Automated Anonymization of Parole Hearing Transcripts

Abed El Rahman Itani
University of Passau
itani01@ads.uni-passau.de

Wassiliki Siskou
University of Konstanz
University of Passau
wassiliki.siskou@uni-konstanz.de

Annette Hautli-Janisz
University of Passau
annette.hautli-janisz@uni-passau.de

Abstract

Responsible natural language processing is more and more concerned with preventing the violation of personal rights that language technology can entail (Weidinger et al., 2022). In this paper we illustrate the case of parole hearings in California, the verbatim transcripts of which are made available to the general public upon a request sent to the California Board of Parole Hearings. The parole hearing setting is highly sensitive: inmates face a board of legal representatives who discuss highly personal matters not only about the inmates themselves but also about victims and their relatives, such as spouses and children. Participants have no choice in contributing to the data collection process, since the disclosure of the transcripts is mandated by law. As researchers who are interested in understanding and modeling the communication in these hierarchy-driven settings, we face an ethical dilemma: publishing raw data as is for the community would compromise the privacy of all individuals affected, but manually cleaning the data requires a substantive effort. In this paper we present an automated anonymization process which reliably removes and pseudonymizes sensitive data in verbatim transcripts, while at the same time preserving the structure and content of the data. Our results show that the process exhibits little to no leakage of sensitive information when applied to more than 300 hearing transcripts.

1 Introduction

The growing need for anonymized datasets in computational social science such as NLP applications in law, criminology, sociology and political science is driven by the importance of ethical compliance, legal requirements, reduction of bias, and, ultimately, by the necessity for data sharing. In the context of spoken and transcribed dialogue data, anonymized datasets are particularly scarce. This holds especially for dialogues in legal settings, such as parole suitability hearings, where inmates who were originally sentenced to life-long imprisonment engage in discussions with a board of commissioners, requesting to be released from prison before the completion of their sentence. In California, the verbatim transcripts of these hearings can be requested at the California Board of Parole Hearings, but sharing them publicly raises ethical concerns: they include personally identifiable information (PII), such as names, inmate IDs, dates and other sensitive details about the people involved, and the participants do not have a choice as to whether they want to take part in the data collection process. Researchers who wish to make these transcripts available to ensure academic transparency face ethical dilemmas, as publishing the data would compromise the privacy of those affected. The contribution of our work is twofold: First, we introduce a robust automatic anonymization process for dialogue transcripts in criminal law, ensuring consistent entity replacement throughout each transcript. Second, we provide an evaluation of our process based on a subset of 100 manually anonymized parole hearing transcripts and show our system's minimal risk of data leakage demonstrated by the systems high precision.

2 Related Work

While anonymization has mainly been applied to data in the legal and clinical domain, researchers from other disciplines also feel the need to protect sensitive information in their datasets. In the specific case of Californian parole hearings, the dataset has not been made available to the public in its entirety (Hong et al., 2021b), with the available data only restricted to individual examples (Todd et al., 2020; Hong et al., 2021a). In a similar case for German, the entire dataset was manually redacted (Espinoza et al., 2024) and double-checked by a second person to ensure correctness.

Proceedings of the Natural Legal Language Processing Workshop 2024, pages 115–128

Regarding automatic anonymization, previous work mainly relies on written legal and clinical documents, applying methods such as masking, falsification or pseudonymization. ANOPPI uses a combination of automatic and semi-automatic processes, utilizing statistics and rule-based Named Entity Recognition (NER) methods to identify and remove personal information in Finnish court documents (Oksanen et al., 2019). It consistently replaces sensitive data with categorical labels, preserving both the semantic meaning and readability of the documents. PSILENCE uses a combination of NER tools and Coreference Resolution to ensure consistent labelling of entities in written legal documents (Cabrera-Diego and Gheewala, 2024). Schamberger (2021) proposes a customization solution to anonymize German legal court rulings using domain-specific NER in order to mask entities according to predefined rules. In the clinical domain, Ribeiro et al. (2023) proposes INCOGNITUS to automatically anonymize clinical notes by using a combination of NER tools like Conditional Random Field. For the anonymization of spoken language, Gardiner et al. (2024) use an Automatic Speech Recognition (ASR) system to generate transcripts from phone and video conversations and then enhance Google Data Loss Prevention service to improve PII detection.

The data underlying this paper falls between audio data (unpredictable and unstructured) and written legal documents (structured and manually curated): parole hearing transcripts exhibit some inherent structure with clearly identified speakers and roles, while dialogue is characterized by repair sentences, filled pauses, elliptical content as well as cut-off and spelled-out names with PII. While the aforementioned work on anonymization of written language does not directly translate to verbatim transcripts, the approach we take in the paper is similar in that we combine Named Entity Recognition and regular expressions with rule-based post-processing to identify sensitive information in the transcripts and replace it with categorical tags[1]. In this way we adjust the approach by INCOGNITUS (Oksanen et al., 2019) and apply it to dialogues from criminal law.

3 Data

The dataset comprises 334 parole suitability hearing transcripts in PDF format, which we officially requested from the California Department of Corrections and Rehabilitation (CDCR)[2]. All hearings took place between August and September 2021. The corpus consists of 21,874 pages and 5,013,156 words in total, with an average of 15,009 words and 65 pages per file. The transcripts share a standardized format, which we can exploit for anonymization: The first page of the PDF contains the names of the participants present in the hearing, the location such as the prison or facility of the inmate, as well as the date and time of the hearing. The second page contains the index, indicating the page number of each section (such as pre-commitment factors, post-commitment factors and decision) of the hearing. The subsequent pages comprise the main body of the transcript, containing the verbatim transcription of the dialogue, uniformly formatted with the speaker tag and the according text. The document also includes the closing statements of all participants and the final decision. The last pages of the document are reserved for a declaration of the transcriber as well as their signature. For anonymization, all transcripts are converted to text files. The index of each transcript is ignored during conversion as it does not contain PII. The final text file is formatted so that each utterance appears on a separate line to facilitate reading and processing.

