RUTGERS

# PII Detection & Removal From Educational Texts

Master's Project
Lovedeep Singh

Under the guidance of Dr. Desmond Lun

# Problem Statement

Develop automated techniques to detect and remove PII from educational data.

Problem:

- PII is barrier to analyze and create open datasets for educational advancement.
- Hinders anonymous reviews / peer feedback.
- Currently, most reliable method is Manual review and cleaning.

How this solution and tool will help?

- Identify, detect and remove PII from educational texts and documents.
- Safeguarding Student Privacy

# Project Work Done

1. Data Collection and Augmentation
2. Data Analysis and Visualization
3. Labelling, Text Tokenization and Encoding
4. Evaluated Different Classifiers - Logistic Regression, LSTM (RNN)
5. Building / Evaluating using Transformers - DeBERT based Token Classification
6. Model Validation
7. Inference

# Datasets

- Kaggle Dataset - 22,000 essays written by students enrolled from massively open online course.
  - All essays written in response to a single assignment prompt
  - JSON format
  - Documents tokenized using the SpaCy English tokenizer.
  - Structure:
    - **(int):** the index of the essay
    - **document** (int): an integer ID of the essay
    - **full_text** (string): a UTF-8 representation of the essay
    - **tokens** (list)
      - (string): a string representation of each token
    - **trailing_whitespace** (list)
      - (bool): a boolean value indicating whether each token is followed by whitespace.
    - **labels** (list)
      - (string): a token label in BIO (Beginning, Inner, Outer) format
        - PII type is prefixed with “B-” when it is the beginning of an entity
        - “I-” if the token is a continuation of an entity
        - “O” for tokens that are not PII

| BIO encoding | Michel | Jordan | would | choose | Bush |
|---|---|---|---|---|---|
| | B-PER | I-PER | O | O | B-PER |

# Dataset

```
EMAIL: 24
ID_NUM: 33
NAME_STUDENT: 891
PHONE_NUM: 4
STREET_ADDRESS: 2
URL_PERSONAL: 72
USERNAME: 5
OTHER: 5862
```

| | document | full_text | tokens | trailing_whitespace | labels |
|---|---|---|---|---|---|
| **0** | 7 | Design Thinking for innovation reflexion-Avril... | [Design, Thinking, for, innovation, reflexion,... | [True, True, True, True, False, False, True, F... | [O, O, O, O, O, O, O, O, O, B-NAME_STUDENT, I-... |
| **1** | 10 | Diego Estrada\n\nDesign Thinking Assignment\n\... | [Diego, Estrada, \n\n, Design, Thinking, Assig... | [True, False, False, True, True, False, False,... | [B-NAME_STUDENT, I-NAME_STUDENT, O, O, O, O, O... |
| **2** | 16 | Reporting process\n\nby Gilberto Gamboa\n\nCha... | [Reporting, process, \n\n, by, Gilberto, Gambo... | [True, False, False, True, True, False, False,... | [O, O, O, O, B-NAME_STUDENT, I-NAME_STUDENT, O... |
| **3** | 20 | Design Thinking for Innovation\n\nSindy Samaca... | [Design, Thinking, for, Innovation, \n\n, Sind... | [True, True, True, False, False, True, False, ... | [O, O, O, O, O, B-NAME_STUDENT, I-NAME_STUDENT... |
| **4** | 56 | Assignment:  Visualization Reflection Submitt... | [Assignment, :,  , Visualization,  , Reflecti... | [False, False, False, False, False, False, Fal... | [O, O, O, O, O, O, O, O, O, O, O, B-NAME_ST... |
| **...** | ... | ... | ... | ... | ... |
| **9328** | ffacbfa2-fd35-4ac1-b975-a2cc41f56544 | During my 15-year career as a developer, I hav... | [During, my, 15, -, year, career, as, a, devel... | [True, True, False, False, True, True, True, T... | [O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, ... |
| **9329** | ffdecdd0-cd86-4281-9271-1dfe26d4a790 | In 2019, a particularly challenging case came ... | [In, 2019, ,, a, particularly, challenging, ca... | [True, False, True, True, True, True, True, Tr... | [O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, ... |
| **9330** | ffdf4428-e76f-4465-9fba-60d5685096f7 | Hello there, I'm Krishna Lopez and I've been a... | [Hello, there, ,, I, 'm, Krishna, Lopez, and, ... | [True, False, True, False, True, True, True, T... | [O, O, O, O, O, B-NAME_STUDENT, I-NAME_STUDENT... |
| **9331** | ffdfbd65-978c-49e8-88ce-b726ce95b26b | As a designer, I've encountered various challe... | [As, a, designer, ,, I, 've, encountered, vari... | [True, True, False, True, False, True, True, T... | [O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, ... |
| **9332** | fff794fe-d3f7-4452-b704-2714c24337ef | Hello, my name is Hiroko Yu and I am a freelan... | [Hello, ,, my, name, is, Hiroko, Yu, and, I, a... | [False, True, True, True, True, True, True, Tr... | [O, O, O, O, O, B-NAME_STUDENT, I-NAME_STUDENT... |

