. Finally, Presidio detected almost 1.4 billion names and 7 million US bank numbers for C4 and almost 3.3 billion names and 70 million US bank numbers for the Pile, indicating that these are highly prevalent – however, we were not able to define a meaningful regular expression baseline for these types so we lack a baseline. Comparing the two datasets, C4 seems to have fewer instances of PI across the board, even though there are more documents in the dataset. However, these counts alone are hard to interpret, since we do not know what the precision and recall are for each approach and each type of PI. This is why proceeded to do a manual verification of the top 100 documents with the most detections, which we describe below.

4.2 Manual Audit of Documents from C4

An ideal exhaustive study of PI in our target datasets could envisage employing crowdworkers to fully annotate every detection made by both tools. However, this would expose personal information publicly, further amplifying and propagating content where consent to share may be missing and there may be harmful ramifications for the identified individuals. To avoid these issues, we, the authors, annotate detections from `C4-en`. [4] First, we investigate documents with large amounts of PI by selecting the 100 documents with the most detected PI, which have between 999 and 6888 in-

[3] We were unable to develop meaningful regular expressions for two of the entities, U.S. BANK ACCOUNT NUMBER and NAME, given the complexity of recognizing them without returning a very high number of false positives. For credit cards, we found specific regular expressions for different companies (e.g., American Express, Visa, etc.), so we employed an ensemble of those to detect credit card numbers.

[4] We selected C4 for a manual audit given that it has gone through less filtering compared to the Pile, and has been used to train more models since its creation.

stances of PI each. We split the instances across all authors, with discussions in cases of uncertainty. We found that from these documents:

- **31%** are large dumps of cell phone numbers from different countries, containing the full name, phone number, and cellphone providers of users.
- Another **8%** of the documents are voter dumps from the US with the full name, address, and voter identification number of individuals in states such as Florida and Michigan.
- A further **8%** of the documents contain extensive lists of IP addresses and their corresponding company name, whereas another **5%** contain family trees or genealogies with name, birth year, and death year.
- Finally, **5%** contain a log of a bank accounts' transactions with amounts, although they do not have the name of the person, they do contain the bank account ID number.

The remaining documents contain lists of numbers (ISBN numbers, product ID numbers, polygon coordinates) that were falsely classified. This initial analysis indicates that over half of the documents that we manually verified did contain extensive amounts of truly sensitive, character-based personal information that can make links between individuals, their contact information, and information such as bank transactions and voting IDs.

There are also types of PI that were not explicitly searched for, but were encountered due to similarity with other types – for instance, patent numbers were found given their similarity to Social Security Numbers and GPS coordinates were flagged for their similarity in structure to US Bank numbers. We also found many highly questionable websites that were included in C4, ranging from a complete index of state-wide voter ID numbers (including full addresses and contact information) to a dump of US Social Security numbers of the deceased, also including their full names and locations. This particular kind of document is disquieting because if it is present in sufficient quantities in the data used for training language models, can then be generated given the right prompt (e.g. producing someone's SSN given their name), putting those individuals at risk, as per the work of Carlini et al. (2020). However, to take our validation further, we also carried out a manual analysis of a random sample of PI instances, to verify the accuracy of the two approaches that we used for our analysis.

4.3 Evaluating a Large Random Sample of PI Detections

While a small number of documents from C4 consisted of large dumps of personal information, the majority of the instances detected by our approaches were interspersed among the 364 million documents of the corpus. In fact, nearly all documents (approximately 98%) with PI have have 6 or fewer detected PI instances, and most documents contain just one type of PI. We therefore took a random sample of 200 instances of each type of PI detected in C4 for each of the two approaches and manually validated them to evaluate the performance of each approach.

Since there is no commonly agreed-upon system for evaluating PI detection, metrics for this task often re-purpose metrics from NER, e.g. partial or fully matching spans alongside the span type (Hathurusinghe et al., 2021). Other metrics that are used include variations on precision, recall, and F1 score (e.g., García-Pablos et al. (2020)) – however, for our evaluation, we cannot measure metrics that require true negatives, as that requires exhaustive PI ground truth annotations, which we lack. Thus, we focus on *precision* and introduce a second metric inspired by work on NER, PII detection, and computer vision: *detection accuracy*.

