The Learning Agency Lab - PII Data Detection.

Aditya Peer 2020355, Ayush Gupta MT23027, Sonu Kumar MT23144

Problem Statement

Identifying and categorizing Personally Identifiable Information(PII) from different essays: Essays usually contain information that could reveal the author's identity. This results in a breach of privacy, and it becomes challenging and costly to provide these essays as public educational datasets. Hence, making an ML model that can detect and classify such information would be highly useful.

We see the problem as threefold: we have to identify the named entities in the text, identify whether they are PIIs, and then classify them.

Data

We are provided two JSON files for test and training data. The training set contains 6807 essays, and the test set contains only 10 essays. The JSON objects have the following keys:

- Document ID
- Full essay
- Token list
- Whitespace list
- Labels(only training data)

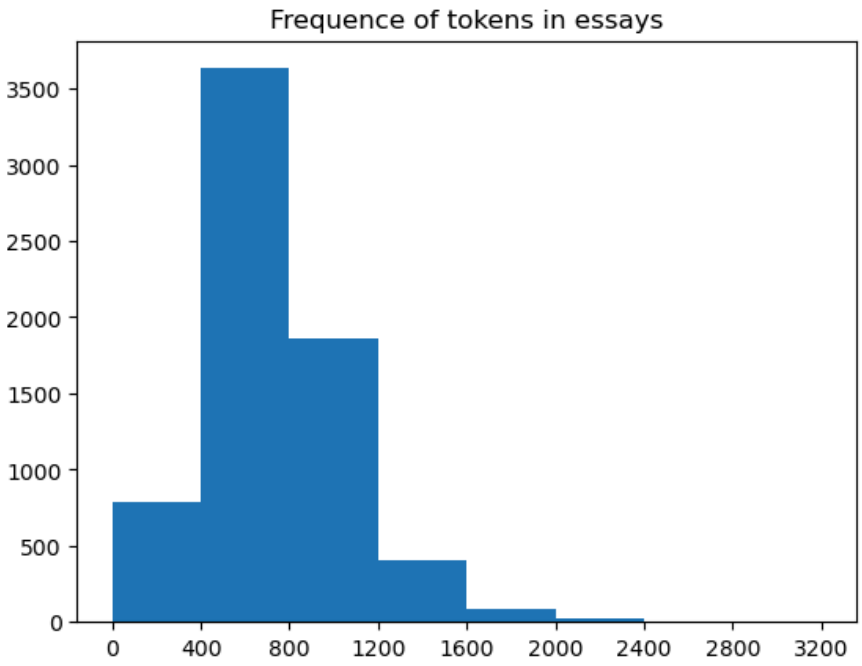
The labels mark each token in the BIO format: ‘B’ is used when an entity stars, ‘I’ is used for an intermediate entity, and ‘O’ is used for non-PII information. PII can be classified into the following classes:-

1. NAME_STUDENT - Name of the student, which may not necessarily be the author
2. EMAIL - Personal email address of student
3. USERNAME - A student’s username on any platform
4. ID_NUM - A number or alpha-numeric sequence of characters that could be used to identify the student, such as a student ID
5. PHONE_NUM - A phone number associated with the student
6. URL_PERSONAL - A URL that might be used to identify a student
7. STREET_ADDRESS - A full or partial street address that is related to the student, such as their home address