9333 rows × 5 columns

# Sample Data

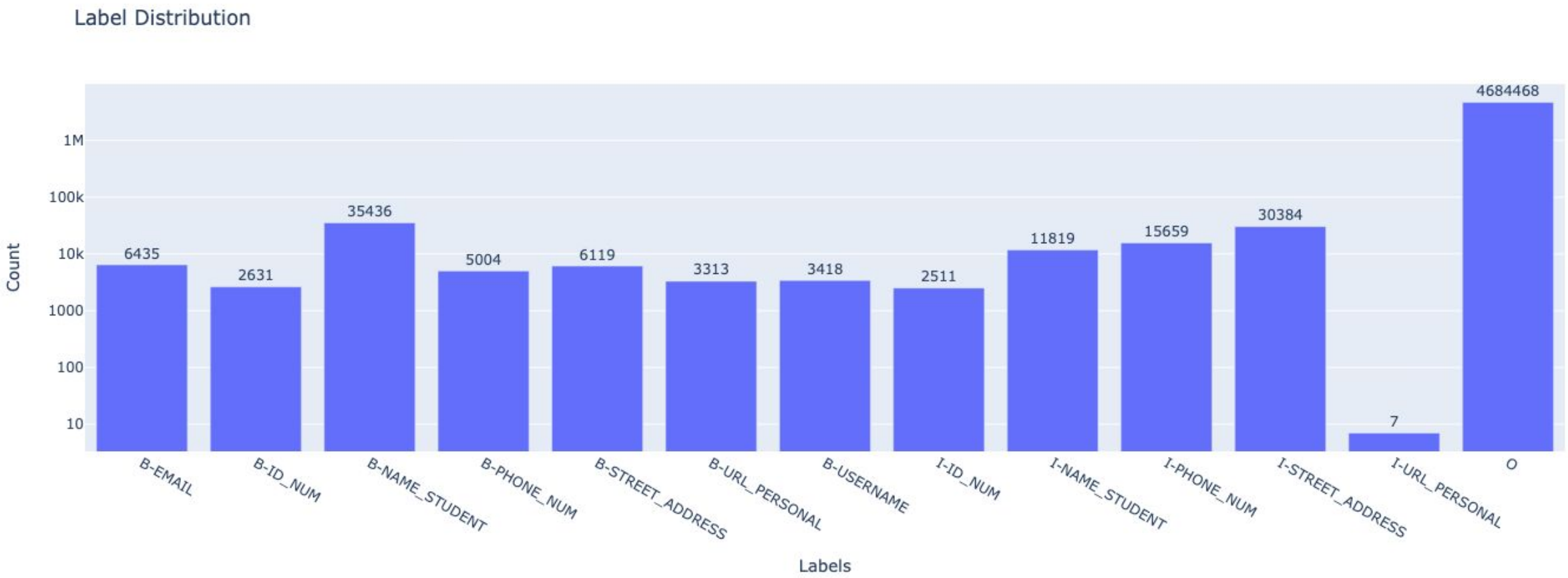
Waseem B-NAME_STUDENT Mabunda I-NAME_STUDENT 591 B-STREET_ADDRESS Smith I-STREET_ADDRESS Centers I-STREET_ADDRESS Apt I-STREET_ADDRESS . I-STREET_ADDRESS 656 I-STREET_ADDRESS I-STREET_ADDRESS
Joshuamouth I-STREET_ADDRESS , I-STREET_ADDRESS RI I-STREET_ADDRESS 95963 I-STREET_ADDRESS ( The Netherlands) 410.526.1667 B-PHONE_NUM vpi@mn.nl