4 Automatic Identification of Sensitive Information

4.1 Categories of Sensitive Information

Pre-defining categories of sensitive information is a key prerequisite for effective automatic data anonymization. The identification of privacy-relevant data categories involved an iterative approach to capture the full spectrum of sensitive information in parole hearing transcripts. This process entailed multiple rounds of manual review, with each iteration refining the list of categories to be anonymized.

Given the standardized format of the hearing

[1]We decided against masking as an anonymization technique, as the sensitive information would have only been redacted (e.g. "****") but not replaced with context-sensitive tags, thereby reducing the semantic and pragmatic expressiveness of the transcript. Additionally, we ruled out falsification, which replaces real data with generated false data, as it carries the risk of generating real names of individuals, potentially causing unintended consequences when publishing the falsified dataset.

[2]Parole hearing transcripts can be requested under the California Public Record Act. https://www.cdcr.ca.gov/bph/psh-transcript/

transcripts, most of the relevant data can be found in the front page of the documents. These include mainly the names of each person present in the hearing, the inmate's identification number (CDCR ID from now), the location, as well as time and date. Other forms of parole hearing specific sensitive data include spelled names, as well as fractions of spelled names (such as "V as in Victor"). We define those as *direct* identifiers and consider those to be the most important information to remove, as they significantly increase the risk of re-identification. Consequently, their removal is a priority.

The above mentioned comprehensive examination of the transcripts yields a list of *indirect* identifiers such as company names, organizations, age, height, nationality, religion, political group, phone numbers, URLs and email addresses that need to be redacted in order to ensure coverage of all privacy-relevant information of the individuals involved. Table 1 provides descriptions of all direct and indirect identifiers.

4.2 Automatic Entity Labeling

We employ a multi-tool approach for Named Entity Recognition to detect the different direct and indirect identifiers in the transcripts. We mainly use Presidio[3](Mendels and Balter, 2020), an open-source tool by Microsoft, to identify most PII in the text as it allows to manually add custom recognizers based on Regular Expressions. We developed a set of regular expressions tailored to detect entities unique to parole hearings, including SPELLED_NAME, SPELLED_OUT_ITEM and CDCR_ID. We additionally use spaCy[4] (Honnibal et al., 2020) and StanfordNER[5] (Finkel et al., 2005) to cover entities like PERSON, LOCATION, ORGANIZATION, as well as TIME and DATE. Although the latter two NERs already search for any occurrence of these entity types, we implemented custom recognizers in Presidio for these categories too, to complement the detection process, therefore reducing the risk of data leakage. Table 1 specifies which tool or combination of tools are used for each entity type.

By using multiple NER tools simultaneously on a single transcript, we leverage the strengths of each tool. While there is overlap in detecting common types such as PERSON and LOCATION, the tools complement each other by expanding coverage, thereby improving the overall accuracy of the anonymization process. In case one tool misses a piece of sensitive information, the likelihood of another tool detecting it gets increased, resulting in a more reliable entity identification process.

4.3 Automatic Filtering

A thorough manual iteration through the generated results shows some common errors in the labeling, e.g., context-specific non-sensitive terms of parole hearings that are needed in the transcript for information preservation such as "Board of Parole Hearings" is frequently misidentified as an ORGANIZATION entity. This process yields a whitelist of 290 named entities to prevent over-anonymization.

The main goal of our cleaning process however is the elimination of duplicate and overlapping labels generated by different NER tools. We define duplicates as those with identical start and end index positions. Formally, given two labeled entities $A_1(s_1, e_1)$ and $A_2(s_2, e_2)$, where s and e represent the start and end positions, we remove A_2 if:

$$s_1 = s_2 \text{ and } e_1 = e_2$$

For overlapping entities, we apply a series of rules for filtering. Labels that share the same start point but differ in end points, we keep the longer label. Formally, given $A_1(s_1, e_1)$ and $A_2(s_1, e_2)$, we keep A_1 if:

$$s_1 = s_2 \text{ and } e_1 > e_2$$

Similarly, for labels with the same endpoint but different start points, we preserve the longer annotation. Given $A_1(s_1, e_1)$ and $A_2(s_2, e_1)$, we keep A_1 if:

$$e_1 = e_2 \text{ and } s_1 < s_2$$

In cases where entities overlap, but do not share start and end index, we adjust their boundaries to create distinct, non-overlapping labels. For $A_1(s_1, e_1)$ and $A_2(s_2, e_2)$, where:

$$e_1 > s_2$$

we modify e_1 to ensure that it precedes s_2:

$$e_1 = s_2 - 1$$

[3] https://github.com/microsoft/presidio

[4] https://spacy.io/models

[5] https://nlp.stanford.edu/software/CRF-NER.html

Entity Type	Description	NER Source
PERSON	Names of individuals involved in the hearings are listed on the first page of the transcript. PERSON entities often include prefixes such as titles (e.g., Doctor, Miss, Commissioner). We remove these titles from the annotations to ensure that only the names themselves are removed in the final transcript.	Presidio, spaCy, StanfordNER
SPELLED_NAME	A custom label from a custom recognizer in Presidio is used to handle cases where names in parole hearings are spelled with letters separated by dashes (e.g., "J-O-H-N")..	Presidio
SPELLED_OUT_ITEM	A custom label from a custom recognizer in Presidio is used to handle cases where names are phonetically spelled out (e.g., "V as in Victor").	Presidio
CDCR_ID	A custom label added in Presidio to detect inmates CDCR ID, typically starting with a letter and followed by a series of numbers (e.g., "V12345").	Presidio
LOCATION	An umbrella term for locations, including states, countries, cities, etc.	Presidio, spaCy, StanfordNER
ORGANIZATION	Includes company names and organizations.	Presidio, StanfordNER
DATE	SpaCy's DATE entity detects dates, durations, ages, and time under a single category. To isolate actual dates, the duration, age, and time data are filtered out and reassigned to their specific entities. Additionally, a custom Presidio recognizer is employed to enhance the detection of typical date formats by assigning them to the DATE entity.	Presidio, spaCy
TIME	This combines a custom Presidio recognizer with spaCy's DATE entity to identify time patterns (such as XX:XX) and label them as TIME entities.	Presidio, spaCy
AGE	Identified using a combination of a custom Presidio recognizer as well as the age and duration data extracted from spaCy's DATE entity.	Presidio, spaCy
HEIGHT	Custom label detected by a custom Presidio recognizer. It detects numbers followed by height units (e.g., feet, inches).	Presidio
NRP	Presidio entity representing Nationality, Religion, or Political group.	Presidio
PHONE_NUMBER	Covers telephone numbers.	Presidio
EMAIL_ADDRESS	Covers email addresses.	Presidio
URL	Covers web addresses.	Presidio

Table 1: Entity types detected during the automatic annotation and anonymization process, alongside a small explanation of each type and the corresponding NER source.

This approach ensures that each word in the text is associated with at most one entity to prevent ambiguities in the anonymization process. A brief example of the filtering process is presented in Appendix A (step 1 to 3).

A total of 573,024 entities were labeled by the NER tools and regular expressions across the entire dataset. Table 2 displays the counts of labels that were filtered out. By applying the whitelist, handling duplicates, and resolving overlapping labels, 372,714 annotations were removed, with PERSON entities accounting for the highest number of removed annotations (306,567). As personal names are the most prevalent in parole hearing transcripts and all three NER tools are tasked with identifying them, this high degree of overlap is expected. Entities such as SPELLED_NAME and CDCR_ID are identified using regular expressions. Due to the transcription guidelines, the transcripts often include instances of stuttering (e.g., "I-I-I"), where repeated letters mimic the format of spelled names (e.g., "J-O-H-N"). This causes the regular expressions to incorrectly label stuttering as spelled names. To avoid incorrect labeling, we check if the repeated letters are identical, and if so, the label is getting removed. As a result, 418 SPELLED_NAMES labels were removed. The removal of certain CDCR_ID labels is due to overlaps where short IDs were detected as part of a larger CDCR_ID. In such cases, the filtering process merges the overlapping IDs into one and discards the redundant labels. Additionally, there are instances where commissioners begin spelling an ID but need to correct themselves partway through, leading to duplicate labels. As a result, 4 CDCR_ID labels needed to be removed.

The remaining 200,310 identified entities are clean, unique and usable annotations.

Annotation Type	Count
PERSON	306,567
LOCATION	14,787
ORGANIZATION	5,072
SPELLED_NAME	418
CDCR_ID	4
DATE	24,918
TIME	4,326
AGE	1,735
URL	4
NRP	360
Numerical values (non-sensitive)	14,523
Total annotations filtered-out	**372,714**
Final correct annotations	**200,310**

Table 2: Filtered annotation counts by entity and results of the automatic filtering process.

4.4 Results

We evaluate the performance of the automatic entity labelling based on a manually annotated subcorpus of 100 parole hearing transcripts. While the results presented in Table 3 primarily reflect the accuracy of the identification of sensitive entities, they directly impact the effectiveness of the anonymization process, as the replacements of each entity are based on these results. The table presents the precision, recall and F1-score for each entity type, based on a gold standard created by one of the authors. The results show a generally strong performance, with several entity types achieving high scores.

Entity Type	Precision	Recall	F1-score
PERSON	0.981	0.989	0.985
LOCATION	0.846	0.946	0.893
ORGANIZATION	0.768	0.739	0.754
SPELLED_NAME	1.000	0.995	0.997
CDCR_ID	0.933	0.996	0.964
DATE	0.883	0.968	0.923
TIME	0.977	0.943	0.960
AGE	0.903	0.926	0.914
HEIGHT	1.000	0.800	0.889
EMAIL_ADDRESS	1.000	0.750	0.857
URL	0.667	1.000	0.800
NRP	0.765	0.830	0.796
SPELLED_OUT_ITEM	1.000	1.000	1.000
PHONE_NUMBER	1.000	1.000	1.000
OVERALL	**0.955**	**0.972**	**0.963**

Table 3: Precision, Recall and F1-score of spaCy, Presidio and StanfordNER combined for different entity types across 100 hearing transcripts.

As stated in §4.1, we prioritize the detection of direct identifiers, which pose a higher risk of re-identification, such as PERSON, SPELLED_NAME, SPELLED_OUT_ITEM, CDCR_ID, LOCATION, TIME and DATE. For all of these entities our entity labeling approach achieves high or very high F1-scores. Notably, we observe an impressive F1-score of 0.985 for the PERSON entity type, showing the effectiveness of our multi-tool approach in accurately detecting individual names. Similarly, entities such as SPELLED_NAME, SPELLED_OUT_ITEM and PHONE_NUMBER achieve perfect or near-perfect scores (F1 $\geq$ 0.997). These high scores in performance can be attributed to the use of custom recognizers and regular expressions, which are particularly suited for the consistent structure and formatting of these entity types.