Our formulation of detection accuracy borrows from the evaluation of "object segmentation accuracy" in computer vision (Everingham et al., 2010), which measures per-pixel intersection-over-union (IOU), also known as Jaccard index, with respect to a ground truth. In our formulation, for every span of overlapping text between the ground truth (GT) and the detected personal information (DPI), we calculate IOU of the DPI *with respect to the detection* as a function of their word indices i:

$$\frac{\sum_i overlap(GT_i, DPI_i)}{(1 + max_i(GT_i, DPI_i) - min_i(GT_i, DPI_i))}$$

Where $overlap(GT_i, DPI_i)$ is an indicator function for a character within both the ground truth and the detected personal information spans. DPI spans without GT overlap receive a score of 0. Note that detection accuracy does not take label information into account at all.

As shown in Table 2, which reports the average of these scores across the selected instances, some PI types have high detection accuracy by both Presidio and regular expressions – this was the case for phone numbers and email addresses, which

Type	Presidio	Regex
PHONE NUMBER	94.4%	90.9%
EMAIL ADDRESS	99.0%	98.3%
US BANK NUMBER	30.0%	N/A
US SSN	46.3%	24.7%
IP ADDRESS	25.9%	49.8%
CREDIT CARD	13.1%	11.9%
IBAN CODE	98.5%	15.2%
NAME	52.3%	N/A

Table 2: Average detection accuracy for detected character-based personal information spans by Presidio and Regular Expression (Regex) approaches.

DPI Label	Presidio	Regex
PHONE NUMBER	95.5%	67.5%
EMAIL ADDRESS	99.0%	74.5%
US BANK NUMBER	0.5%	N/A
US SSN	0%	0%
IP ADDRESS	26.0%	38.5%
CREDIT CARD	2.5%	0.5%
IBAN CODE	98.5%	7.5%
NAME	52.0%	N/A

Table 3: Precision@.5 for detected character-based personal information spans by Presidio and Regular Expression (Regex) approaches.

both had detection accuracies in the 90s, with near-perfect accuracy for email addresses. These two types have very rigid syntax, naturally lending themselves to detection via rule-based methods like regular expressions. Other PI types, such as IBAN codes, were also well detected by Presidio (with 98.5% accuracy), but much less so via regular expressions (15.2%), which are more prone to false positives for this type because they do not include an IBAN checksum (see the large counts for regex IBANs in Table 1), which is used to separate IBAN codes from strings with similar patterns. False positives include ISBN numbers, hash values, and article id numbers. IP addresses had an opposite pattern, with regular expressions performing *better* than Presidio (49.8% versus 25.9%) with roughly comparable amounts detected. We found that Presidio often detects a single colon and labels it as an IP address, leading to many false positives. This simple error suggests there may be "low hanging fruit" to improving PI detection.

Results on label classifications are shown in Table 3 for precision at an average detection accuracy of 0.5. All thresholds for each of the fields produce similar results, even at a threshold of 1.0. This indicates that when a type of PI is correctly labeled, the predicted span tends to be correct. We find that Presidio is very precise at labeling phone numbers, email addresses, and IBAN codes, all with precision over 95%. The regular expressions did not have as high precision even in cases with high detection accuracy (phone numbers and email addresses). For IP addresses, regular expressions were more precise than Presidio (38.5% vs. 26.0%), similar to the high detection accuracy of this type discussed above. For US bank numbers, US social security numbers, and credit cards, neither method was particularly precise and often led to numerous false positives such as ISBN numbers, MLS numbers, article numbers, phone numbers, and miscellaneous manufacturing part numbers. In addition, US bank numbers, US social security numbers, and credit cards have detection accuracies that are much higher than their respective precisions because many of the detected results are other types of PI, such as phone numbers, leading to accurate spans, but incorrect labels for those spans.