Mind Mapping, Challenge: For several years I have been working for an Asset manager in the Netherlands. During this period I have been involved in many projects. Certainly in the world of asset management, much has changed in recent years in the area of Law and Regulations. What I mainly experience in these projects is that all departments have a different interest in starting a new project. This certainly does not benefit the project. How do you get everyone to complete a project in the common interest and how do you motivate everyone who participate in the project? Selection: An improvement project can be approached in different ways. The most common way is the scrum approach. We work in multidisciplinary teams that work in short sprints, with a fixed length of 1 to 4 weeks. Cooperation is very important and everyone must be able to respond quickly to changing circumstances. Scrum is based on the theory of empirical process control, or empiricism. Empiricism assumes that knowledge arises from experience and making decisions based on what is known. I chose mind mapping because I am looking for a way to show the creativity colleagues always have at the start of a project, to keep this up to date and very important to keep it visible. But also with the thoughts to keep colleagues motivated and to show how their creativity contributes to the project. So I want to see if scrum can be combined with Design Thinking and especially with Mind mapping. Application: When starting a new project at work, I checked whether it is workable to combine the scrum approach with Mind Mapping. The central theme was to increase the STP (Straight through processing) rate for a specific product that we trade with an x percentage. As a scrum team, we have tried to provide insight into the various topics related to the 'increase STP rate' via a paper diagram. Each team member could indicate in this diagram his or her creativity which related to increasing the STP rate. After this we went to see if there was a connection between certain ideas. We quickly learned that certain ideas could be combined and that certain steps in the project could be skipped. By combining scrum work and mind mapping, we were able to go live with implementation faster and increase STP speed step by step. By making the project visible through a diagram, colleagues also indicated that this gave them more energy to participate in the project.

D e s i g n T h i n k i n g

Insight: The insight I got to combine scrum with mind mapping (Design thinking) is that if you make everyone's creativity and thinking visible through Mind mapping, you will come sooner to a solid solution to complete a project. The feedback we received is that it also gives more energy to colleagues who have participated in this project. The biggest challenge was to create support for this new way of working. At the beginning of the project, we showed a short video of how mind mapping works. This gave us immediate support from our fellow team members to combine scrum with mind mapping. https://www.youtube.com/watch?v=tIBN9VJ0S4a B-URL_PERSONAL The conclusion is that you definitely can combine scrum and Design thinking. Approach: In terms of approach, I wouldn't be much different from what I did in this project. I only see advantages of combining scrum with mind mapping. As described in the alinia insight, there are only benefits.

# Distribution with more Data



# Tokenization and Labels

```
original_labels = ["B-EMAIL", "B-ID_NUM", "B-NAME_STUDENT", "B-PHONE_NUM",
          "B-STREET_ADDRESS", "B-URL_PERSONAL", "B-USERNAME",
          "I-ID_NUM", "I-NAME_STUDENT", "I-PHONE_NUM",
          "I-STREET_ADDRESS","I-URL_PERSONAL","O"]
```

```
('Nathalie', 'B-NAME_STUDENT')
('Sylla', 'I-NAME_STUDENT')
('Nathalie', 'B-NAME_STUDENT')
('Sylla', 'I-NAME_STUDENT')
('Nathalie', 'B-NAME_STUDENT')
('Sylla', 'I-NAME_STUDENT')
**********
('N', 'B-NAME_STUDENT')
('atha', 'B-NAME_STUDENT')
('lie', 'B-NAME_STUDENT')
('_S', 'I-NAME_STUDENT')
('ylla', 'I-NAME_STUDENT')
('N', 'B-NAME_STUDENT')
('atha', 'B-NAME_STUDENT')
('lie', 'B-NAME_STUDENT')
('_S', 'I-NAME_STUDENT')
('ylla', 'I-NAME_STUDENT')
('N', 'B-NAME_STUDENT')
('atha', 'B-NAME_STUDENT')
('lie', 'B-NAME_STUDENT')
('_S', 'I-NAME_STUDENT')
('ylla', 'I-NAME_STUDENT')
```