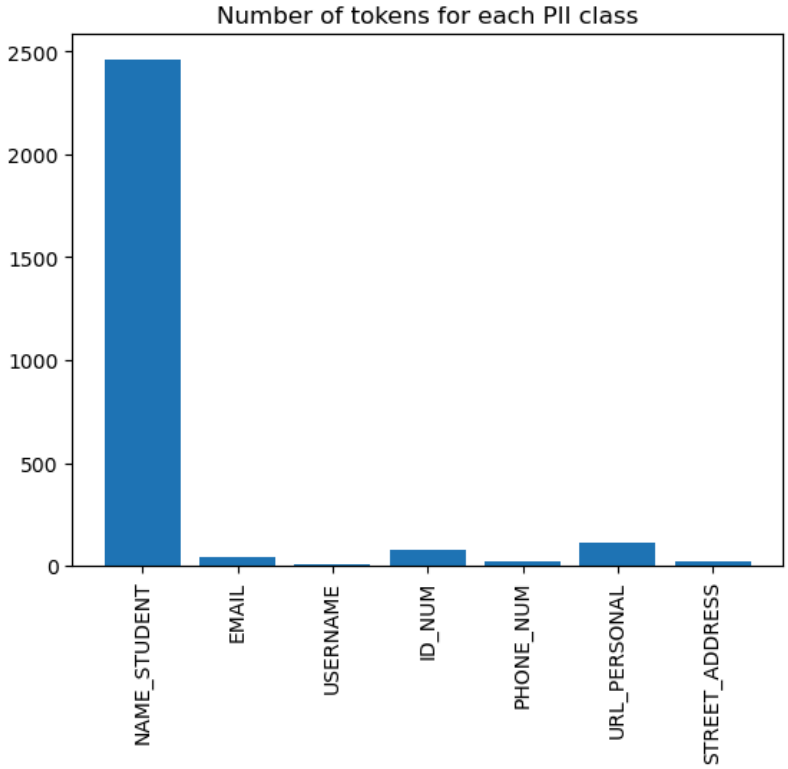
EDA

Since we are dealing with textual data, we have to perform EDA for the tokens, which are basic syntactical units of the language. Tokens exclude whitespace, control characters, etc. First, we analyzed the data on an essay level; we found out the frequency distribution of tokens in the

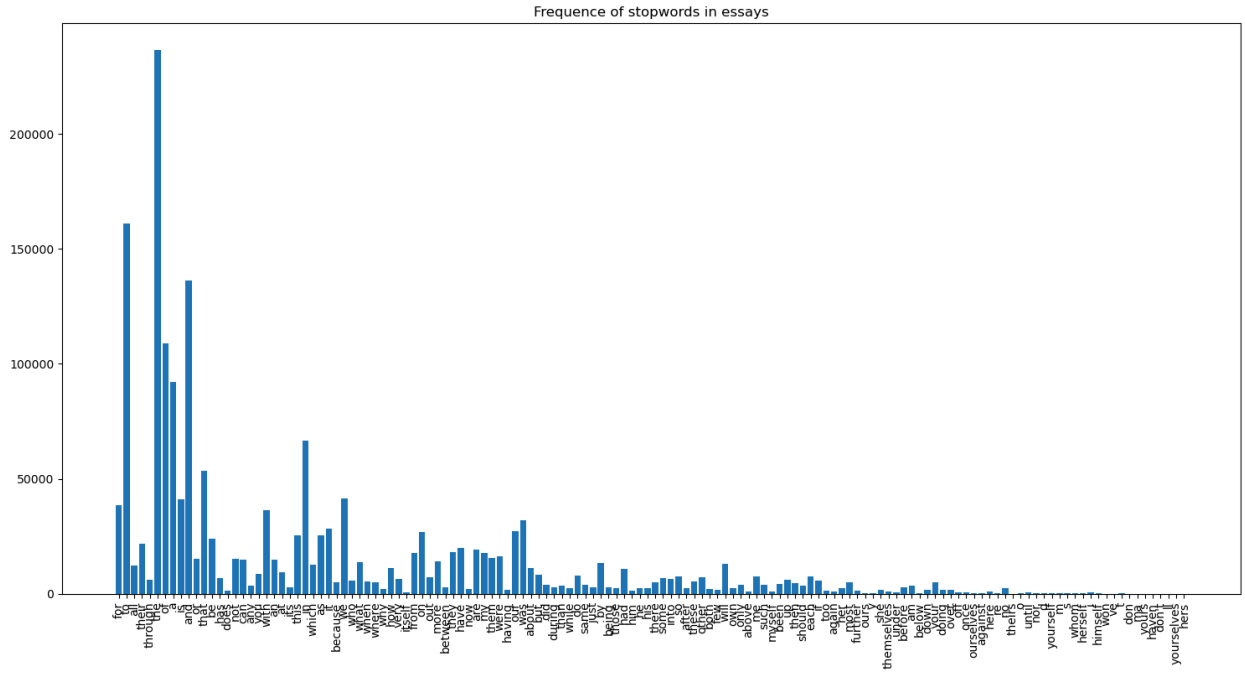
essay. More than 3500 essays were found to have tokens in the 400-800 range, meaning most of them are short.



