NERPII: A Python Library to Perform Named Entity Recognition and Generate Personal Identifiable Information

Simona **Mazzarino**[1], Andrea **Minieri**[1] and Luca **Gilli**[1]

[1]*Clearbox AI, Turin, Italy*

Abstract

Nowadays, the convergence of Artificial Intelligence and data privacy is of crucial importance. This paper introduces NERPII, a Python library utilizing Named Entity Recognition (NER) and synthetic data generation to identify and protect Personal Identifiable Information (PII). We discuss the architecture of NERPII and provide a concise tutorial on its application, demonstrating how to extract entity information from datasets containing personal data and synthesize new PII while preserving data characteristics. Additionally, the study discusses the library's potential contributions and implications for future research. In conclusion, NERPII emerges as a practical tool for addressing ethical concerns related to data privacy in the AI domain.

Keywords

Named Entity Recognition, Personal Identifiable Information, Synthetic Data, Data Privacy, Python library

1. Introduction

In the era of data-driven insights and technological progress, the intersection of Artificial Intelligence (AI) and data privacy has assumed an increasingly important role. AI has profoundly reshaped societal dynamics, influencing everything from finance and healthcare to transportation and entertainment. However, this progress has not been devoid of ethical concerns, particularly those related to the exposure of Personal Identifiable Information (PII) [2]. PII refers to any data or information that can be used to identify, contact, or locate a specific individual [3]. PII includes a wide range of data elements, and it can be either sensitive or non-sensitive. Sensitive PII includes information such as Social Security numbers, driver's license numbers, financial account numbers, and medical records, while non-sensitive PII may include names, addresses, phone numbers, email addresses, and other information that, when combined, can be used to identify a person.

The protection of PII is a critical aspect of privacy and data security, as it involves safeguarding individuals' personal information from unauthorized access, use, or disclosure. Various laws

and regulations, such as the European Union's General Data Protection Regulation (GDPR) [4] and the United States' Health Insurance Portability and Accountability Act (HIPAA) [5], govern the collection, storage, and handling of PII to ensure individuals' privacy rights are respected and their information is kept secure.

In this perspective, some Natural Language Processing (NLP) techniques, such as Named Entity Recognition (NER), a challenging learning problem that involves processing a text and identifying certain occurrences of words or expressions as belonging to particular categories of Named Entities (NE) [6, 7], if combined with the potential of synthetic data generation, emerge as a tool that can be used with an ethical purpose, that is, to identify and categorize entities, including names, phone numbers, credit card number, and other potentially sensitive information that then can be replaced with synthetic data. The synthesis of data is a technique that involves generating artificial data while preserving the statistical characteristics of real data [8, 9]. So, combining synthetic data with NER can be a strong strategy to mitigate the risk of data breaches and preserve data privacy.

So, NER, as mentioned before, is a technique often used on unstructured data, namely texts. However, companies or institutions often possess entire structured datasets, such as Excel or CSV files, containing numerous PII records. The challenge, therefore, was to create a tool capable of performing NER on tabular data in order to recognize PII and replacing it with synthetic data.

In literature, numerous tools have been created to associate semantic types with table columns. For instance, Hulsebos et al. [10] introduced Sherlock, a deep learning model that employs neural networks to analyze various feature sets, including word embeddings, character embeddings, and global statistics derived from individual column values. Extending this work, Zhang et al. [11] developed Sato, which enhances Sherlock by incorporating table context and structured output prediction to more effectively capture the correlations between columns within the same table. In addition, there are tools that make use of pre-trained language models to annotate columns: for instance, Deng et al. [12] developed TURL, a Transformer-based pre-training framework for table understanding tasks, while Suhara et al. [13] developed Doduo, a multi-task learning framework designed to take the entire table as input and uses a single model to predict column types and relationships.

Although these tools have achieved high performance on state-of-the-art benchmarks, their architecture makes them computationally expensive. Furthermore, these tools only enable the recognition of entities and relationships within a table, without the capability to regenerate a potential PII once identified in order to ensure privacy.