4.4 Extrapolated Results

Based on the results of our manual verification, we can estimate the total quantity of each type of personal information present in C4 [5]. We can multiply our estimate of the proportion of true detections in Table 1 by the precision at .5 from our manual validation in Table 3 to arrive at an estimate of the total amount of personal information in C4. Using this method, we estimate C4 contains millions of phone numbers and IP addresses, according to both Presidio and regular expressions, as well as significant number of IP addresses (around half a million). This also estimates thousands of of IP addresses, credit card numbers, and IBAN codes. Our extrapolated results indicate that, even with limited methods that only cover a small subset of personal information, there are millions of examples of personal information openly available and non-anonymized in C4. We note that even though all manually checked detections of US Social Security Numbers were false positives, there likely exists some in the corpus. In addition, our tools may be ill-equipped to detect some instances.

While these numbers are estimates based on the detection counts and the accuracies that we calculated based on our random sample, they still indicate that there are significant quantities of personal information in C4 and the Pile, which are being

[5]We did not extrapolate for the Pile because we did not manually audit it, but we expect similar detection accuracies.

used to train LMs that are deployed in real-world settings ranging from customer service to predictive text generation. This opens the door to models parroting personal sensitive information such as credit card numbers, phone numbers, and email addresses without accounting for issues like privacy and consent. We discuss related endeavors in disciplines ranging from NLP to privacy and socio-technical studies in the next section.

4.5 Linked Instances of PI

Although character-based instances of PI are extreme-risk, when multiple instances of different types (e.g. US SSN, email address, and name) are close together, risk increases significantly. To analyze this, for each detected instance of PI by Presidio, we compute the number of other unique types of PI that are present within a 200 character window on both sides, including spaces. In C4, less than 2.7% of the detected instances were types other than a person's name. Despite that small percentage, almost 2.5% of the total instances had at least one other type of PI in its immediate vicinity, indicating much higher risk than originally thought. These instances were often a name coupled with another type. These trends are exacerbated for the Pile, where nearly 3.4% of the total instances were linked, compounding their risks.

This is particularly problematic because work by Latanya Sweeney, the founder of the Data Privacy Lab, used a combination of quasi-identifiers like gender, birth dates and postal codes to uniquely identify individuals, and concluded that the combination of all three is sufficient to identify 87% of individuals in the United States (Sweeney, 2000). This brings up the question of how this information can be used to identify a unique individual based on a single record with different types of PI.

5 Related Work

5.1 Creating and Documenting NLP Corpora

Before the advent of large language models (LLMs) requiring massive quantities of data, mindful curation was still possible for many linguistic corpora, which were manually collected using approaches involving adequate anonymization and consent, taking into account potential ethical issues (De Pauw, 2006) and respecting aspects such as copyright and autonomy (McEnery, 2019). Even though initial usages of the Common Crawl often involved some degree of manual curation and filtering (e.g. (Radford et al., 2019)), the amount of human intervention gradually tapered off in recent years, replaced by automatic filtering using approaches such as fuzzy deduplication (Brown et al., 2020) and perplexity scoring (Wenzek et al., 2019), despite their limited efficacy in filtering out problematic content such as hate speech and pornography (Luccioni and Viviano, 2021). Even despite these filtering techniques, Caswell et al. (2021) show that audits of numerous automatically crawled corpora are of very poor quality, with many corpora being completely erroneous and less than 50% of sentences being of acceptable quality.

The C4 corpus is actually one of the primary sources of training data for AI models, as well as one of the largest language datasets that currently exist, consisting of over 156 billion tokens collected from the Internet by Raffel et al. (Raffel et al., 2019) and used for training models such as T5 (Raffel et al., 2019) and the Switch Transformer (Fedus et al., 2021). A recent study by Dodge et al. found that a large portion of the domains represented in C4 comes from patent documentation and US military websites, as well as sources such as Wikipedia and newspapers, and that it contains machine-generated text, text from from benchmark NLP datasets, as well as a slew of demographic biases (2021). Other related work has also pursued other topics of analysis, either with the purpose of detecting undesirable context like hate speech and pornography (Luccioni and Viviano, 2021) or for filtering corpora (Wenzek et al., 2019). Given the sheer size of the web corpora and the frequency at which they are updated, in-depth analyses are challenging for researchers and practitioners alike, and there are many types of content of the corpus, such as personal information, that remain under-explored.