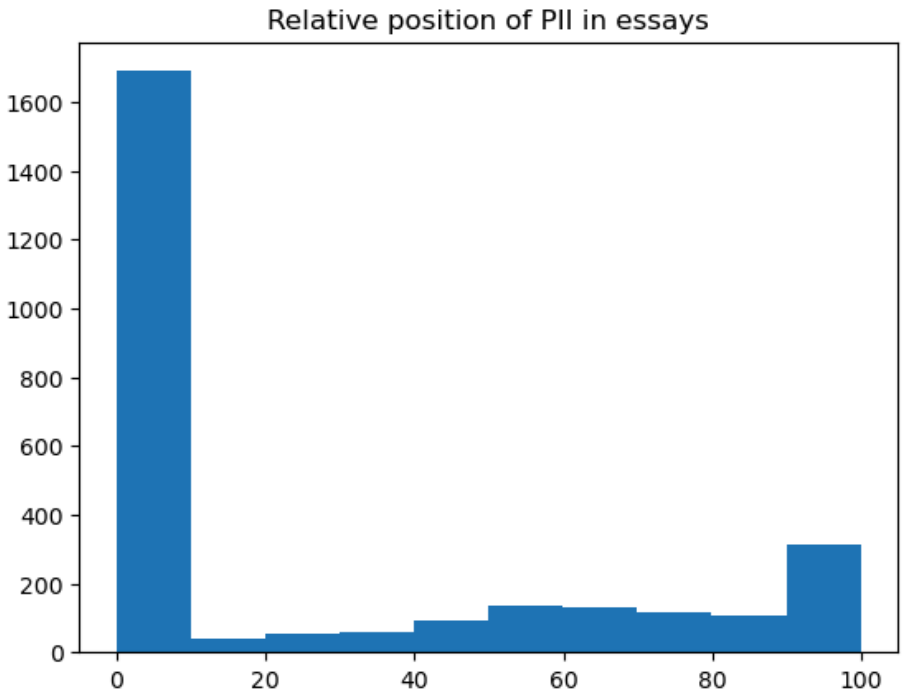
Then, we checked class imbalance in training data. We found out that there is a high-class imbalance with most PII tokens belonging to the NAME_STUDENT class.



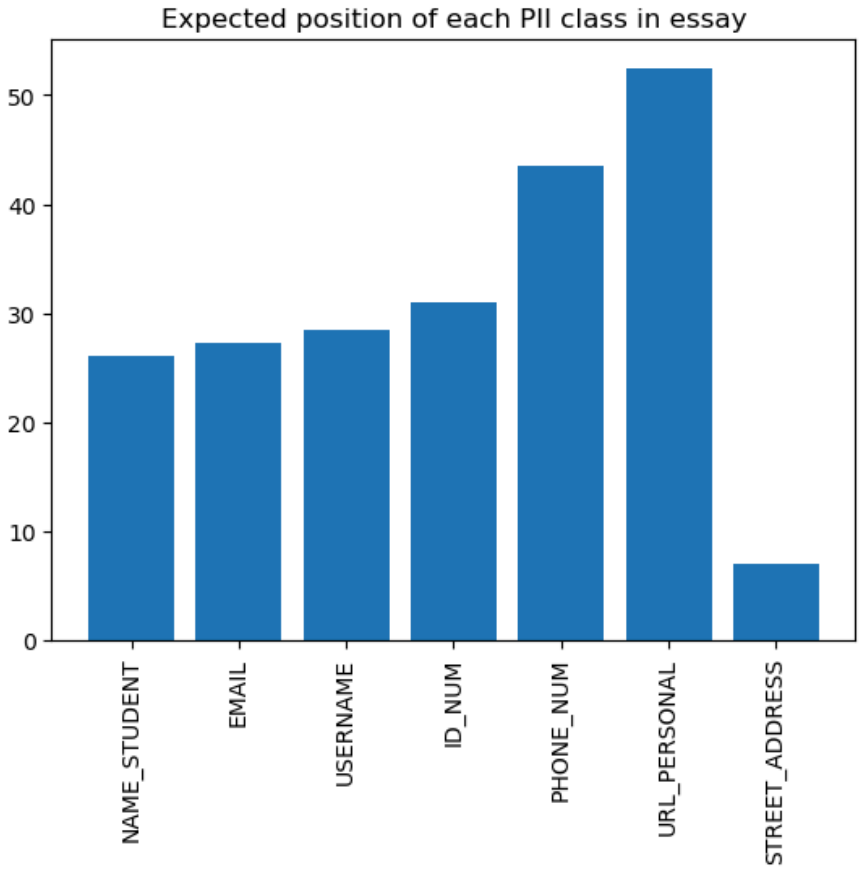
Then, we used a list of stop words(words that are usually not useful for our model) to plot a frequency distribution for those stop words.



