Identifying PII in Educational Datasets

Team: Sydney Lister, Lucy Herr, Torrey Trahanovsky

Class: W266 - Natalie (Wednesday)

Term: Spring 2024

Berkeley SCHOOL OF INFORMATION

Overview

- **Motivating Question:** What is the best solution for identifying personally identifiable information (PII) in student essays?
  - Student data often requires manual labelling
  - Potential for rich data sets collected from educational data
  - **Do state-of-the-art techniques sufficiently identify student PII?**

Background

- Named entity recognition (NER) well-researched field
  - Transformers models standard
    - SpaCy provides open-source library for NER

- Identifying *student* PII rather than PII generally
  - Ex: “In 1865, Abraham Lincoln…” vs. “My friend Natalie…”

Dataset

- ~7k student essays written by students enrolled in a massively open online course
  - Responding to one single assignment prompt

- Original PII replaced by surrogate identifiers using partially automated process.

- Target PII types: student names and usernames, B-I-O format

Challenges

- Text classification —> Token classification
  - Handling output dimension of:

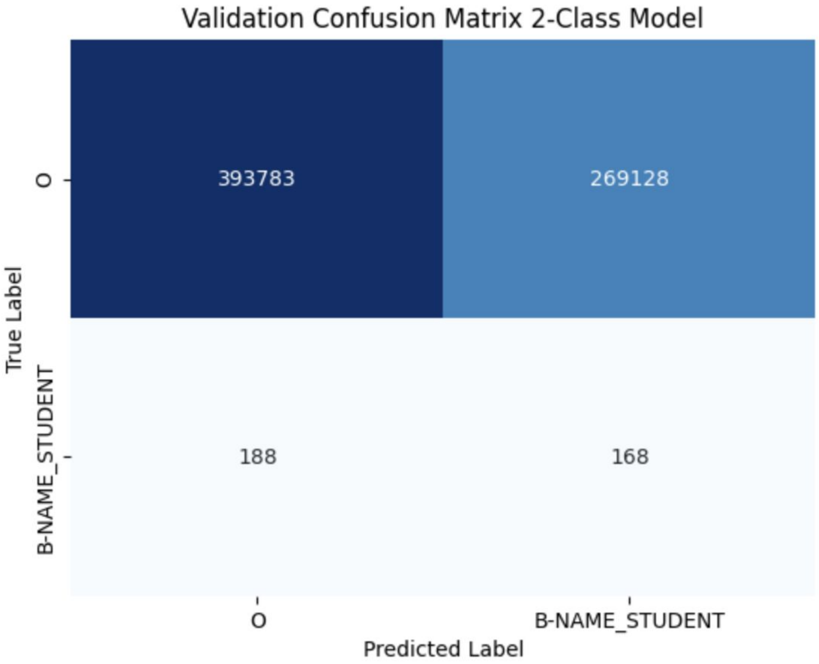
    (batch size, sequence length, number of classes)
  - Aligning labels with valid text to account for transformers special tokens and padding

- Implementation of weighted loss function to account for class imbalance

Models and Results

Weighted F1 Scores of Evaluated Models

- SpaCy: 0.997
- BERT: 0.24
- DeBERTa:
  - Pre-trained model and continue pre-trained model: 0.97
  - Fine-tuned: 0.12
  - Fine-tuned two-class (binary) model: 0.74


Validation Confusion Matrix 2-Class Model

Conclusions

- NLP systems show great performance detecting student PII
  - SpaCy shows superior performance
  - Pre-trained deBERTa also good performance

- Future directions:
  - Address class imbalances:
    - Weighted loss function
    - Incorporate more PII data
  - Evaluate performance of bidirectional LSTM