5.2 Detecting Personal Information

To date, the detection and removal of personal information has predominantly attracted attention in domains such as cybersecurity and privacy studies, and its presence has been detected in different parts of the Internet in the form of willful and accidental data dumps and records (Liu et al., 2020; Floyd et al., 2016; An, 2016). Despite the risks that the dissemination of PI entails (which we discuss in more detail in Section 2.2), its sharing on the Internet continues to grow, fueled by the increased usage of social media (Irshad and Soomro, 2018)

and the 'data market,' which gathers user data for targeted advertisements (Ullah et al., 2020).

In the field of ML in particular, the detection of personal information has not become a mainstream practice. Despite the abundance of data used to train LMs, the subject of PI detection has not received much attention compared to other tasks such as deduplication or filtering 'low-quality' data (Wenzek et al., 2019). While there has been work in detecting (and, eventually, removing) PI from training corpora, this has mostly been explored in contexts such as emails (Bier and Prior, 2014), health records (Murugadoss et al., 2021) and biographical information in Wikipedia (Hathurusinghe et al., 2021). Systematically detecting PI in written text remains an open question, given the diversity of types and source of PI that exist (as discussed previously in Section 2.1). Recent work has also proposed autoencoder-based approaches for transforming textual corpora while preserving privacy (Krishna et al., 2021), although the extent to which this works is still under debate (Habernal, 2021).

6 Future Work

Given the results of our case study on the C4 corpus and the Pile, we propose several recommendations for dataset creators and users that can help reduce the risks and harms due to the dissemination of personal information, whether it be via dataset sharing or model training.

Detecting and Removing Personal Information Both when creating new corpora and when using existing ones (such as the Common Crawl or C4), it is crucial to do due diligence surrounding PI. While there are limited tools that exist for culturally-specific personal information, programmatic approaches such as regular expressions can be viable, since they can be written given the specific types of information that is relevant to a given context (e.g. addresses or phone numbers from a specific country or region). Running an out-of-the-box tool such as Presidio is the bare minimum that should be run on all new and existing corpora; manually labelling a small sample of documents, such as we did in the current study, can be a valuable complement to that approach. Replacing names by `<NAME>` and credit card numbers by `<CREDIT CARD>` can be used as a fail safe when PI is detected in corpora. Initiatives such as the Workshop on Private NLP (Feyisetan et al., 2020) are working towards this goal, pursuing the creation of privacy-preserving datasets.

Practicing Consent Scraping data automatically from the Web can be tempting given the amount of information available online; however, it often sidesteps the issue of consent and 'opting in' to sharing ones information (which we discussed in Section 2.2). Collecting datasets in a way that is more respectful of individuals' rights is a direction that our field should be moving in, and we hope that future corpora collection efforts will offer individuals the option to 'opt in', rather than assuming that they do so by default. Including data providers and data owners in the collection process grants them more agency in the process and helps ensure that goals and expectations are maximally aligned.

Developing Tools for Detecting Personal Information Memorization A complementary approach to reducing exposure in already trained models is testing them for the existence of PI. There are no existing approaches that can do this systematically, but there are some tools that can be of use — for instance, Carlini et al (2020) share their code for extracting memorized training data from GPT-2 (Radford et al., 2019), which can be modified for other models and data sources. However, running this code necessitates a set of prompts (i.e. personal information) that first need to be gathered from the training corpus itself. Developing better approaches (i.e. unit tests) to detect memorization, more specifically PI memorization in trained models is vital, since deployed ML models in sensitive contexts (e.g. finance and healthcare) can divulge sensitive information and expose individuals and communities to potential harm.

We hope that the approach to defining and structuring PI, as well as the case study, described in the current article present a compelling case to our community that the topic of personal information is under-explored (and its impact is under-estimated). Our goal is to start a conversation and spur action around this important topic, and to contribute to developing tools and approaches, both ML-centric and rule-based, to detect and remove PI in both models and datasets. We believe that this will be useful for communities above and beyond our own, spanning from legal studies to socio-technical ones, who can benefit from such tools in their own initiatives to improve the state of privacy preservation on the Internet and beyond.

