MARI: Semi-Automated, Human-in-the-Loop Redaction of Text Corpora

Emma I. C. Peterson
University of Chicago

Valerie Zhao
University of Chicago

Dan Byrne
University of Chicago

Blase Ur
University of Chicago

Abstract

Large datasets consisting of unstructured text can provide valuable research insights, yet require redaction to protect participant privacy. Unfortunately, manual redaction of large datasets is impractical, while existing tools both overzealously redact information that provides valuable insights and miss information that makes human subjects identifiable. To address this gap, we present a novel human-in-the-loop redaction framework informed by interviews with 13 researchers who steward such datasets. We designed the MARI (Mostly Automated Redaction of Identifiers) tool, which merges classification techniques with knowledge bases and linguistic inference to find identifiable data in naturalistic text.

1 Introduction and Background

In the social sciences, research questions can require large datasets of unstructured text to answer. For example, one can study linguistic patterns by transcribing naturalistic recordings of conversations. To protect participant privacy, *redaction* de-identifies participants by concealing or removing personal information, such as names and locations. Unfortunately, datasets of large unstructured text are typically too large for manual redaction. Consequently, such datasets—despite their value—are rarely shared with the scientific community. To answer related research questions, researchers must collect similar datasets on their own, incurring significant costs.

Automated data loss prevention tools, such as Google's Cloud DLP [2] and Amazon's Macie [1], can redact identifiers in large datasets, but are inflexible to nuances of personal information because they mostly rely on pattern matching, list comparisons, and simple classifiers. They may fail to detect items such as unique names, idiosyncratic phrases, or identifiers that vary across contexts (e.g., of gender-fluid participants). Stricter settings may inadvertently remove information that does not make human subjects identifiable, yet provides critical research insights (e.g., verbal utterances of fictional names that mark the participant's language development). Lastly, these tools can miss transcription errors a human can infer (e.g. "Mike" may be mistranscribed as "bike").

We address these gaps by designing a semi-automatic, human-in-the-loop framework for redacting research data. To this end, we first interviewed 13 principal investigators who have stewarded large datasets of unstructured text, and we present findings about their redaction needs. We then designed—and are currently implementing—**MARI, a semi-automatic redaction tool** utilizing a human-in-the-loop framework informed by the interview findings. With feature engineering geared towards personal information, linguistic analysis, and information-theoretic scoring, our approach tries to maximize data utility, generalization, and coverage.

2 Tool Design Goals

In our interviews, PIs concurred that the resource cost of manual redaction hindered their ability to share large datasets with the broader scientific community. While the majority of interviewees agreed with our taxonomy of potentially sensitive information to redact (Table 1), they also highlighted that data sharing and redaction concerns differed slightly by field methodology and philosophy, as well as by historically and politically fraught demographic categories such as sexuality and gender expression and identification.

Given these insights, we designed MARI, a tool to efficiently assist researchers in redacting unstructured text data. Our design had the following goals. To avoid costs of manual labor, the tool should redact information automatically. However, it must also be sensitive to the users' specific research needs, which may vary based on the discipline, methodology,

USENIX Symposium on Usable Privacy and Security (SOUPS) 2023.
August 6–8, 2023, Anaheim, CA, USA

Table 1: Our taxonomy of identifying information to redact, which was iterated upon based on the interviews.

Taxonomy Category	Taxonomy Sub-Categories	Examples
Identifiers	Personal Names, Nicknames, Personal Identity, Numbers	*Legally given name, Diminutives, SSN, EIN*
Demographics	Age, Sex, Gender, Pronouns, Sexuality, Race & Ethnicity, Education, Profession, Health Status	*Date of Birth, LQBTQ+, Niche Job, Rare Disease, HIV Positive, Mixed Race*
Locations	Country, State, City, Postal Code, Address, Landmark, Business	*United States, Illinois, Chicago, 5307 S Woodlawn Ave, The Bean, Jimmy's Tap*
Dates & Events	Publicly Recognized Dates & Events, Personal Dates & Events	*Thanksgiving, Christmas, Cancer-Remission Anniversary*
Linguistic Patterns	Regional Dialects, Code-Switching, Unique Vocabularies, Idiosyncratic Expressions	*African American Vernacular English, Scots-English, Spanglish, Parmesan Cheese == "Pasta Sugar"*
Personal Interests & Activities	Traditions, Group Membership, Cultural References, Popular Culture Participation, Hobbies	*University / school traditions, belonging to native tribe or military, member of a small fandom*

and research questions. Furthermore, as automatic tools may be prone to mistakes regardless, the user should also be able to adjust the final redactions as needed.

3 MARI Redaction Tool Implementation

MARI is tailored to unstructured text data and builds on existing libraries. It runs locally and will be open-sourced.

To effectively analyze unstructured text, we first teach the tool how to parse and model words. The structures we use are *tokens*, *types*, and *chunks*. Tokens consist of attributes pertaining to a given instance of a sequence of characters within a document. These features consist largely of items we can glean through natural language processing. Types—also referred to as global tokens—encapsulate *all* tokens of a given sequence of characters (case insensitive) as long as the items are homographs. We also include type-specific attributes in addition to those given by token instances. While some of these are gained by processing the text and aggregating features from individual tokens, we gather attribute insights from the transcription itself, as well as from a curated knowledge base (a custom subset of WikiData). Chunks consist of a sequence of multiple tokens and will be analyzed separately for linguistic patterns (e.g., code-switching, idiosyncratic speech).

Once the text is processed, we pass tokens to a series of classifiers, one for each category in our taxonomy. Each classifier determines the probability that a given token or type is identifiable based on its inherent constraints pertaining to the classifier's category. We note that the use of the term "classifier" for each algorithm is not entirely based on logistic regression classifiers; rather, we perform logistic regression on a *partial* set of attributes, and we may use the other attributes in a more nuanced analysis driven by linguistic intuition. We can also examine the *textual context* in which the tokens are presented. For instance, certain non-discrete properties from the knowledge base can impact how a term is viewed.

A user can review MARI's automated redactions, as well as customize settings to fit their data. MARI highlights not only terms that are redacted, but also those that are close to our identifiability thresholds. This allows users to examine edge cases and adjust granularity for redaction (i.e., controlling strictness). Users can add or remove terms to redact. They can also select the taxonomic categories to redact or preserve. We account for potential typos and mistranscriptions by flagging similar IPA phonetic transcriptions and calculating each token's Levenshtein distance to redacted terms.

4 Discussion

Our tool focuses on filling these needs in context of our taxonomy. With a separate classifier for each category, we can detect more nuanced instances of identifiable data. Preliminary evaluations of proof-of-concept classifiers on subsets of large, naturalistic speech corpora show low rates of over-redaction and missed terms, pointing to a promising future in semi-automated redaction of naturalistic data. Eventually, our framework should be able to generalize to most naturalistic data of human subjects studies. The tool will enable researchers to efficiently and accurately redact data at far less cost of labor and time than manual redaction, thus enabling more data-sharing for the benefit of the scientific community.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. CNS-2210193 and by a Meta Privacy-Enhancing Technologies Research Award. We thank Susan Goldin-Meadow, Marisa Casillas, Chenhao Tan, and Ruthe Foushee for guidance, as well as Oliver Cai and Brianna Zeng for helping with MARI's implementation.

References

[1] Amazon. Macie: Discover and protect your sensitive data at scale, 2023. https://aws.amazon.com/macie/.

[2] Google. Cloud Data Loss Prevention (DLP), 2020. https://cloud.google.com/dlp/.