# Metrics

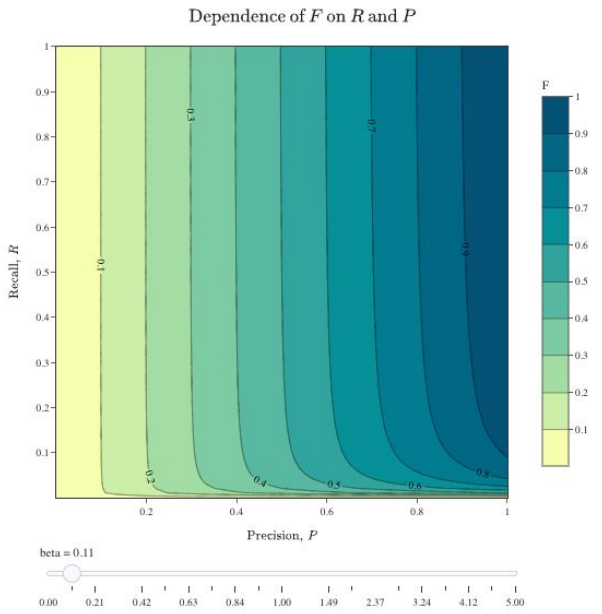
- Metrics:
  - Recall, Precision, F1 score, Accuracy, MCC
- For our use case:

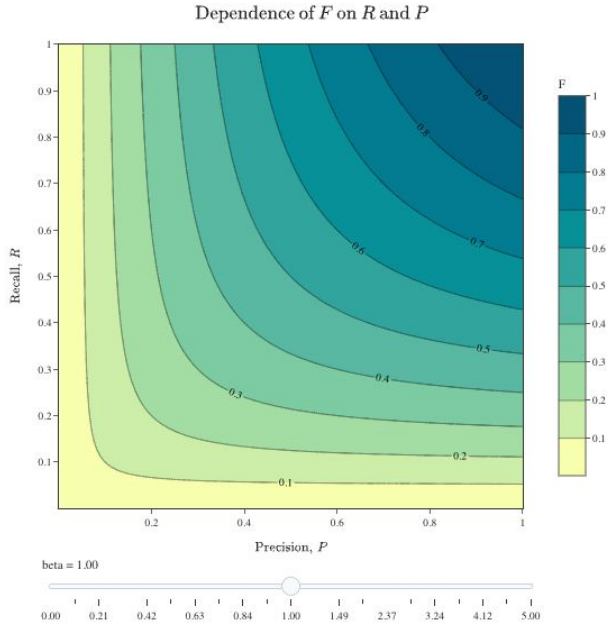
  f1_score = (1 + 5*5) * recall * precision / (5*5*precision + recall)
- Penalizes not identifying True Positives heavily

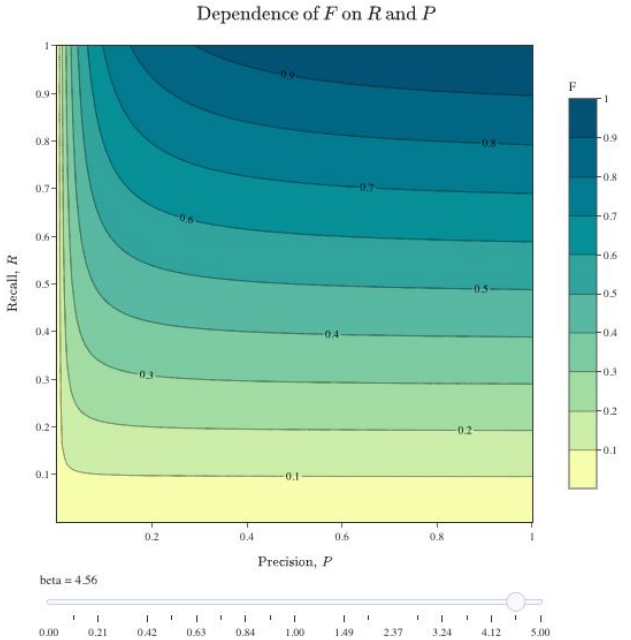
$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