It is important to note that entities such as EMAIL_ADDRESS, URL and PHONE_NUMBER are quite rare, with HEIGHT and SPELLED_OUT_ITEMS, being the only ones occurring more than 20 times in the 100 transcripts analyzed. Additionally, the transcription conventions of parole hearings ensure a standardized format for these entities, which makes their detection through regular expressions straightforward. CDCR_ID and TIME entities both achieved F1-scores above 0.95. Among the direct identifiers entities related to LOCATION (F1 = 0.893) and DATE (F1 = 0.923) are the ones that show moderate performance but leave room for improvement.

The ORGANIZATION entity type posed significant challenges, resulting in the lowest F1-score of 0.754. We attribute this underperformance to the excessive use of abbreviations for programs, procedures and acts within the hearings, which are often misclassified by the NERs as organizations. For example, the abbreviation ”CBA” for ”Criminal Behavior Assessment” is incorrectly labelled as an organization, leading to confusing anonymization results in the end. The URL entity is amongst the rarest entities in the transcripts. It has a precision of 0.667 and an F1-score of 0.8 due to a single case of false positive, thus reaching to the conclusion that the automatic annotation process struggles with detecting URL entities.

Although some indirect identifiers exhibit lower F1-scores, this does not undermine the effectiveness of our approach. The high precision in recognizing direct identifiers significantly mitigates the risk of re-identification, ensuring the protection of privacy even if indirect identifiers are not perfectly detected.

Table 5 in Appendix B presents the performance of each NER tool when run individually on the subcorpus of 100 transcripts, with Presidio serving as baseline for comparison. Notably. Presidio achieves an F1-score of 0.989 for the PERSON entity, thus outperforming both StanfordNER and spaCy in this specific category. However, Presidio's performance falls short in other categories, with spaCy demonstrating better results for temporal data such as DATE, TIME and AGE. Presidio relies on regular expressions for such entities and therefore only serves to boost spaCy's results when used in combination. The F1-scores for entities such as LOCATION and ORGANIZATION are also lower when the tools are used by themselves. For instance, StanfordNER and Presidio show a low performance in the ORGANIZATION entity with F1-scores less than 0.600. In contrast, Table 3 demonstrates higher F1-scores when all tools are combined by improving the F1-score by 0.154 for the ORGANIZATION category.

This comparison underscores the complementary nature of the multi-tool approach, where the combination of tools compensates for weaknesses of the individual tools. Overall, an F1-score of 0.963 is achieved across all entities with the multi-tool approach. The values reported show a strong performance across most entity types, indicating that, once anonymized, the final transcripts will effectively protect individual's privacy and make re-identification difficult.

Our results are comparable to those reported in previous work by Schamberger (2021) and their domain-specific NER models achieving an F1-score between 0.802 and 0.811 for the identification of personal names. Our process achieves a higher F1-score for PERSON entities (0.985), indicating improved handling of names within legal settings, though it performs lower in detecting LOCATION and ORGANIZATION data, with F1-scores of 0.893 and 0.754, respectively. We attribute the higher performance for PERSON entities to the document format, where the names of participants are listed on the cover page of each transcript. In contrast, the lower scores for LOCATION and ORGANIZATION are likely due to domain-specific abbreviations, which lead to misclassifications (see also §5.1).

We conducted an ablation study, to evaluate the impact of the information given by the first page of the transcripts. We therefore executed the automatic annotation process without incorporating the names of the participants, the inmate's name, as well as the time and date of the hearing, typically found on the first page. Table 7 (Appendix D) shows the results of this experiment. A standard run (i.e. with first page information included) of the anonymization process yields a total of 573,024 unfiltered annotations, while running the code without the incorporation of the first page information only resulted in 447,216 unfiltered annotations. Integrating the names and organizations from the first page into Presidio improves it's accuracy of name detection, a step that particularly proves valuable in identifying names that were spelled or appear in incomplete form in the transcript.

The impact of incorporating the first page information into Presidio is evident in Table 8 in Appendix D. This table illustrates the differences in F1-scores for PERSON, LOCATION and ORGANIZATION entities. Without the first page information, these scores were 0.944, 0.713, and 0.630 respectively. These values are lower compared to the scores achieved when the anonymization process includes the first page data.

During a standard run, there is less confusion between the entities, resulting in more accurate annotations.

5 Pseudonymization

Pseudonymization involves assigning unique labels to each distinct entity within the dataset for the purpose of anonymization, meaning that reappearing entities are consistently replaced by the same tag. This is done through the use of a dictionary, which stores the original entity along with their pseudonymized category label. The primary function of this dictionary is to ensure the correct tag is consistently assigned throughout the transcript to the specific entities. In practice, entities are anonymized by combining the entity type with a sequential number. For example, names within the transcript are replaced by tags like [PERSON_1], [PERSON_2] and so forth. This approach is applied not only to names but to all recurring entities, ensuring consistent labeling across the dataset. Appendix A (step 4) showcases a practical illustration of how the anonymized final transcript appears after the replacement of PII by categorical tags.

For PERSON entities, all full names are extracted and each part of the full name is assigned

a unique tag, which is then stored in a dictionary to ensure consistency across the transcript. We decided to use this approach, as commissioners often refer to the inmates by their last name alone. Each part of a PERSON entity is labeled with a sequential number, and the corresponding SPELLED_NAME is normalized and assigned the same sequential number as the matching PERSON entity. For CDCR ID entities, each unique value is assigned an individual tag. Example (1) demonstrates how these direct identifiers (in 1a) are anonymized within the transcripts (in 1b):

(1) a. **Original:**
We have a John Doe and the victim is Jane Smith. That's D-O-E. Case ID M23515.

b. **Anonymized:**
We have a [PERSON_1] [PERSON_2] and the victim is [PERSON_3] [PERSON_4]. That's [SPELLED_NAME_2]. Case ID [CDCR_ID_1].