Considering this, we introduce NERPII[1], a Python library that leverages NER methods to effectively identify PII within structured data formats, such as CSV files, and subsequently regenerate them in a privacy-preserving manner. To perform NER, we used Presidio [14], a Microsoft SDK that provides a fast identification for private entities such as credit card numbers, names, locations, phone numbers, financial data and more, and a BERT model (*dslim/bert-base-NER*) [15] to identify additional entities. The use of the BERT model thus makes the tool more robust, allowing for the expansion of the entities recognized by the library. Moreover, in order to generate synthetic PII, we use Faker [16], a Python library created to generated fake PII. The

---

[1]The library is available at the following link: https://github.com/Clearbox-AI/nerpii

idea is to assign an entity to each column which contains PII in a dataset, and then replace each values in that column with a coherent fake PII.

In this paper, we describe the architecture of NERPII and we show a short tutorial on how to use the library to extract entity information from a dataset containing personal information and to synthesize new PII that maintains data characteristics while safeguarding privacy. Finally, we discuss conclusion and future directions.

2. NERPII Architecture

The NERPII library consists of two distinct classes: NamedEntityRecognizer and FakerGenerator.

NamedEntityRecognizer is used to perform Named Entity Recognition on structured data, typically in the form of a CSV file. An instance of NamedEntityRecognizer requires several parameters: a Pandas DataFrame containing the data to be subjected to NER, an optional count indicating the desired number of samples to be processed (with a default value of 500), and a NaN filler—a string used to fill NaN values in the dataset (set to '?' by default). Within the NamedEntityRecognizer, a Presidio analyzer is initialized, which attempts to assign a named entity from those supported by the library to each column of the dataset, and the BERT model that tries to assign the organization entity to the columns. To assign an entity to each column, the analyzer and the model first assign an entity to each value within the column. The entity ultimately assigned to the column is the one that has been most frequently assigned to the values contained within that column. When an instance of NamedEntityRecognizer is created, the NER is performed on the dataset, resulting in a dictionary accessible through the instance's attribute *dict_global_entities*. Within this dictionary, for each column, the assigned entity and the confidence score with which that entity has been assigned are recorded.

FakerGenerator is used to generate synthetic PII. The parameters of a FakerGenerator object include the dataset previously analyzed by the NamedEntityRecognizer and the associated *dict_global_entities* dictionary. The generator divides columns for which the recognizer was able to assign an entity from those without an associated entity. For those columns with an associated entity, the generator replaces each value within the column with synthetic data. The FakerGenerator can regenerate the following Named Entities: address, phone number, email address, first and last name, city, state, URL, zipcode, credit card number, Social Security Number (SSN), and country.

Consequently, the output of the generator is a new dataset in which the originally present PII has been substituted with synthetic PII.

3. How to use NERPII

Suppose you have a dataset containing personal information of several people, such as first name, last name, phone number, address, e-mail, etc., like the one shown in Table 1. You need to anonymize the PII contained in the dataset, so firstly, you have to install the library by cloning the github repository or by simply installing it via pip.

```
pip install nerpii
```

Once you have installed the library, you can import the class NamedEntityRecognition by using the following line of code.

```python
from nerpii.named_entity_recognizer import NamedEntityRecognizer
```

Then, you can create a recognizer passing as parameter the path to the CSV file that contained your dataset or directly your dataset in Pandas DataFrame format.

```python
recognizer = NamedEntityRecognizer('./csv_path.csv')
```

Once you have created your recognizer, you can performed NER using the following functions.

```python
recognizer.assign_entities_with_presidio()
recognizer.assign_entities_manually()
recognizer.assign_organization_entity_with_model()
```

These functions assign an entity to most of the columns. The final output is a dictionary, like the one shown below, accessible with

```python
recognizer.dict_global_entities
```

in which column names are given as keys and assigned entities and a confidence score as values.

```python
{'first name': {'entity': 'PERSON', 'confidence_score': 0.9127725856697819},
 'last name': {'entity': 'PERSON', 'confidence_score': 0.8625},
 'address': {'entity': 'ADDRESS', 'confidence_score': 0.8926174496644296},
 'city': {'entity': 'LOCATION', 'confidence_score': 0.8731343283582089},
 'state': {'entity': 'LOCATION', 'confidence_score': 0.976},
 'zip': {'entity': 'ZIPCODE', 'confidence_score': 1.0},
 'phone': {'entity': 'PHONE_NUMBER', 'confidence_score': 0.888},
 'email': {'entity': 'EMAIL_ADDRESS', 'confidence_score': 1.0}}
```

