The Learning Agency Lab - PII Data Detection

I. Introduction

In the modern era of technology, there is a vast amount of educational data available from sources like educational technology platforms, online learning spaces, and academic studies. This abundance of information has great potential to enhance education research and application. However, it also presents a major challenge with regards to protecting Personally Identifiable Information. It is crucial to prioritize safeguarding private student details to prevent potential misuse or harm while upholding individual privacy.

The demand to minimize the potential dangers linked with disclosing educational information in public has led to the creation of effective methods for examining and purifying datasets to identify and eliminate Personally Identifiable Information. Though manual review continues to be the most reliable way of detecting PII, it is laborious and expensive, which hinders the expansion of educational datasets. Automated techniques, especially those using Named Entity Recognition, present possible solutions; however, they frequently encounter challenges when it comes to distinguishing accurately between sensitive names and non-sensitive ones.

This project utilizes modern Natural Language Processing model DistilBERT to improve the effectiveness and precision of PII detection. By harnessing this state-of-the-art technology, goal is to pave the way for generating high-quality public education datasets while preserving privacy.

At the core of initiative is a strong drive to unleash the extensive possibilities of educational data while protecting individuals' privacy rights. By automating PII detection, we aim to grant researchers access to previously unattainable data, empowering them to create innovative tools and interventions that can support both educators and students. Joint endeavors highlight a mutual dedication to utilizing advanced technology for the benefit of education and society at

large. Collectively, we are committed to tackling the issues presented by PII in educational datasets and enabling significant progress in learning-based tools and programs.

II. Literature review and Related work

Detecting and protecting Personally Identifiable Information from extensive, unorganized text collections presents significant challenges in upholding data privacy and security. Kulkarni and Cauvery N[1] have introduced an innovative approach that merges topic modeling with Byte mLSTM, a sequential model, to effectively categorize and identify PII details within email archives. Their study illustrates the efficiency of their method in accurately identifying PII, highlighting its relevance in business settings while emphasizing proactive measures for managing personal information to mitigate privacy risks and prevent data breaches.

Mitra *et al.* examined various machine learning algorithms for the classification of personally identifiable information within text data. The models included Support Vector Machines, Random Forest, Logistic Regression, Long Short-Term Memory, and Multi-Layer Perceptron, all trained on TF-IDF features. Their results emphasized LSTM as the optimal option for efficient PII identification, while suggesting that ensemble learning could enhance both accuracy and efficiency further.

Jaikishan J. *et al*.[3] propose a technique that integrates machine learning and regular expressions to detect and categorize PII in different types of documents. This approach allows organizations to improve privacy protection measures and meet regulatory requirements by applying methods like feature engineering, adaptive learning, and training with substantial datasets.

Paula S [4] propose an integrated method that combines Natural Language Processing methods with machine learning algorithms for identifying personally identifiable information. Their study demonstrates the success of a neural network-based MLP model, which is capable of generating estimated results for new data inputs, leading to effective management of diverse datasets. In comparison, the RF model, utilizing decision trees, may encounter difficulties when presented with unfamiliar inputs.

Payne B *et al.*[5] examines techniques for anonymizing data by evaluating four Named Entity Recognition libraries. The research emphasizes the effectiveness of word embeddings and stacked word embedding methods, but notes that they are time-consuming. In contrast, Conditional Random Fields offer practicality with minimal training time. Although SpaCy shows improvements through re-training, it lacks support for certain entities, affecting its suitability for privacy protection. In a privacy-preserving context, recall is identified as the most crucial metric due to its role in mitigating privacy breaches and underlining the importance of model evaluation.

III. Data exploration

The project's dataset comprises around 22,000 student-authored essays, all in response to a specific assignment prompt within a massively open online course. The data is formatted as JSON and includes fields such as document identifiers, complete essay texts, tokenized words, and token annotations. Important aspects of the data exploration process involve studying the distribution of essay lengths to comprehend the dataset's variability, examining the occurrence of various types of personally identifiable information, and reviewing the tokenization method and labels used. It is also essential to identify any missing values, class imbalances or anomalies in the dataset to ensure its quality and integrity before proceeding with further analysis and model development.