Given that specific information such as the offender's name, hearing date and time are publicly available online on the CDCR's hearing calendar web page[6], we remove numerical values, such as date, time and age, by replacing the original data with fine-grained labels, using manually crafted rules consistently across all transcripts. The full date value is split into individual components and replaced by a type-specific label based on specific conditions. For instance, ordinal numbers ("1st", "2nd", ...) in the context of dates are replaced with the label [DAY]. Months and days of the week are detected using regular expressions and replaced with [MONTH] and [DAY_OF_WEEK] labels respectively. Four digit numbers under the DATE entity represent years and are replaced with [YEAR]. Decades ("20s", "30s", ...) are replaced with the label [DECADE]. Formatted dates that resemble patterns like "MM/DD/YYYY" are simply replaced by [DATE]. Any other numbers that do not satisfy the aforementioned conditions are replaced with [NUMBER]. Example (2) illustrates the handling of these specific entities.

(2) a. **Original:**
Today is 05/13/2012, 10:30, he was convicted back on Monday the 15th of June, 2011 at the age of 33 years old.

b. **Anonymized:**
Today is [DATE], [TIME], he was convicted back on [DAY_OF_WEEK] the [DAY] of [MONTH], [YEAR] at the age of [AGE] years old.

To generate more fine-grained and accurate pseudonymization labels for NRP, LOCATION and ORGANIZATION, we use BART, a zero shot classification model by Facebook (Facebook, 2024). BART is trained on the Multi-Genre Natural Language Inference (MultiNLI) dataset (Williams et al., 2018), which includes a diverse range of written and spoken data sources, including letters, Oxford University Press, press releases from government websites as well as transcriptions of face-to-face conversations and telephone calls.
The NRP entity, derived from Presidio, combines an individual's Nationality, Religion, and Political group affiliations – three distinct yet interrelated types of sensitive information. Similarly, LOCATION entities are complex, including diverse geographical information such as states, counties, cities, and countries.
BART's role is to distinguish and categorize the data within these multifaceted entities. By doing so, it enhances the contextual relevance and overall utility of the final dataset. This approach allows for more precise and meaningful pseudonymization while maintaining the analytical value of the data. Example (3) illustrates the conversion of each entity type into appropriate category labels, demonstrating the granularity and accuracy achieved through this method.

(3) a. **Original:**
He lived in Connecticut but then moved to California. He is a Canadian citizen from Canada and works with the California City Police Department.

b. **Anonymized:**
He lived in [STATE_1] but then moved to [STATE_2]. He is a [NATIONALITY_1] citizen from [COUNTRY_1] and works with the [POLICE_DEPARTMENT_1].

[6] https://www.cdcr.ca.gov/bph/2024/02/07/august-2024-hearing-calendar

For the remaining entities, each unique occurrence is replaced with the specific category label and their corresponding sequential number.

Appendix C shows that PERSON entities are the most frequent throughout parole hearing transcripts, accounting for 34.92% per 1000 tokens. This result is none of a surprise, as discussions typically revolve around the inmate, their victims, accomplices and family members. DATE, AGE and LOCATION entities, while important, appear at varying frequencies with only DATE occurring just over 3% per 1000 tokens. This reflects the importance of discussing the inmate's age at specific life events, such as the crime or key moments during incarceration, as well as references to locations related to their past or to future parole plans. Similarly, the frequent mention of DATE and LOCATION entities can be attributed to discussions about important milestones in the inmate's history or potential locations for future parole plans.

5.1 Challenges & Limitations

Despite the promising results we obtained from the detection of PII through NER tools and the pseudonymization technique, certain issues and constraints still need to be addressed. By law, all parole hearings transcripts are required to provide verbatim records of the dialogue. This standardized format presents both advantages and challenges to the anonymization process. We identified two primary categories of issues: (1) errors stemming from the NER tools that result in misclassified entities, and (2) entities that are completely missed by the NER tools, leading to unintended leakage of sensitive information.

Table 4 shows the number of misclassified entities we observed in the 100 manually analyzed transcripts. The most common misclassifications occur between LOCATION and PERSON. This is mainly caused by names that refer to both people and places (e.g. Georgia or Dallas) and are often incorrectly tagged as LOCATION by the NER tools.

Classification errors between ORGANIZATION and PERSON entities occurred 45 times, primarily because people's names appeared within official organization names. Misclassifications between CDCR ID and LOCATION entities are due to the regular expressions for CDCR ID entities matching zipcodes and post office box (PO Box) numbers. The least frequent mismatches, such as NRP | ORGANIZATION and EMAIL_ADDRESS | PERSON come from specific cases where an abbreviation was incorrectly misclassified by the NER tools as an NRP and a part of the email including a person's name is misclassified as a PERSON.

Mismatched entity pairs	**Count**
LOCATION | PERSON	127
ORGANIZATION | PERSON	45
LOCATION | ORGANIZATION	26
AGE | DATE	33
CDCR_ID | LOCATION	20
DATE | TIME	3
AGE | TIME	3
NRP | ORGANIZATION	1
EMAIL_ADDRESS | PERSON	1

Table 4: Counts of misclassified entity labels in 100 analyzed transcripts. In the format "LABEL A" | "LABEL B", the first label represents the incorrect classification by the NER, while the second label indicates the correct classification.