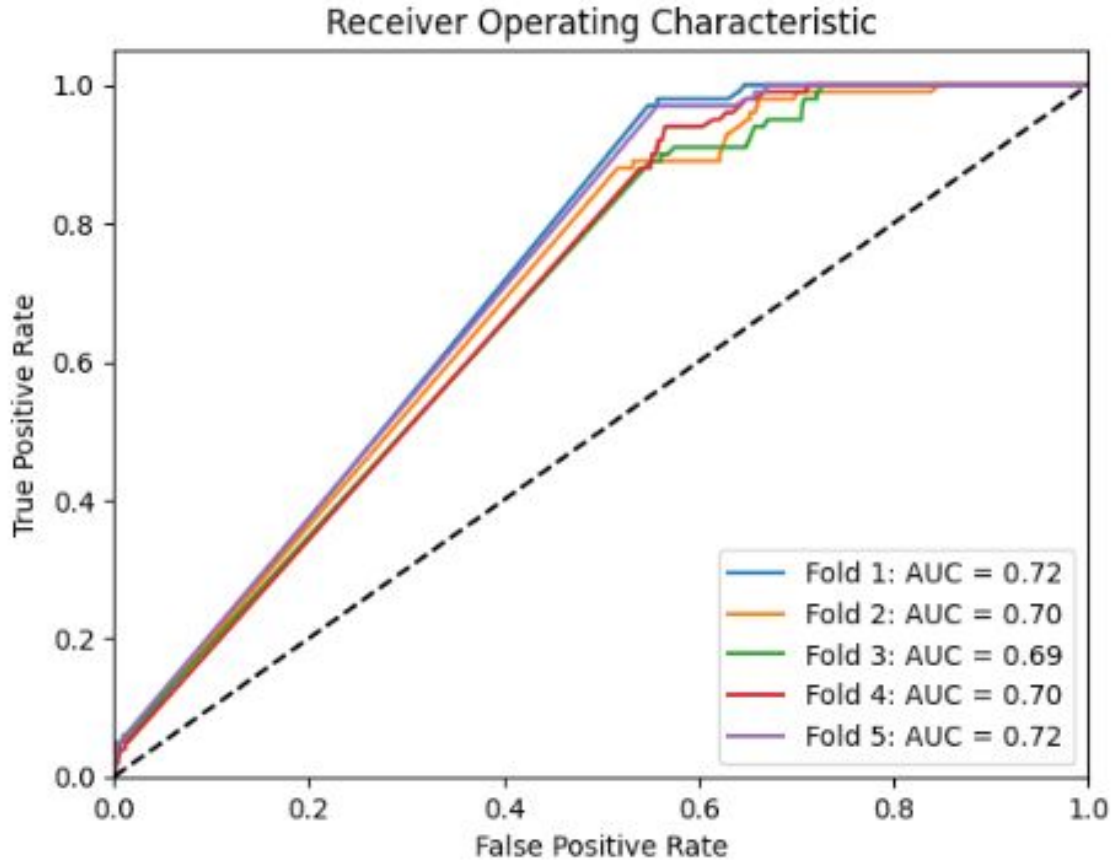
$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$



