PII/Sensitive Information Identification - Capstone Project

This document serves as a place to gather relevant background literature and data sources, as well as a place to document proposal ideas and discussion.

Please feel free to add resources, make comments, and edit sections.

Background Literature

This section captures any relevant background literature for the project. We have broken this up into two sections, non-technical background and ML/Stats/Technical background literature.

Non-Technical Lit Review:

- Commercial text extraction and PII detection services:
  - Micro Focus IDOL: https://www.microfocus.com/documentation/idol/IDOL_12_8/KeyviewFilterSDK_12.8_Documentation/Guides/pdf/KeyViewFilterSDK_12.8_CProgramming.pdf and https://www.microfocus.com/documentation/idol/IDOL_12_8/EductionSDK_12.8_Documentation/Guides/pdf/Eduction_12.8_UserProgramming_en.pdf
  - PII Tools: https://documentation.pii-tools.com/
  - AWS: https://docs.aws.amazon.com/comprehend/latest/dg/how-pii.html
  - Google: https://cloud.google.com/dlp#section-5
  - Microsoft : https://docs.microsoft.com/en-us/azure/search/cognitive-search-skill-pii-detection
  - Gretel (Privacy Ops): https://gretel.ai/blog/automate-detecting-sensitive-personally-identifiable-information-pii-with-gretel

Technical Lit Review:

https://en.wikipedia.org/wiki/Named-entity_recognition

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

[1] [Dissertation Thesis] Detecting and Protecting Personally Identifiable Information through Machine Learning Techniques by Carlos Jorge Augusto Pereira da Silva

[2] The European Parliament and the Council of the European Union. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation), 2016.

[3] [Paper] Ishna Neamatullah, Margaret M Douglass, Li wei H Lehman, Andrew Tomas Reisner, Mauricio Villarroel, William J Long, Peter Szolovits, George B Moody, Roger Greenwood Mark, and Gari D Clifford. Automated de-identification of free-text medical records. BMC Medical Informatics and Decision Making, 8(1):32–32, 2008.

[4] JRC-Names Database

[5] [Paper] Georgios Petasis, Alessandro Cucchiarelli, Paola Velardi, Georgios Paliouras, Vangelis Karkaletsis, and Constantine D. Spyropoulos. Automatic adaptation of proper noun dictionaries through cooperation of machine learning and probabilistic methods. In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, pages 128–135, 2000

[6] Ralph Grishman and Beth Sundheim. Message understanding conference-6: a brief history. In COLING '96 Proceedings of the 16th conference on Computational linguistics - Volume 1, volume 1, pages 466–471, 1996.

[7] [Paper] Eneko Agirre, Elena Garcia, Mikel Lersundi, David Martinez, and Eli Pociello. The basque task: Did systems perform in the upperbound? In Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems, pages 9–12, 2001.

[8] [Paper] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In Proceedings of CoNLL-2003. Edmonton, Canada, 2003.

[9] Ralph Weischedel, Ada Brunstein. BBN Pronoun Coreference and Entity Type Corpus. 2005.

[10] [Paper] George R. Doddington, Alexis Mitchell, Mark A. Przybocki, Lance A. Ramshaw, Stephanie M. Strassel, and Ralph M. Weischedel. The automatic content extraction (ace) program tasks, data, and evaluation. In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC-2004), 2004.

Patrice Bellot, Véronique Moriceau, Josiane Mothe, Eric SanJuan, and Xavier Tannier. Inex tweet contextualization task: Evaluation, results and lesson learned. Information Processing & Management, 52(5):801 – 819, 2016.

[12] [Paper] Krisztian Balog, Pavel Serdyukov, and Arjen P. de Vries. Overview of the trec 2010 entity track. In NIST Special Publication, 2010.

[13] [Paper] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. Journal of Machine Learning Research, 12(76):2493–2537, 2011.

[14] [Paper] Valentin Barriere and Amaury Fouret. May I check again? — a simple but efficient way to generate and use contextual dictionaries for named entity recognition. application to French legal texts. In Proceedings of the 22nd Nordic Conference on Computational Linguistics, pages 327–332, Turku, Finland, September–October 2019. Linköping University Electronic Press.