Additionally, the investigation will include evaluating options for incorporating external datasets to enhance the training data, especially because a large part of the essays is set aside for the test set. The phase of data exploration will establish a solid foundation for subsequent project stages such as model training, evaluation, and ultimately creating an effective solution for detecting PII in unstructured text collections.

IV. Exploratory Data Analysis

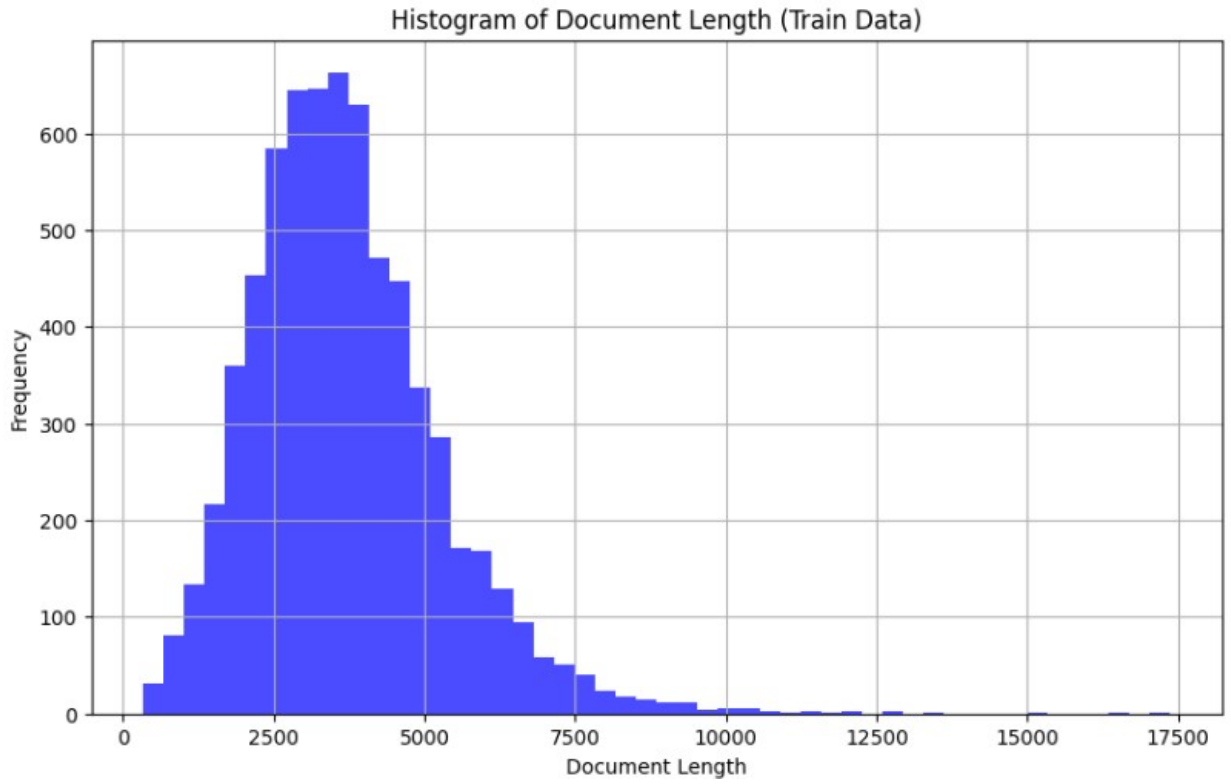
In modern data-driven research, understanding dataset characteristics is essential for developing effective solutions. In educational data analysis, thorough exploration and exploratory data analysis is crucial for building robust models to detect Personally Identifiable Information. This section explores the dataset provided for PII detection in student essays using visualization

techniques to understand its composition and distribution. Detailed exploration aims to reveal patterns, anomalies, and trends in the data for informing model development and evaluation. The goal is to devise strategies for safeguarding student privacy while leveraging educational datasets for research and innovation.

