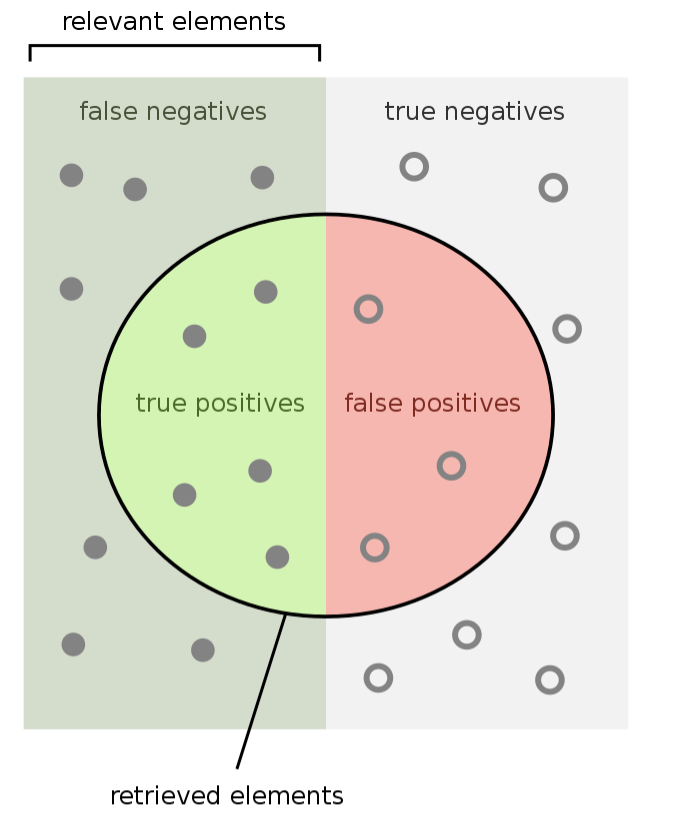
1. Introduction

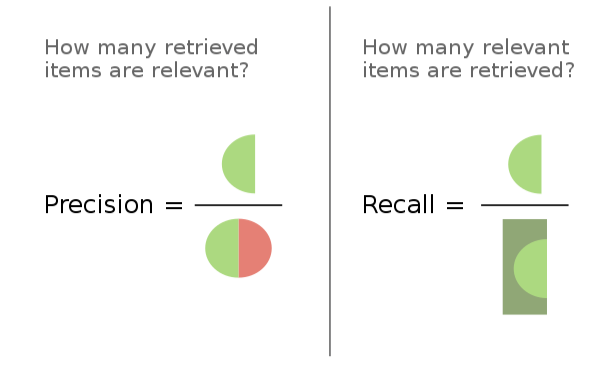
In today’s era of abundant educational data from sources such as ed tech, online learning, and research, widespread PII is a key challenge.

PII’s presence is a barrier to analyze and create open datasets that advance education because releasing the data publicly puts students at risk. To reduce these risks, it’s crucial to screen and cleanse educational data for PII before public release, which data science could streamline.

Manually reviewing the entire dataset for PII is currently the most reliable screening method, but this results in significant costs and restricts the scalability of educational datasets. The cost of releasing educational datasets can be decreased by automating the detection and removal of PII (Personal Identifiable Information).

The goal of this competition is to develop a model that detects PII in student writing. This will further more support learning science research and the development of educational tools

1. Evaluation



Submissions are evaluated on micro 𝐹𝛽, which is a classification metric that assigns value to recall and precision. The value of 𝛽 is set to 5, which means that recall is weighted 5 times more heavily than precision.

1. Submission File

For each document in the test set, you must predict which token values have a positive PII label. You should only include predictions of positive PII label values. Outside labels O should not be included. Each row in the submission should correspond to a single label found at a unique document-token pair. Additionally, the evaluation metric requires a row\_id with an enumeration of predicted labels.

1. Problem Statement

What is the token?

In linguistics, a token is an instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit for processing. For example, in the sentence “The cat sat on the mat”, there are six tokens.

In natural language processing (NLP), the sentence "Hello, world!" might be broken down into 4 tokens like ["Hello", ",", " world", "!"].

Ref: <https://platform.openai.com/tokenizer>

What is the character?

Characters are the basic units of text. Characters are used to construct strings, each character is typically stored in one or more bytes, depending on the encoding standard (ASCII, UTF-8). For example, in the string “Hello, world!”, each of the following is an individual character: H, e, l, l, o, ,w, o, r, l, d and !.

Characters are important in text processing and natural language processing (NLP), as operations such as tokenization, encoding, and decoding all deal with characters and their representations in computers.

What is the named entity recognition (NER) task?

Named Entity Recognition (NER) is a subtask of information extraction that seeks to locate and classify named entities mentioned in unstructured text into pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.

For instance:

"Elon Musk, the CEO of SpaceX, announced that the company's new spacecraft would be ready by 2023. The event took place in Los Angeles on Wednesday. The price is estimated to be around $55 million."

After NER system:

[

("Elon Musk", "PERSON"), # Person's name

("CEO", "TITLE"), # Title or position of a person

("SpaceX", "ORGANIZATION"), # Organization's name

("2023", "DATE"), # Date

("Los Angeles", "LOCATION"), # Location

("Wednesday", "DATE"), # Date

("$55 million", "MONEY") # Monetary value

]

What is the tokenizer? (eg: tokenizer = AutoTokenizer.from\_pretrained(TRAINING\_MODEL\_PATH))

Tokenizers are used to convert text into tokens, which are small pieces of text that the model can work with (like words or subwords). The AutoTokenizer class can automatically detect the correct tokenizer class to use based on the pretrained model we’re using.

What is PII?

PII stands for Personally Identifiable Information. It refers to any data that could potentially identify a specific individual. This includes names, social security numbers, passport number, driver’s licence number, credit card number, street address, email address, fingerprints, student grades, transcripts.

What is padding?

In Machine Learning (ML), “PAD” often refers to a technique used to make sequences of data uniform in length, which is essential for certain types of models that require input data in fixed sizes. (i.e. Natural language processing)

1. Data Analysis

Dataset Description

The competition dataset comprises approximately 22,000 essays written by students enrolled in a massively open online course. All of the essays were written in response to a single assignment prompt, which asked students to apply course material to a real-world problem. The goal of the competition is to annotate personally identifiable information (PII) found within the essays.

PII Types

The competition asks competitors to assign labels to the following seven types of PII:

NAME\_STUDENT - The full or partial name of a student that is not necessarily the author of the essay. This excludes instructors, authors, and other person names.

EMAIL - A student’s email address.

USERNAME - A student's username on any platform.

ID\_NUM - A number or sequence of characters that could be used to identify a student, such as a student ID or a social security number.

PHONE\_NUM - A phone number associated with a student.

URL\_PERSONAL - A URL that might be used to identify a student.

STREET\_ADDRESS - A full or partial street address that is associated with the student, such as their home address.