After performing NER on your dataset, you can generate new PII using Faker. You can import the class FakerGenerator by using the following command.

```python
from nerpii.faker_generator import FakerGenerator
```

Then, you can create a generator as follows.

```python
generator = FakerGenerator(dataset, recognizer.dict_global_entities)
```

Finally, to generate new PII you can run this command line.

```python
generator.get_faker_generation()
```

At the end of the whole process you will have obtained a dataset, identical to the original (see Table 2), where the values in the various columns will have been replaced with synthetic PII.

Overall, the library has been tested on two openly available and two proprietary data sets. The Classic Models data set represents a Customer Relationship Management database [2], while the AWS Honeypot data set [3] comes from the cybersecurity domain. The proprietary data sets were a synthetic table depicting a financial fraud detection use case and a table containing user data collected by an IT department. The following table contains an overview of the metrics achieved on the aforementioned tests.

---

[2]Classic Models data set: https://relational.fit.cvut.cz/dataset/ClassicModels
[3]AWS Honeypot data set: https://datadrivensecurity.info/blog/pages/dds-dataset-collection.html

first name	last name	address	city	state	zip	phone	email
James	Butt	6649 N Blue Gum St	New Orleans	LA	70116	504-621-8927	jbutt@gmail.com
Josephine	Darakjy	4 B Blue Ridge Blvd	Brighton	MI	48116	810-292-9388	josephine_darakjy@darakjy.org
Art	Venere	8 W Cerritos Ave #54	Bridgeport	NJ	8014	856-636-8749	art@venere.org
Lenna	Paprocki	639 Main St	Anchorage	AK	99501	907-385-4412	lpaprocki@hotmail.com
Donette	Foller	34 Center St	Hamilton	OH	45011	513-570-1893	donette.foller@cox.net

Table 1

The original dataset containing personal information that need to be synthesized. (https://www.briandunning.com/sample-data/)

first name	last name	address	city	state	zip	phone	email
Jon	Rogers	34980 Johnson Island Apt. 135	Lake Darrell	WY	68167	468.314.5065x909	tonikelly@yahoo.com
Elizabeth	Warner	6039 Beth Coves	North Oliviatown	NC	50586	+1-470-728-1129x641	emma88@gmail.com
Daniel	Watson	23613 Taylor Circles	Amberport	KS	92801	(302)182-0322x122	melissa50@hotmail.com
Joseph	Perry	77488 Miller Field	Patriciamouth	NJ	87509	7614314889	kimberlyedwards@hotmail.com
Simone	Black	85429 Walker Pines Apt. 122	New Heidi	SC	04704	(594)629-1757	nramirez@yahoo.com

Table 2

The synthetized dataset.

Database	Precision (%)	Recall (%)	F1-Score (%)
Classic Models	100.	93.	97.
AWS Honeypot	89.	100.	93.
Sysadmin	100.	93.	97.
Synthetic Data	100.	100.	100.

Table 3

Performance of the NERPII Library on Different Data Sets

4. Conclusion and Future Directions

The evolution of AI has undeniably transformed numerous sectors, contributing to the creation of several new technologies and showing rapid progress. However, this transformation has not been without its ethical implications, particularly regarding the exposure of sensitive personal data. The synthesis of data coupled with NLP techniques, offers a promising solution that ensure the protection of PII.

NERPII, as detailed in this paper, tries to combine the remarkable potential of AI with the importance of data privacy. By employing NER methods and harnessing the power of synthetic data generation, the library adeptly identifies and categorizes PII within structured data formats. The integration of Presidio and a BERT model in the identification process, along with the usage of the Faker library for synthetic data generation, demonstrates a comprehensive and innovative approach.