Limitations

To our knowledge, the current study is the first effort in the ML community aiming to define PI and estimate how much of it is present in two major training corpora, C4 and the Pile. However, we recognize that there are ways in which our study can be improved, and directions in which future studies can be conducted. To start with, when annotating the PI found both by using Presidio and regular expressions, we observed that new forms of PI have appeared with the advent of the Internet, but have yet to be considered in traditional definitions (e.g. Facebook events URLs), despite their potential for risk. Also, given the diversity of types of PI that exist, it is unsurprising that systematically detecting them remains a challenge. As we reported Section 4, we found that both Presidio and regular expressions were able to detect certain types of PI, such as emails and phone numbers, relatively well, but failed on other types, such as SSNs and credit cards; however, without access to ground truth annotations, measuring and characterizing false negatives is impossible.

Other limitations of both types of approaches is that they are language- and often country-specific, and need to be adapted to contexts of application and languages. This can quickly become complex, because the format of common types of PI such as bank account numbers varies immensely depending on its country of provenance. Finally, linguistic characteristics of individual languages make it difficult for multi-lingual PI detection since features that are relevant towards PI detection in some languages are not relevant for others; more work on developing more modular and extensive PI detection tools would be an important contribution to many communities and endeavors, and it is conceivable that ML-based approaches can contribute to these efforts.

Broader Impact Statement

Our work endeavors to help the NLP community better understand and quantify the types and quantity of personal information contained in popular training corpora. In order to strive towards this goal, we manually annotated a subset of the personal information detected in C4, which constitutes a dataset that could be valuable to the community. However, given the quantity of high-risk personal information that this sample contains, we do not feel comfortable disseminating it. We are, however, working on methods for developing synthetic and lower-risk labeled corpora to help develop better methods for detecting PI. As large language model development is increasing dramatically, more models will be trained on these data sources, so its becoming increasingly important to quantify and characterize the personal information present in datasets as well as help practitioners develop better PI detection methods.

References

Johanna An. 2016. A responsible de-identification of the real data corpus: building a framework for pii management. Technical report, Naval Postgraduate School Monterey United States.

Tuomas Aura, Thomas A Kuhn, and Michael Roe. 2006. Scanning electronic documents for personally identifiable information. In *Proceedings of the 5th ACM workshop on Privacy in electronic society*, pages 41–50.

Marianne Bertrand and Sendhil Mullainathan. 2004. Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review*, 94(4):991–1013.

Christoph Bier and Jonas Prior. 2014. Detection and labeling of personal identifiable information in e-mails. In *IFIP International Information Security Conference*, pages 351–358. Springer.

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, et al. 2022. Gpt-neox-20b: An open-source autoregressive language model. *arXiv preprint arXiv:2204.06745*.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*, pages 267–284.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2020. Extracting training data from large language models. *arXiv preprint arXiv:2012.07805*.

Isaac Caswell, Julia Kreutzer, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Auguste Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios Gonzales, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Rubungo Andre Niyongabo, Toan Q. Nguyen, Mathias Muller, Andr'e Muller, Shamsuddeen Hassan Muhammad, Nanda Firdausi Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, M. Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine cCabuk Balli, Stella Rose Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi N. Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2021. Quality at a glance: An audit of web-crawled multilingual datasets. *ArXiv*, abs/2103.12028.

Commonwealth Consolidated Acts. 1988. Privacy act 1988 - section 6 interpretation. [Online; accessed 15-January-2022].

Creemers, Rogier and Webster, Graham. 2021. Translation: Personal information protection law of the people's republic of china – effective nov. 1, 2021. [Online; accessed 15-January-2022].

Guy De Pauw. 2006. *Developing Linguistic Corpora—A Guide to Good Practice Martin Wynne (ed.).* EADH: The European Association for Digital Humanities.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305.

Dtaa Protection Act. 2018. United kingdom general data protection regulation. [Online; accessed 16-January-2022].

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338.

Executive Office of the President. 2006. Safeguarding personally identifiable information. [Online; accessed 15-January-2022].

William Fedus, Barret Zoph, and Noam Shazeer. 2021. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity.

Oluwaseyi Feyisetan, Sepideh Ghanavati, and Patricia Thaine. 2020. Workshop on privacy in NLP (PrivateNLP 2020). In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 903–904.