i. Histogram of Document Length (Train Data):

The histogram illustrates how document lengths are distributed in the training dataset, offering valuable information about the variation in essay lengths. On the x-axis, we have document length and on the y-axis, we can see the frequency of essays falling into different length ranges. The histogram demonstrates a diverse range of document lengths, with most essays concentrated around specific intervals. Analyzing this distribution is essential to detect any unusual or atypical cases and to maintain integrity and quality in the dataset.

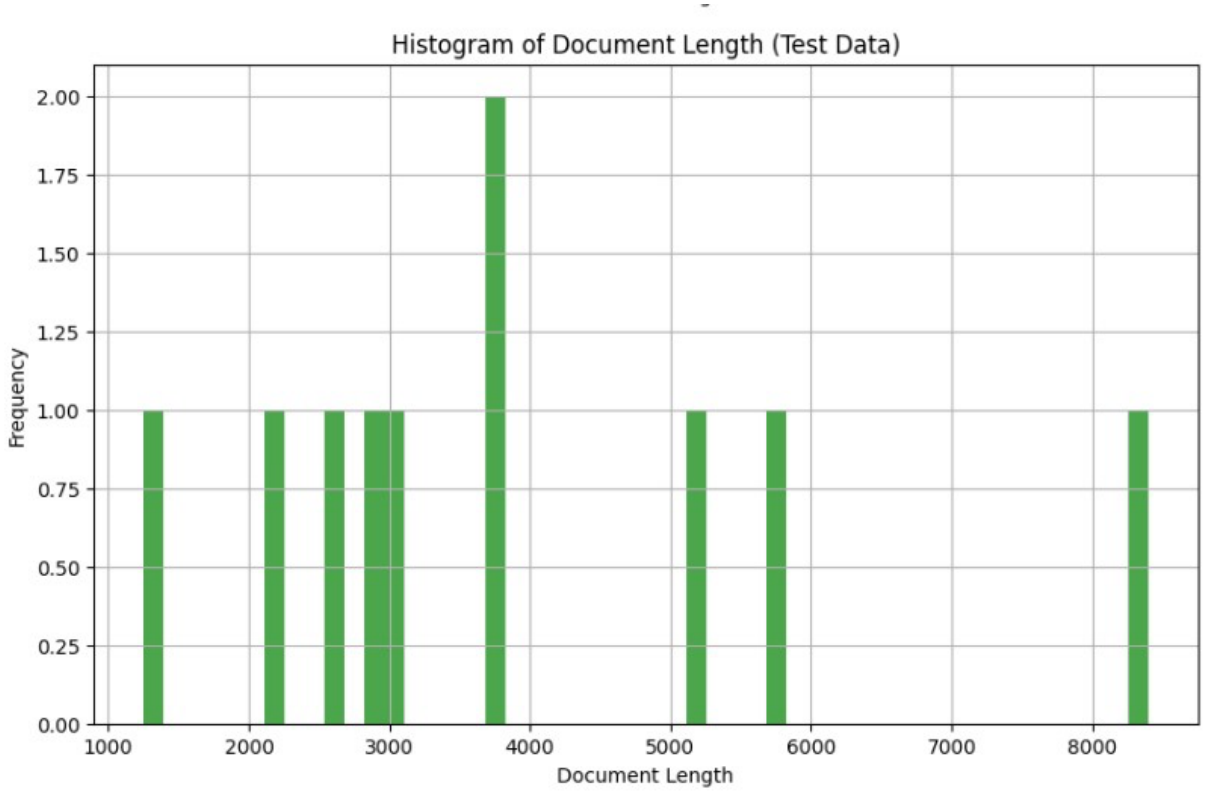
Fig 1: Histogram of Document length in Training Data



ii. Histogram of Document Length (Test Data):

Similar to the histogram representing the training data, this visualization depicts the distribution of document lengths in the test dataset. Comparing the distributions of document lengths between these two datasets allows researchers to evaluate their models' generalizability and uncover any disparities or biases that may be present. The histogram is instrumental in pinpointing noteworthy variations in document lengths between the training and test datasets, which could affect model performance and overall applicability.