[15] [Paper] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In ICML '01 Proceedings of the Eighteenth International Conference on Machine Learning, pages 282–289, 2001.

[16] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. CoRR, abs/1603.01360, 2016.

[17] Alexei Baevski, Sergey Edunov, Yinhan Liu, Luke Zettlemoyer, and Michael Auli. Clozedriven pretraining of self-attention networks. In 2019 Conference on Empirical Methods in Natural Language Processing, 2019.

[18] [Google Cloud] De-identification and re-identification of PII in large-scale datasets using Cloud DLP

[19] [Google Cloud] Cloud DLP

[20] Detecting Personal Data within API Communication Using Deep Learning

[21] Huang et al., Removing Personally Identifiable Information from Shared Dataset for Keystroke Authentication Research

[22] [Paper] Gregory J. Matthews and Ofer Harel, Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy, Statistics Surveys, 2011

[23] [Paper] Sweeney, L., 2002b. k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge Based Systems 10 (5), 557–570.

[24] How to Search Datasets for Personally Identifiable Information

[25] Data privacy and GDPR

[26] How to Build Realistic but Fake PII

[27] Automate Detecting Sensitive Personally Identifiable Information (PII)

[28] Using Powershell to report on files containing PII (Personally Identifiable Information)

[29] Finding personally identifiable information (PII) with PowerShell.

[30] Search PII Using Powershell

[31] GDPR discovery, protection, and reporting in the dev/test environment

[32] Using CLI for Microsoft 365 and PowerShell to report on SharePoint Document Library files containing PII

[33] Prevent PII Harvesting with PerimeterX

[34] Bier and Prior, Detection and Labeling of Personal Identifiable Information in E-mails, 2014

Data Sources:

In this section we discuss both public and private options for sourcing data.

Public API’s/Solutions/Open Source Code:

In this section we list and discuss publicly available solutions or open source code bases relevant to the project.
See the above listing in the non-technical review section.

Project Discussion

Let's use this section to bring up any questions, proposed directions, or ideas for discussion.

Given the sheer number of off the shelf PII detection and redaction libraries available commercially. Should we focus on a specific problem space e.g. detection of PII / PHI in messaging (Slack, Teams, Google Mail etc.) or source code repos or perhaps even API calls to Falcon?

Some Narrower Project Ideas:

1. Preservation Of Training Data:

A major use case for PII identification is the preservation of data that is potentially useful in training machine learning models, but would otherwise need to be deleted due to the potential inclusion of PII. Examples could include script contents, documents, command line contents, filenames etc.

This is both an important use case for Crowdstrike, but also presents some interesting avenues for the project. Evaluation data is readily available internally, so any model that the team produced could be applied to real cyber PII identification use-cases. The team could take this in many different directions:

- Simply train general PII/NER model on public data and evaluate against CS internal cyber use cases or
- Explore optimal transfer learning architecture. How to generate model using one dataset that in expectation will work best on the unobserved CS internal data
- Explore optimizing for machine learning model signal preservation: because the end goal is to preserve data for ML model training, we want to remove PII while preserving any important signals that could indicate malicious activity etc. How to optimize for this slightly different target?
- Explore active learning/model optimization strategies using limited feedback from internal evaluation. Can you build a system that can learn from limited information and metrics produced by evaluation on internal data sets with out access to the full data set?

2. Improved Synthesis of PII Data for Training:

Many NER and PII models are readily trained on semi-synthetic data. For example, the names databases discussed above (real data) can be injected into documents to make synthetic labeled PII data.

The team could explore improving generation of synthetic PII data, with many directions to take this in as well:

- As part of any of the training data preservation projects listed above

- With a specific focus on improving model performance on cyber-security use cases
- Train generative models to build realistic PII. This could be part of a GAN's style approach for building a better PII detector as well.
-

3. Beat external benchmarks:

A simple approach would be to aim to simply beat publicly available API's and benchmark systems for PII identification. This is both straightforward, but difficult...