We looked at the relative position of each PII token in its respective essay. The position is calculated as a percentile.



This tells us that most PII tokens are found at the start of the essay, with a noticeable uptick at the end. This can help us differentiate between entities and the PII. Then, we look at the expected position of each PII class:



The STREET_ADDRESS class is usually found at the beginning, and the URL_PERSONAL is usually found at the halfway point.

Literature Review

Using NLP and Machine Learning to Detect Data Privacy Violation: Paulo Silva, Carolina Goncalves, Carolina Godinho, Nuno Antunes, Marilia Curado
Year: 2020 | Conference Paper | Publisher: IEEE
https://ieeexplore.ieee.org/xpl/conhome/9146478/proceeding

The paper begins by explaining what PII and NER are. Personal Identifiable Information(PII) is considered to identify or can be used to identify, contact, or locate a person. It is usually identified using Named Entity Recognition(NER). It includes sensitive information like phone numbers, addresses, names, etc. The detection and classification of PII in textual data become critical to safeguard individual privacy and prevent unauthorized access or misuse of personal information.

Named Entity Recognition(NER) is a fundamental task in natural language processing that involves identifying and categorizing named entities in text. NER in various domains, including healthcare, finance, and social media. The specific task of identifying PII entities within NER systems requires additional considerations, as PII can encompass a wide range of information beyond standard named entities.

PII Detection

Detecting PII involves identifying patterns or structures in the text that indicate the presence of personal information. Traditional approaches to PII detection often rely on rule-based systems or regular expressions to match predefined patterns such as email addresses, phone numbers, social security numbers, and credit card numbers

PII Classification

PII entities are identified in the text; the next step is to classify them into different categories based on their sensitivity and potential impact on privacy. Classification tasks may include determining whether an entity is a name, address, email, or other type of PII and assigning appropriate labels or tags accordingly.

The paper aims to do the following:-

1) Evaluation of NLP tools' performance with general purpose and multi-dimensional data sets;
2) Manual labeling (gold standard) of publicly available datasets with entities such as names, addresses, employment, organizations, and others;
3) Analysis of NLP tools' performance on correctly retrieving entities classified as PII on publicly available data;
4) Presentation of proof of concept NER models for PII monitoring and discussion of its applicability as an Enhancing Technology.

Conclusion

In The Learning Agency Lab - PII Data Detection problem, we use three well-known natural language processing tools (NLTK, Standard cost NLP and SpaCy). First, we assess the effectiveness of the tools with a generic dataset. Then, machine learning models are trained and evaluated with datasets built on data that contain personally identifiable information

The results show the model's performance was highly positive in accurately classifying generic and more context-specific data. We observe the relationship between the datasets' training size

and respective performance and estimate the appropriate size for model training within this context.

Plan to solve

Since our problem involves recognizing named entities, identifying PIIs out of them, and then classifying them, we will begin by first recognizing the named entities using NER(Named Entity Recognition) models. Libraries like ‘NLTK’ provide native solutions to NER and can be used as a starting point. After recognizing named entities, we can use features like the position of the token, the length of the token, and the surrounding tokens to identify if the NE is a PII or not. After that, we would have to classify the PII, which could be done by mapping the value provided by our NER module, using techniques like regular expression, or making a model specifically to train it.