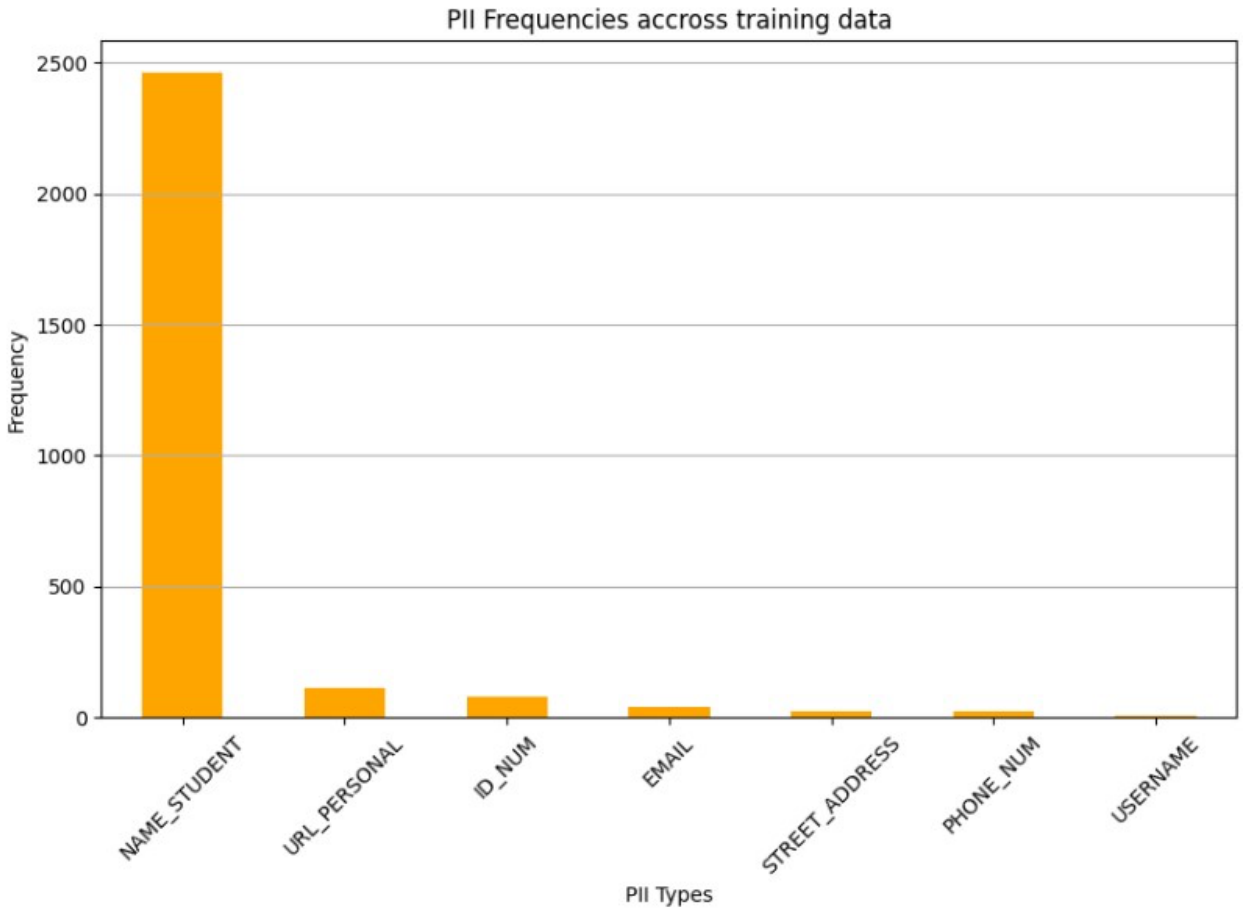
Fig 2: Histogram of Document Length of Testing data



iii. Bar Plot of PII Frequencies Across Training Data:

The bar chart illustrates the occurrences of various categories of personally identifiable information in the training dataset. Each bar corresponds to a specific type of PII, including names, email addresses, usernames, and phone numbers, along with their respective frequencies. Examining the distribution of PII types within the dataset offers valuable insights into the prevalence of sensitive information. From this bar plot we can observe that “Name_Student” category is relatively more when compared to the other categroies.

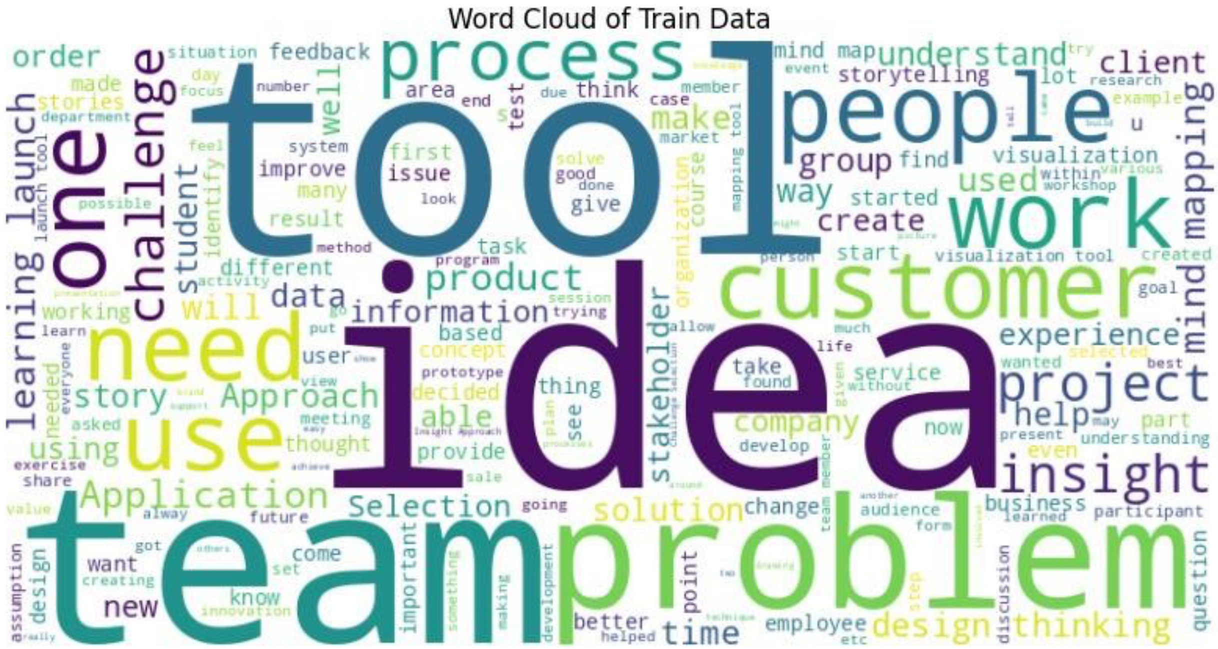
Fig 3: Bar Plot of Frequencies of PII categories in Training data



iv. Word Cloud of Training data:

The Word Cloud representation method provides an engaging insight into the most frequently used words in the essays of the training dataset. The presented Word Cloud, which is created from the combined text of all essays, highlights specific terms displayed in varying font sizes based on their frequency. Words like "one," "challenge," "tool," "process," "work," "people," "idea," and "team" are recurring themes within the essays. Furthermore, certain terms such as “Tool,” “idea,” and “team” stand out more prominently due to their higher frequency, as reflected through their larger size in the visualization. This visualization offers valuable understanding of prevalent topics and concepts covered in the essays, providing a glimpse into students' viewpoints, concerns, and focal points. Such insights play a crucial role in guiding subsequent analyses and model development efforts for effectively addressing underlying themes while