As already reported on in §4.4, the misclassification of non-sensitive entities that do not require anonymization, such as abbreviations of parole hearing-specific terms, frequently results in their attribution to the ORGANIZATION category, resulting in an over-anonymization of the data. Expanding the whitelist can help address this challenge.

Another issue was found in cases where the CDCR ID's first letter was spelled out phonetically by using a corresponding name. For example, "Victor 12345" was used to indicate that the CDCR ID begins with "V", resulting in the full CDCR ID "V12345". The NER tools misclassified "Victor" as a PERSON entity, leading to incorrect pseudonymization. Changing PERSON entities followed by CDCR IDs to just CDCR ID might seem like a straightforward solution, but is complicated by the fact that the inmates' names are often immediately followed by their CDCR ID, without any separating punctuation. We decided to accept these minor errors in the pseudomized text and will address the requirement of a more nuanced approach for phonetically spelled out names in the future.

The pseudonymization approach relies on a dictionary to assign distinct tags to each identified PERSON entity (e.g. PERSON_1, PERSON_2). This procedure ensures maintenance of privacy, while simultaneously allowing different individuals to remain distinguishable. However, this method faces challenges whenever a name that has already

been assigned a label is later misspelled and therefore not found in the dictionary. As a result, a new and incorrect label is generated, leading to multiple labels for the same person, compromising the consistency and reliability of the redacted transcript. Example (4) showcases the erroneous anonymization of names due to typos. Even though the same person is referenced in both sentence fragments, two different tags are generated for the last name due to the misspelling. This issue might not leak PII, but compromises the data's integrity and underscores the dependence on error-free transcriptions for anonymization.

(4) a. **Original name:**
[Mark] [Stevenson] is present.
[Mark] [Stevenston] here is...

b. **Anonymized name:**
[Person_1] [Person_2] is present.
[Person_1] [Person_3] here is...

Due to the nature of spoken language, the transcripts include passages where multiple speakers talk simultaneously, leading to fragmented utterances appearing on separate lines. This can cause sensitive information to be split between lines and potentially remain undetected by the NERs.

While incorrect category labels reduce the utility of the anonymized transcripts and can lead to confusion, the consequences of missed entities are more severe, as they result in the direct leakage of PII.

Our approach to anonymization faces several challenges that highlight the inherent trade-offs between data privacy and and analytical utility. One notable decision we made was to not anonymize gender information, including gender specific pronouns in the text, given that the majority of inmates seeking parole in California are male.

Another significant limitation stems from the temporal context of the hearings. In the specific case of our dataset, many transcripts contain references to the COVID-19 pandemic, which inadvertently narrows the timeframe of the hearings to 2019 and 2021. This temporal information, while valuable for understanding the unique circumstances of conducting the hearings via video conferencing, also increases the potential for re-identification.

The same applies to high-profile cases that received significant media attention, such as parole hearings for individuals involved in the Manson murders. For these instances, achieving complete anonymization is especially difficult. The risk of re-identification cannot be entirely eliminated without significantly compromising the analytical utility of the transcripts.

While some of the identified limitations are inherent to the nature of the data and cannot be fully resolved, we hope to address the remaining in future work by enhancing the automatic detection of sensitive data. However, to ensure that no sensitive information has been overlooked, a final manual review before publication of the data is essential to prevent unintended data exposure and maintain ethical standards.

6 Conclusion

We introduced a novel approach for the anonymization of direct and indirect identifiers in parole hearing transcripts, offering a way to protect sensitive personal information while preserving the utility of the data for different kind of research purposes. Our methodology combines Named Entity Recognition tools with pseudonymization techniques and addresses the challenges posed by this specific type of legal dialogue. Despite the limitations of current NER tools, leading to misclassified entity types and errors arising from misspelled names, our approach successfully cleans the transcripts from sensitive data in the majority of cases. While our approach provides a strong methodology to reliably pseudonymize parole hearing transcripts, a thorough manual review of the transcripts before publication is still mandatory in order to avoid any unintended data leakage. Future research should focus on addressing the remaining limitations, with the ultimate goal of enhancing data privacy without sacrificing data utility.

Ethical considerations

While unanonymized parole hearing transcripts can be officially requested via email from the Board of Parole Hearings in California, our goal is to protect the privacy of the individuals involved. At the same time, we want to enable researchers to investigate linguistic strategies in parole hearings, which could lead to improved understanding of decision-making processes and potentially contribute to more equitable outcomes. However, we acknowledge that despite our best efforts at anonymization, a small risk

of re-identification remains. This is especially true for high-profile cases. We would like to note that while we are using a technically publicly available dataset, we cannot guarantee that all participants, especially victims and their next of kin, are fully aware that these transcripts can be requested by anyone, regardless of scientific usage or other purposes. This underscores our commitment to robust anonymization and ethical handling of the data.

Despite the promising results of our anonymization process, we still wait for official confirmation that the anonymized dataset can be published online. We will also seek ethical clearance before releasing the anonymized dataset to confirm compliance with relevant regulations and standards. We have published the code for the anonymization process on GitHub[7].

Acknowledgements

We want to thank the California Department of Corrections and Rehabilitation (CDCR) for providing us with the parole hearing transcripts.
The work reported on in this paper was funded by the Deutsche Forschungsgemeinschaft (DFG – German Research Foundation) under Germany's Excellence Strategy – EXC-2035/1 – 390681379 as part of the project "Inequality in Street-level Bureaucracy: Linguistic Analysis of Public Service Encounters".

References

Luis Adrián Cabrera-Diego and Akshita Gheewala. 2024. PSILENCE: A Pseudonymization Tool for International Law. In *Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024)*, pages 25–36. Association for Computational Linguistics.