The strength of NERPII lies in its ability to work with structured data, which is often overlooked by other solutions that focus more on unstructured data such as texts. Additionally, with NERPII, the privacy of personal data is guaranteed, without sacrificing the semantic component

of the data. Indeed, there are other solutions, like Presidio itself, that allow for the identification of PII in texts and their anonymization using predefined or customizable tags. However, in our view, this approach results in a significant loss of informative content in the data. On the other hand, the use of synthetic data enables complete anonymization without any loss of meaning.

A promising direction for the future development of this library is to adapt it to other languages in addition to English, such as Italian, using Faker providers for the Italian language. Currently, the library is designed to recognize Named Entities in structured data in English and to regenerate PII in the American format (for example, if it needs to regenerate a Social Security Number, it will do so according to the American nine-digit format). Moreover, it could be useful to expand the number of Named Entities recognized by the NamedEntityRecognizer and those regenerated by the FakerGenerator. Finally, a future development could examine the aspect of coherence among different columns: in the current version of the library, data synthesis occurs independently for each column, despite it being evident that some columns are semantically correlated (for example, the column containing the city with the one containing the state).

In conclusion, therefore, NERPII can represent a practical solution for addressing some of the ethical issues related to data privacy in the field of artificial intelligence.

References

[1] E. Bassignana, D. Brunato, M. Polignano, A. Ramponi, Preface to the Seventh Workshop on Natural Language for Artificial Intelligence (NL4AI), in: Proceedings of the Seventh Workshop on Natural Language for Artificial Intelligence (NL4AI 2023) co-located with 22th International Conference of the Italian Association for Artificial Intelligence (AI* IA 2023), 2023.

[2] L. Floridi, M. Taddeo, What is data ethics?, 2016.

[3] A. Narayanan, V. Shmatikov, Myths and fallacies of" personally identifiable information", Communications of the ACM 53 (2010) 24–26.

[4] European Parliament, Council of the European Union, Regulation (EU) 2016/679 of the European Parliament and of the Council, 2016. URL: https://data.europa.eu/eli/reg/2016/679/oj.

[5] Centers for Medicare & Medicaid Services, Online at http://www.cms.hhs.gov/hipaa/, 2009.

[6] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, Neural architectures for named entity recognition, arXiv preprint arXiv:1603.01360 (2016).

[7] D. Nadeau, S. Sekine, A survey of named entity recognition and classification, Lingvisticae Investigationes 30 (2007) 3–26.

[8] K. El Emam, L. Mosquera, R. Hoptroff, Practical synthetic data generation: balancing privacy and the broad availability of data, O'Reilly Media, 2020.

[9] T. E. Raghunathan, Synthetic data, Annual review of statistics and its application 8 (2021) 129–140.

[10] M. Hulsebos, K. Hu, M. Bakker, E. Zgraggen, A. Satyanarayan, T. Kraska, Ç. Demiralp, C. Hidalgo, Sherlock: A deep learning approach to semantic data type detection, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 1500–1508.

[11] D. Zhang, Y. Suhara, J. Li, M. Hulsebos, Ç. Demiralp, W.-C. Tan, Sato: Contextual semantic type detection in tables, arXiv preprint arXiv:1911.06311 (2019).

[12] X. Deng, H. Sun, A. Lees, Y. Wu, C. Yu, Turl: Table understanding through representation learning, ACM SIGMOD Record 51 (2022) 33–40.

[13] Y. Suhara, J. Li, Y. Li, D. Zhang, Ç. Demiralp, C. Chen, W.-C. Tan, Annotating columns with pre-trained language models, in: Proceedings of the 2022 International Conference on Management of Data, 2022, pp. 1493–1503.

[14] O. Mendels, C. Peled, N. Vaisman Levy, T. Rosenthal, L. Lahiani, et al., Microsoft Presidio: Context aware, pluggable and customizable pii anonymization service for text and images, 2018. URL: https://microsoft.github.io/presidio.

[15] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[16] D. Faraglia, Other Contributors, Faker, 2014. URL: https://github.com/joke2k/faker.

[17] P. M. Schwartz, D. J. Solove, The pii problem: Privacy and a new concept of personally identifiable information, NYUL rev. 86 (2011) 1814.