Travis Floyd, Matthew Grieco, and Edna F Reid. 2016. Mining hospital data breach records: Cyber threats to us hospitals. In *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, pages 43–48. IEEE.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Aitor García-Pablos, Naiara Perez, and Montse Cuadros. 2020. Sensitive data detection and classification in Spanish clinical text: Experiments with BERT. *arXiv preprint arXiv:2003.03106*.

Ivan Habernal. 2021. When differential privacy meets nlp: The devil is in the detail. *arXiv preprint arXiv:2109.03175*.

Rajitha Hathurusinghe, Isar Nejadgholi, and Miodrag Bolic. 2021. A privacy-preserving approach to extraction of personal information through automatic annotation and federated learning. In *Proceedings of the Third Workshop on Privacy in Natural Language Processing*, pages 36–45, Online. Association for Computational Linguistics.

Michael A Hogg. 2020. *Social identity theory*. Stanford University Press.

Shareen Irshad and Tariq Rahim Soomro. 2018. Identity theft and social media. *International Journal of Computer Science and Network Security*, 18(1):43–55.

Sonia K Kang, Katherine A DeCelles, András Tilcsik, and Sora Jun. 2016. Whitened résumés: Race and self-presentation in the labor market. *Administrative Science Quarterly*, 61(3):469–502.

Satyapriya Krishna, Rahul Gupta, and Christophe Dupuy. 2021. Adept: Auto-encoder based differentially private text transformation. *arXiv preprint arXiv:2102.01502*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *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)*, abs/2101.00190.

Yizhi Liu, Fang Yu Lin, Zara Ahmad-Post, Mohammadreza Ebrahimi, Ning Zhang, James Lee Hu, Jingyu Xin, Weifeng Li, and Hsinchun Chen. 2020. Identifying, collecting, and monitoring personally identifiable information: From the dark web to the surface web. In *2020 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 1–6. IEEE.

Alexandra Sasha Luccioni and Joseph D Viviano. 2021. What's in the box? an analysis of undesirable content in the common crawl corpus. *arXiv preprint arXiv:2105.02732*.

Tony McEnery. 2019. *Corpus linguistics*. Edinburgh University Press.

Microsoft. 2021. Presidio - data protection and anonymization api. [Release Version 2.2.23, released on Nov 16, 2021].

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

National Congress of Brazil. 2020. Brazilian general data protection law (lgpd), english translation. [Online; accessed 18-January-2022].

Office of the Secretary of Defense. 2007. Memorandum for the office of management and budget. subject: Personally identifiable information. [Online; accessed 15-January-2022].

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

South African Parliament. 2013. Protection of personal information act (popi act). [Online; accessed 15-January-2022].

Nishant Subramani, Samuel R. Bowman, and Kyunghyun Cho. 2019. Can unconditional language models recover arbitrary sentences? In *NeurIPS*.

Nishant Subramani and Nivedita Suresh. 2020. Discovering useful sentence representations from large pretrained language models. *ArXiv*, abs/2008.09049.

Nishant Subramani, Nivedita Suresh, and Matthew Peters. 2022. Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 566–581, Dublin, Ireland. Association for Computational Linguistics.

GDPR Summary. 2020. Gdpr summary. [Online; accessed 18-January-2022].

Latanya Sweeney. 2000. Simple demographics often identify people uniquely. *Health (San Francisco)*, 671(2000):1–34.

John C Turner. 2010. Towards a cognitive redefinition of the social group. In *Research Colloquium on Social Identity of the European Laboratory of Social Psychology, Dec, 1978, Université de Haute Bretagne, Rennes, France; This chapter is a revised version of a paper first presented at the aforementioned colloquium.* Psychology Press.

Imdad Ullah, Roksana Boreli, and Salil S Kanhere. 2020. Privacy in targeted advertising: A survey. *arXiv preprint arXiv:2009.06861*.

UN High-Level Committee on Management (HLCM). 2018. Personal data protection and privacy principles. [Online; accessed 3-December-2021].

United States Department of Defense. 2007. Department of defense privacy program. [Online; accessed 12-January-2022].

Michael Volske, Martin Potthast, Shahbaz Syed, and Benno Stein. 2017. TL;DR: Mining Reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 59–63, Copenhagen, Denmark. Association for Computational Linguistics.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2019. Ccnet: Extracting high quality monolingual datasets from web crawl data. *arXiv preprint arXiv:1911.00359*.