Ingrid Espinoza, Steffen Frenzel, Laurin Friedrich, Wassiliki Siskou, Steffen Eckhard, and Annette Hautli-Janisz. 2024. PSE v1.0: The First Open Access Corpus of Public Service Encounters. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13315–13320, Torino, Italia. ELRA and ICCL.

Facebook. 2024. BART-Large-MNLI: A Zero-Shot Classification Model. https://huggingface.co/facebook/bart-large-mnli. Accessed: 2024-08-04.

[7] https://github.com/abedit/Automated-Anonymization-of-Parole-Hearing-Transcripts

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 363–370. https://nlp.stanford.edu/software/CRF-NER.html.

Shayna Gardiner, Tania Habib, Kevin Humphreys, Masha Azizi, Frederic Mailhot, Anne Paling, Preston Thomas, and Nathan Zhang. 2024. Data Anonymization for Privacy-Preserving Large language Model Fine-Tuning on Call Transcripts. In *Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024)*, pages 64–75, St. Julian's, Malta. Association for Computational Linguistics.

Jenny Hong, Derek Chong, and Christopher Manning. 2021a. Learning from Limited Labels for Long Legal Dialogue. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 190–204, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jenny Hong, Catalin Voss, and Christopher Manning. 2021b. Challenges for Information Extraction from Dialogue in Criminal Law. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 71–81, Online. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Omri Mendels and Avishay Balter. 2020. Presidio: Context Aware, Pluggable and Customizable Data Protection and De-identification SDK for Text and Images. Microsoft.

Arttu Oksanen, Minna Tamper, Jouni Tuominen, Aki Hietanen, and Eero Hyvönen. 2019. ANOPPI: A Pseudonymization Service for Finnish Court Documents. In *Legal Knowledge and Information Systems*, pages 251–254. IOS Press.

Bruno Ribeiro, Vitor Rolla, and Ricardo Santos. 2023. INCOGNITUS: A Toolbox for Automated Clinical Notes Anonymization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 187–194, Dubrovnik, Croatia. Association for Computational Linguistics.

Tom Schamberger. 2021. Customizable Anonymization of German Legal Court Rulings using Domain-specific Named Entity Recognition. Master's thesis, Department of Mathematics, Technical University Munich. Master's Thesis.

Graham Todd, Catalin Voss, and Jenny Hong. 2020. Unsupervised anomaly detection in parole hearings using language models. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 66–71, Online. Association for Computational Linguistics.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 214–229, New York, NY, USA. Association for Computing Machinery.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

A Automatic Annotation Example

The following is an extract from one of the hearing transcripts. To preserve the privacy of the people involved, the names and IDs have been altered. The labels are visually marked according to the tool that detected them: labels coming from spaCy are in blue, labels from Presidio are in green, and StanfordNER labels are in red.

1. PRESIDING COMMISSIONER JONES: All right. Good afternoon. Today's date, September 1st, 2021. Time is, uh, 1:30 PM. This is the initial parole suitability hearing for inmate- Correction. This is the first subsequent parole suitability hearing for inmate Kevin Richardson, R-I-C-H-A-R-D-S-O-N, CDCR number L90314. Inmate Richardson is not present at the hearing room at San Quentin State Prison. Uh, we were notified today that the inmate is currently out at the hospital and, uh, is currently unavailable for his hearing. Uh, so, uh, let's uh, take appearances. Uh, we are conducting this hearing by video conference. So, let's take appearances on who's here today. Uh, we'll have the Panel members go first. My name is Alyssa Jones J-O-N-E-S, Commissioner with the Board of Parole Hearings.

In this step, invalid labels are filtered out and the remaining labels are cleaned. The identification of "Board of Parole Hearing" as ORGANIZATION is dropped, as it is a non-sensitive term. The same applied to the DATE label of "Today's", since it does not contain a numerical component.

2. PRESIDING COMMISSIONER JONES: All right. Good afternoon. Today's date, September 1st, 2021. Time is, uh, 1:30 PM. This is the initial parole suitability hearing for inmate- Correction. This is the first subsequent parole suitability hearing for inmate Kevin Richardson, R-I-C-H-A-R-D-S-O-N, CDCR number L90314. Inmate Richardson is not present at the hearing room at San Quentin State Prison. Uh, we were notified today that the inmate is currently out at the hospital and, uh, is currently unavailable for his hearing. Uh, so, uh, let's uh, take appearances. Uh, we are conducting this hearing by video conference. So, let's take appearances on who's here today. Uh, we'll have the Panel members go first. My name is Alyssa Jones J-O-N-E-S, Commissioner with the Board of Parole Hearings.

Finally, any overlapping labels are separated.

3. PRESIDING COMMISSIONER JONES: All right. Good afternoon. Today's date, September 1st, 2021. Time is, uh, 1:30 PM. This is the initial parole suitability hearing for inmate- Correction. This is the first subsequent parole suitability hearing for inmate Kevin Richardson, R-I-C-H-A-R-D-S-O-N, CDCR number L90314. Inmate Richardson is not present at the hearing room at San Quentin State Prison. Uh, we were notified today that the inmate is currently out at the hospital and, uh, is currently unavailable for his hearing. Uh, so, uh, let's uh, take appearances. Uh, we are conducting this hearing by video conference. So, let's take appearances on who's here today. Uh, we'll have the Panel members go first. My name is Alyssa Jones J-O-N-E-S, Commissioner with the Board of Parole Hearings.

The pseudonymization method is applied. The changed entities are in bold.