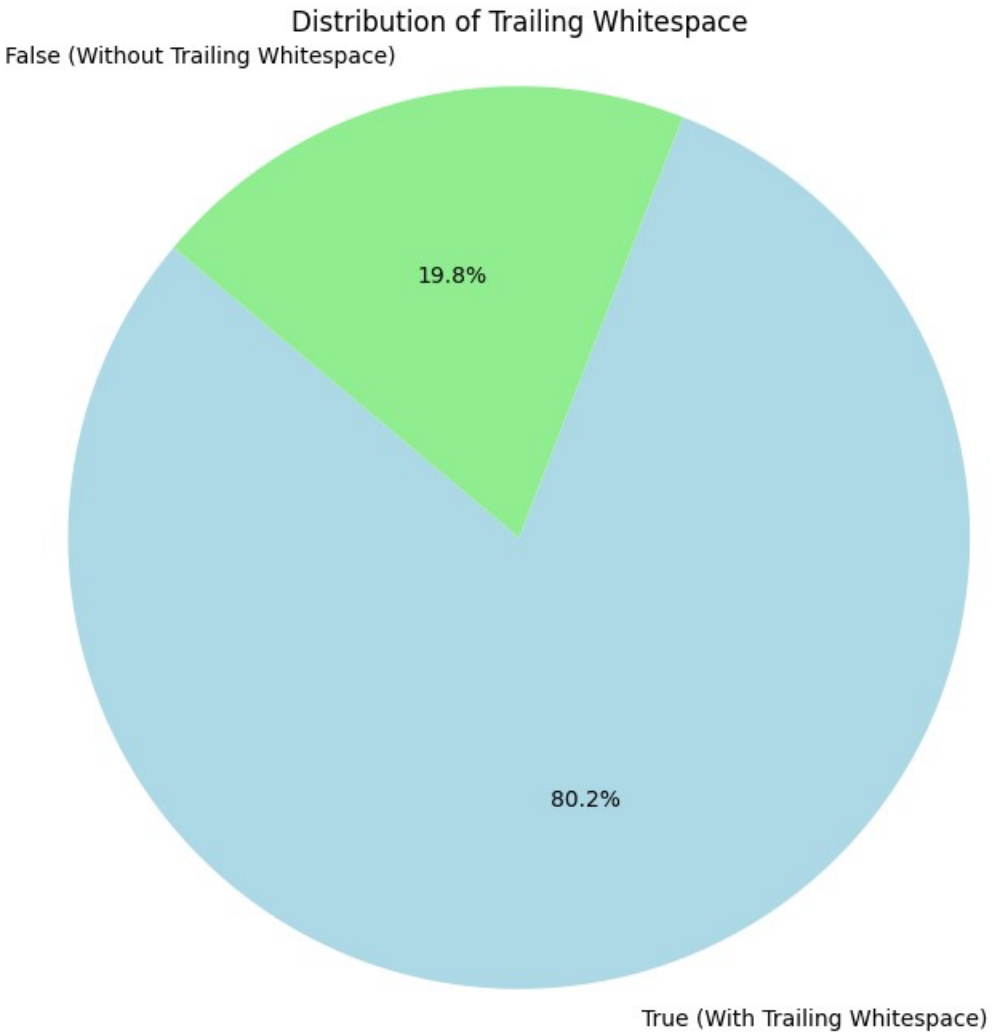
navigating complexities related to PII detection and student privacy protection. **Fig 4:** Word cloud of the training data



v. Distribution of Trailing Whitespace:

This pie chart illustrates how whitespace at the end of tokenized words is distributed in the dataset. Trailing whitespace, which denotes extra space characters after a token, can affect tasks like natural language understanding and machine learning. The pie chart divides tokens into two categories: those with trailing whitespace and those without it. Analyzing this distribution assists researchers in recognizing potential data preprocessing difficulties and maintaining uniformity in tokenization throughout the dataset. From the pie chart we can observe that 19.8% of text is not having trailing whitespace.