Wikipedia contributors. 2021. Luhn algorithm — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Luhn_algorithm&oldid=1058129193. [Online; accessed 14-January-2022].

Supplementary Materials

Regular Expressions

Here are the regular expressions we used to find personal information in C4 and Pile.

```
IP address:
{"(25[0-5]|2[0-4][0-9]|1[0-9][0-9]|[1-9]?[0-9])\.
(25[0-5]|2[0-4][0-9]|1[0-9][0-9]|[1-9]?[0-9])\.
(25[0-5]|2[0-4][0-9]|1[0-9][0-9]|[1-9]?[0-9])\.
(25[0-5]|2[0-4][0-9]|1[0-9][0-9]|[1-9]?[0-9])"}

IBAN code:
{"[a-zA-Z]{2}[0-9]{2}[a-zA-Z0-9]{4}[0-9]{7}([a-zA-Z0-9]?){0,16}"}

US SSN:
"(?!000|.+0{4})(?:\d{9}|\d{3}-\d{2}-\d{4})"

email addresses: "(?:[a-z0-9!#$%&'*+/=?^_`{|}~-]+(?:\.[a-z0-9!#$%&'*+/=?^_`{|}~-]+)*|\"
                 (?:[\x01-\x08\x0b\x0c\x0e-\x1f\x21\x23-\x5b\x5d-\x7f]|\\
               [\x01-\x09\x0b\x0c\x0e-\x7f])*\")@(?:(?:[a-z0-9](?:[a-z0-9-]*[a-z0-9])
                 ?\.)+[a-z0-9](?:[a-z0-9-]*[a-z0-9])?|\[(?:(?:
            (2(5[0-5]|[0-4][0-9])|1[0-9][0-9]|[1-9]?[0-9]))\.){3}(?:(2(5[0-5]|[0-4][0-9])
                 |1[0-9][0-9]|[1-9]?[0-9])|[a-z0-9-]*[a-z0-9]:
                 (?:[\x01-\x08\x0b\x0c\x0e-\x1f\x21-\x5a\x53-\x7f]|\\
                 [\x01-\x09\x0b\x0c\x0e-\x7f])+)\])"

phone numbers: "\s*\(?(\d{3})\)?[-\. ]*(\d{3})[-. ]?(\d{4})"

amex_card: "3[47][0-9]{13}"
bcglobal: "(6541|6556)[0-9]{12}"
carte_blanche card: "389[0-9]{11}"
diners_club_card: "3(?:0[0-5]|[68][0-9])[0-9]{11}"
discover_card: "65[4-9][0-9]{13}|64[4-9][0-9]{13}|6011[0-9]{12}
                 |(622(?:12[6-9]|1[3-9][0-9]|[2-8][0-9][0-9]
                 |9[01][0-9]|92[0-5])[0-9]{10})"
insta_payment_card: "63[7-9][0-9]{13}"
jcb_card: "(?:2131|1800|35\d{3})\d{11}"
"korean_local_card": "9[0-9]{15}"
"laser_card": "(6304|6706|6709|6771)[0-9]{12,15}"
"maestro_card": "(5018|5020|5038|6304|6759|6761|6763)[0-9]{8,15}"
"mastercard": "(5[1-5][0-9]{14}|2(22[1-9][0-9]{12}|2[3-9][0-9]{13}|[3-6][0-9]{14}
|7[0-1][0-9]{13}|720[0-9]{12}))"
"solo_card": "(6334|6767)[0-9]{12}|(6334|6767)[0-9]{14}|(6334|6767)[0-9]{15}"
"switch_card": "(4903|4905|4911|4936|6333|6759)[0-9]{12}|(4903|4905|4911|4936|6333|6759)
[0-9]{14}|(4903|4905|4911|4936|6333|6759)[0-9]{15}|564182[0-9]{10}|564182[0-9]{12}
|564182[0-9]{13}|633110[0-9]{10}|633110[0-9]{12}|633110[0-9]{13}"
"union_pay_card": "(62[0-9]{14,17})"
"visa_card": "4[0-9]{12}(?:[0-9]{3})?"
```