4. PRESIDING COMMISSIONER **[PERSON_2]**: All right. Good afternoon. Today's date, **[MONTH] [DAY]**, **[YEAR]**. Time is, uh, **[TIME]** PM. This is the initial parole suitability hearing for inmate- Correction. This is the first subsequent parole suitability hearing for inmate **[PERSON_7] [PERSON_8]**, **[SPELLED_NAME_PERSON_8]**, CDCR number **[ID_1]**. Inmate **[PERSON_8]** is not present at the hearing room at **[PRISON_1]**. Uh, we were notified today that the inmate is currently out at the hospital and, uh, is currently unavailable for his hearing. Uh, so, uh, let's uh, take appearances. Uh, we are conducting this hearing by video conference. So, let's take appearances on who's here today. Uh, we'll have the Panel members go first. My name is **[PERSON_1] [PERSON_2]**, **[SPELLED_NAME_PERSON_2]**, Commissioner with the Board of Parole Hearings.

B Metrics of Presidio, spaCy and StanfordNER

Entity Type	Presidio			spaCy			StanfordNER		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
PERSON	0.982	0.995	0.989	0.984	0.899	0.939	0.978	0.789	0.873
LOCATION	0.932	0.905	0.919	0.903	0.590	0.714	0.769	0.848	0.807
ORGANIZATION	0.976	0.341	0.505	—	—	—	0.688	0.487	0.571
SPELLED_NAME	1.000	0.995	0.997	—	—	—	—	—	—
CDCR_ID	0.933	0.996	0.964	—	—	—	—	—	—
DATE	0.647	0.611	0.628	0.921	0.780	0.845	—	—	—
TIME	0.880	0.621	0.728	0.973	0.796	0.875	—	—	—
AGE	0.931	0.126	0.222	0.903	0.921	0.912	—	—	—
HEIGHT	1.000	0.800	0.889	—	—	—	—	—	—
EMAIL_ADDRESS	1.000	0.750	0.857	—	—	—	—	—	—
URL	0.667	1.000	0.800	—	—	—	—	—	—
NRP	0.750	0.830	0.788	—	—	—	—	—	—
SPELLED_OUT_ITEM	1.000	1.000	1.000	—	—	—	—	—	—
PHONE_NUMBER	1.000	1.000	1.000	—	—	—	—	—	—
Overall	**0.953**	**0.885**	**0.918**	**0.971**	**0.801**	**0.878**	**0.951**	**0.629**	**0.757**

Table 5: Precision (P), Recall (R), and F1-score (F1) for the automatic labeling process run by each tool separately. Blank cells are due to spaCy and StanfordNER not covering certain entities, while Presidio covers every entity.

C Pseudonymization Statistics

Entity Type	Total	Average	Frequency
PERSON	153,495	459.57	34.9242%
LOCATION	8,410	25.18	1.9135%
ORGANIZATION	5,758	17.24	1.3101%
SPELLED_NAME	2,141	6.41	0.4871%
CDCR_ID	1,565	4.69	0.3561%
DATE	13,735	41.12	3.1251%
TIME	3,194	9.56	0.7267%
AGE	11,092	33.21	2.5237%
HEIGHT	103	0.31	0.0234%
EMAIL_ADDRESS	4	0.01	0.0009%
URL	17	0.05	0.0039%
NRP	735	2.20	0.1672%
SPELLED_OUT_ITEM	57	0.17	0.0130%
PHONE_NUMBER	4	0.01	0.0009%

Table 6: Number of entities pseudonymized, as well as the average of each entity per file and how frequently each entity is pseudonymized per 1000 tokens.

D Ablation Study

Entity Type	Standard process	Excluding 1st page
PERSON	460,062	333,223
LOCATION	23,197	25,245
ORGANIZATION	10,830	9,826
SPELLED_NAME	2,559	2,559
CDCR_ID	1,569	1,569
DATE	38,653	38,653
TIME	7,520	7,520
AGE	12,827	12,827
HEIGHT	103	103
EMAIL_ADDRESS	4	4
URL	21	21
NRP	1,095	1,082
SPELLED_OUT_ITEM	57	57
PHONE_NUMBER	4	4
Numerical values	14,523	14,523
Total	**573,024**	**447,216**

Table 7: Comparison of entity detection results: standard process vs. process excluding first page information.

Entity Type	Standard process			Excluding 1st page		
	P	R	F1	P	R	F1
PERSON	0.981	0.989	0.985	0.979	0.911	0.944
LOCATION	0.846	0.946	0.893	0.592	0.895	0.713
ORGANIZATION	0.768	0.739	0.754	0.721	0.560	0.630
SPELLED_NAME	1.000	0.995	0.997	1.000	0.995	0.997
CDCR_ID	0.933	0.996	0.964	0.933	0.996	0.964
DATE	0.883	0.968	0.923	0.883	0.968	0.923
TIME	0.977	0.943	0.960	0.977	0.943	0.960
AGE	0.903	0.926	0.914	0.903	0.926	0.914
HEIGHT	1.000	0.800	0.889	1.000	0.800	0.889
EMAIL_ADDRESS	1.000	0.750	0.857	1.000	0.750	0.857
URL	0.667	1.000	0.800	0.667	1.000	0.800
NRP	0.765	0.830	0.796	0.765	0.830	0.796
SPELLED_OUT_ITEM	1.000	1.000	1.000	1.000	1.000	1.000
PHONE_NUMBER	1.000	1.000	1.000	1.000	1.000	1.000
Overall	**0.955**	**0.972**	**0.963**	**0.930**	**0.907**	**0.918**

Table 8: Performance metrics comparison: standard process vs. process excluding first page information on a subcorpus of 100 transcripts.