Identifying Personal Identifiable Information (PII) in Student Essays

Manu Achar | Pratik Chaudhari | Cody Ledford | Advised by Dr. Anca Doloc-Mihu | Georgia Gwinnett College

INTRODUCTION

The personally identifiable information (PII) detection competition is hosted by The Learning Agency Lab on Kaggle, and by the request of our client, Dr. Gunay, we aim to detect sensitive PII, like names or emails, in student writing. This is necessary to screen and clean educational data so that when released to the public for analysis and archival, the students' privacy risks are mitigated.

OBJECTIVES

- Understand the distribution of PII in student essays
- Detect PII
- Report showing its location

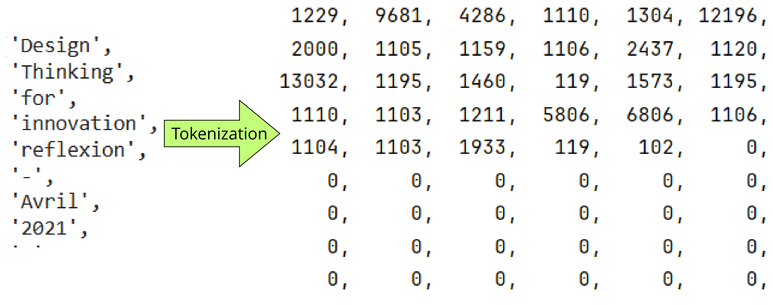
TECHNOLOGIES

Taipy, PyTorch, Pre-trained BERT, Jupyter Notebook

RESULT 1: PREPROCESSING

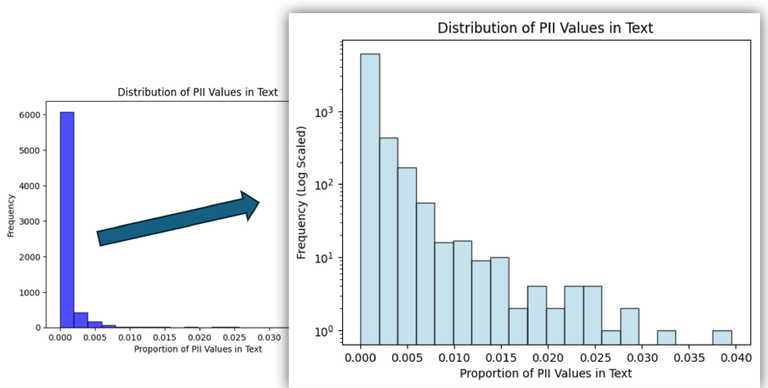
Transforming data to BERT-friendly structure

For Example:

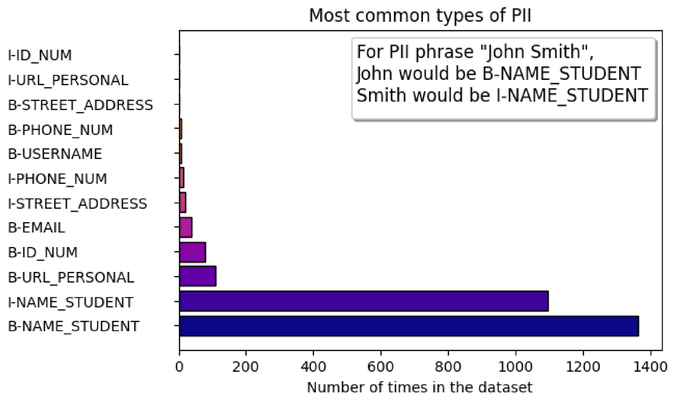


RESULT 2: STATISTICS

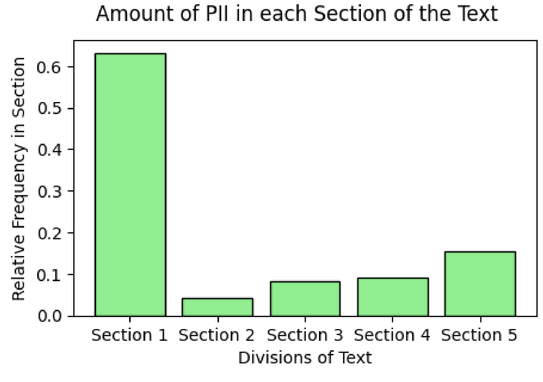
Distribution of PII Values in Text



Most common type of PII are names

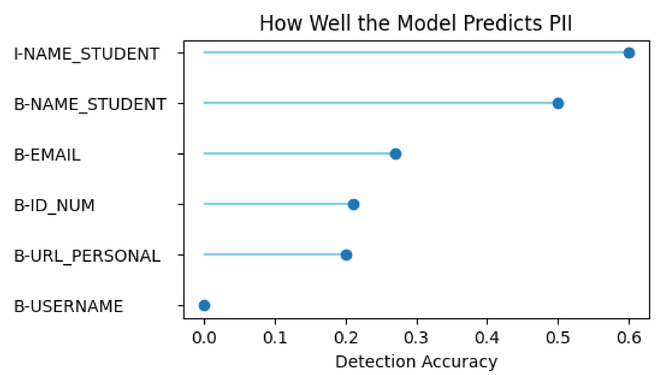


Most PII occurs in the first 1/5th of the essay



RESULTS 3: MODEL

- Token Classification by fine-tuning BERT model
- 5000 Training essays, 1000 Testing essays
- Model Accuracy: 56%, still improving



RESULTS 4: REPOSITORY

Work in progress

Project Github

Web App

CONCLUSION

- PII is likely to exist (p-value < 0.0001)
- Names are most likely to appear as PII in an essay (90%)
- PII appears most often in the 1st fifth of a text

Future work:

- Re-balancing, Name Classes, Truncation

RELATED LINKS

BERT Paper arXiv:1810.04805, taipy.io