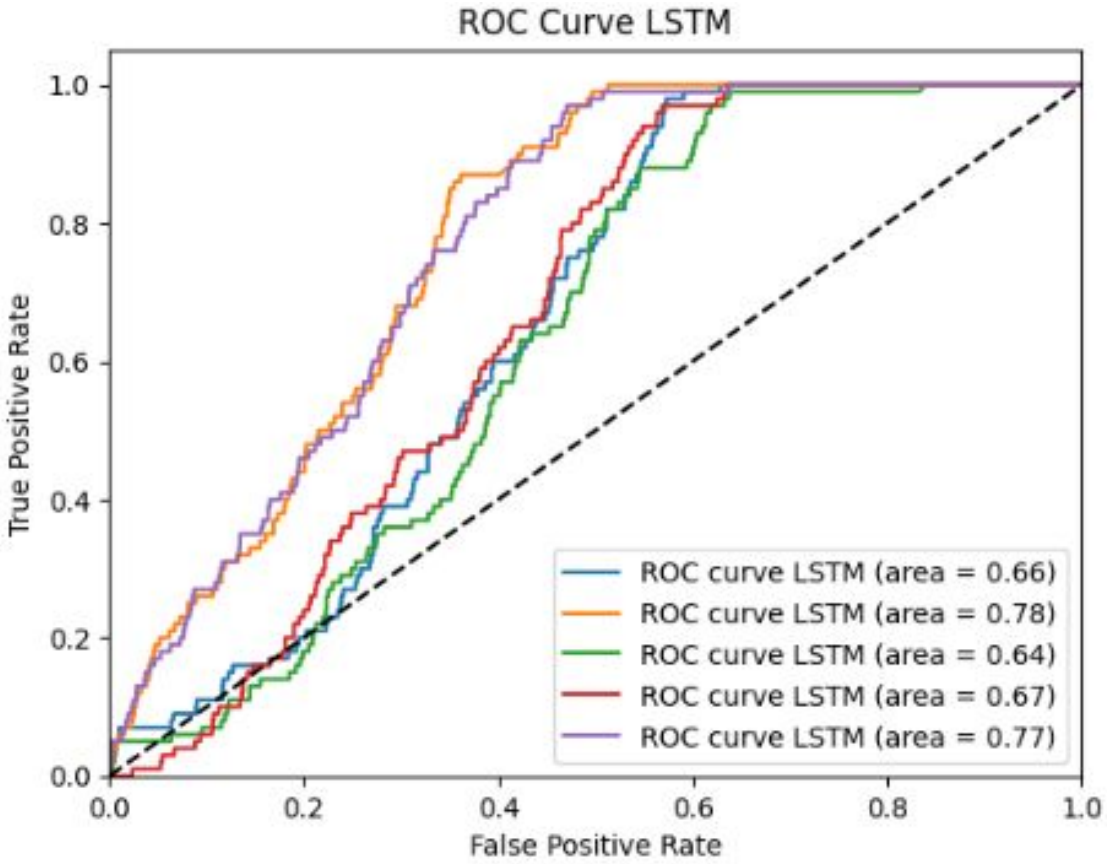



# Classifier - Logistic Regression




```
Average MCC: 0.4502010162994595
Average Recall: 0.39476923076923076
Average Precision: 0.5977135258199183
Average Accuracy: 0.4271663519446496
Average F1 Score: 0.40101993318212187
Average ROC AUC: 0.707427425323621
```

# Classifier - LSTM




```
LSTM Average MCC: 0.42393505412554944
LSTM Average Recall: 0.41353846153846147
LSTM Average Precision: 0.5446901205264025
LSTM Average Accuracy: 0.4474840018232172
LSTM Average F1 Score: 0.412668327978371
LSTM Average ROC AUC: 0.7025574284061181
```

# Transformer - DeBERTa

| Epoch | Training Loss | Validation Loss | Recall | Precision | F1 |
|---|---|---|---|---|---|
| 1 | 0.011100 | 0.008297 | 0.981462 | 0.969037 | 0.980978 |
| 2 | 0.004800 | 0.006080 | 0.985113 | 0.977889 | 0.984833 |
| 3 | 0.003600 | 0.005513 | 0.986301 | 0.979785 | 0.986049 |

```
CPU times: user 34min 38s, sys: 10min 10s, total: 44min 49s
Wall time: 44min 42s
TrainOutput(global_step=2100, training_loss=0.07011932701049817, metrics={'train_runtime': 2682.16, 'train_samples_per_second': 6.262,
'train_steps_per_second': 0.783, 'total_flos': 6363291018572928.0, 'train_loss': 0.07011932701049817, 'epoch': 3.0})
```

RUTGERS

# DEMO

# Key Challenges

- Data Collection and Generation
  - Augmenting with more data
  - Identifying False Positives
- Learned how to configure Cuda, CuDNN, GPU's for running / Training with Language Models
- Training takes a lot of time.
- Could have done a few more iterations to fine tune transformer by changing learning rates.
- More visualization and metrics using Weights & Biases, learn more about ML Ops.

RUTGERS

# THANK YOU