Fig 5: Pie chart depicting the distribution of trailing whitespaces



V. Problem Formulation

The problem of PII (Personally Identifiable Information) detection in text documents is framed as a sequence labeling task. Each token in a sequence (i.e., a document) must be classified as either being part of a PII entity or not. The categories include labels for different types of PII, such as names or identifiers, with distinctions made between the beginning of a PII entity (B-) and continuation of a PII entity (I-). This formulation allows the model to leverage context from surrounding tokens, which is essential for accurately identifying and classifying entities in text.

VI. Model

DistilBERT

Model Selection:

DistilBERT is a lighter version of BERT that retains most of its predecessor's performance capabilities but is faster and more efficient in terms of memory and processing time. retains about 97% of BERT's performance on language understanding benchmarks while being 40% smaller and 60% faster. This efficiency is achieved through a technique known as knowledge distillation, where the distilled model (DistilBERT) is trained to replicate the behavior of the larger pre-trained model (BERT).

It is built on the transformer mechanism, specifically employing a multi-headed selfattention mechanism, which allows it to efficiently handle dependencies in input data. It has 6 transformer layers (compared to BERT's 12 in its base variant), with each layer comprising a selfattention layer followed by a feed-forward neural network. Each of these components includes layer normalization and residual connections, which help in stabilizing the learning process across deeper architectures.

It uses a tokenization system that splits input text into subwords, which enables the handling of out-of-vocabulary words more effectively. The model is trained on a task of predicting masked tokens in the input, similar to BERT's pre-training, but the distillation process involves transferring knowledge from BERT by mimicking its hidden states, attention distributions, and output. And it is particularly well-suited for environments where there are constraints on computational resources, such as in deployment scenarios or where quick iteration is needed during development.

During training, the model is fine-tuned using a labeled dataset where each token in the documents is annotated with a PII category. The training process involves adjusting the weights of the pretrained network using a relatively small learning rate to gradually adapt the model to the specifics of the PII detection task.

Preprocessing:

In the context of PII detection using the DistilBERT model, preprocessing involves below key steps:

Text Normalization:

The raw text from educational documents is normalized to ensure consistency. This typically includes converting all text to lower case, removing, or replacing special characters, and standardizing word contractions and punctuation. These steps help reduce the variability of the input data, which can improve model performance.

Tokenization:

Utilizing the DistilBertTokenizerFast, the text is broken down into tokens. This tokenizer handles sub-word tokenization, which is crucial for handling out-of-vocabulary words by breaking them down into smaller, recognizable subunits.

Alignment of Labels with Tokens:

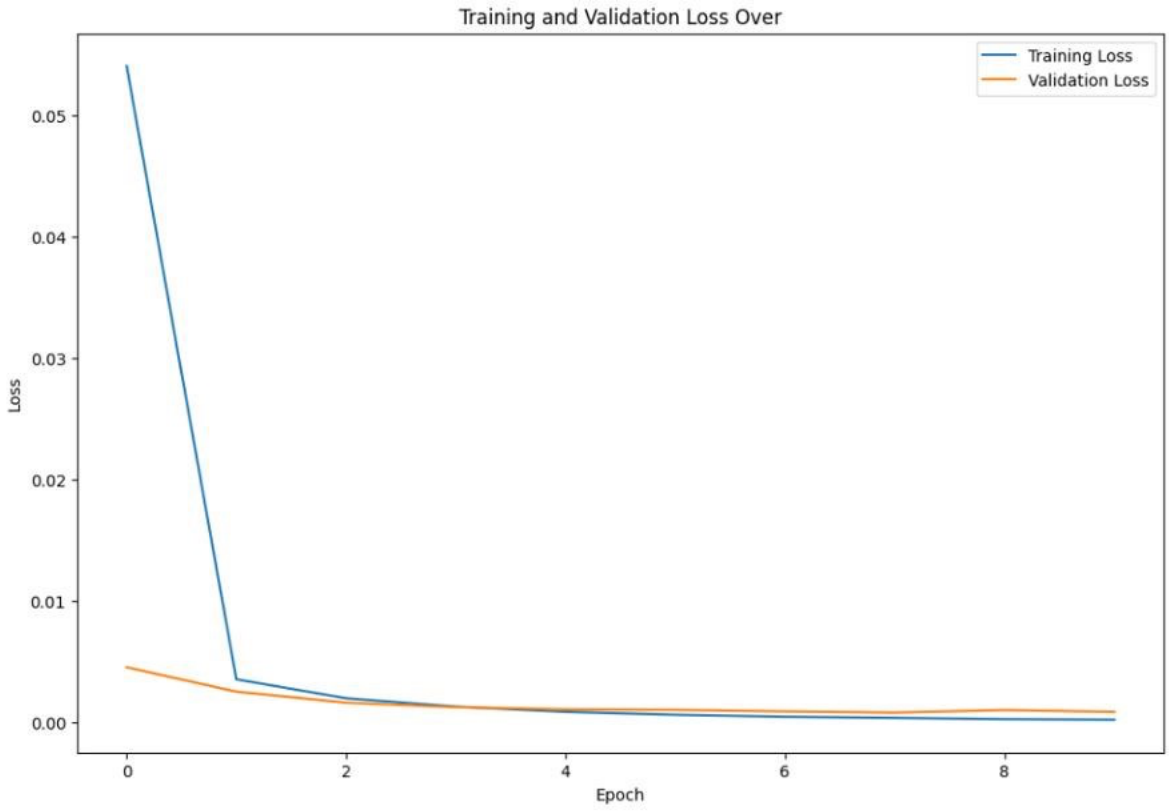
As the model uses sub-word tokenization, it's essential to align the original labels with the generated tokens. This is particularly challenging as each word may be split into multiple tokens, and not all tokens correspond directly to a word in the original text. The preprocessing script ensures that each token receives the correct label, with continuation sub-tokens either inheriting the label from the preceding token or being marked to ignore during loss calculation, typically using a label such as -100.

Model evaluation:

We have evaluated model over a 10 training epochs and configured with a learning rate of 2e-5 and utilizing the AdamW optimizer with Categorical Crossentropy as loss function, the model's setup was aimed at optimizing the balance between quick convergence and avoiding local minima, critical for maintaining the generalization ability of the model on unseen data. Over the course of the training, the model demonstrated a marked and consistent decrease in both training and validation losses, indicative of effective learning and adaptation to the nuances of the dataset. Importantly, the validation loss closely mirrored the training loss across all epochs, suggesting that

the model was not overfitting to the training data. A batch size of 16 was used for both training and validation loaders.

Fig 7: Model evaluation of DistilBERT



Result Analysis:

After successfully training the model and achieving high performance metrics, we have saved and subsequently deployed to infer on the test data. The saved model was utilized effectively, demonstrating its robust capability to identify PII within the test dataset. It effectively identified PII tokens in the test data, with correct classifications as shown in the sample output below. The tokens are correctly identified with their respective labels, and the predictions align well with the ground truth, indicating the model's effectiveness.

Fig 8: Result analysis of DistilBERT

	document	token	label	token_str	row_id
0	7	12	B-NAME_STUDENT	nathalie	0
1	7	15	I-NAME_STUDENT	sylla	1
2	7	474	B-NAME_STUDENT	nathalie	2
3	7	477	I-NAME_STUDENT	sylla	3
4	10	1	B-NAME_STUDENT	diego	4
5	10	2	I-NAME_STUDENT	estrada	5
6	10	444	B-NAME_STUDENT	diego	6
7	10	445	I-NAME_STUDENT	estrada	7
8	16	4	B-NAME_STUDENT	gilberto	8
9	16	6	I-NAME_STUDENT	gamboa	9
10	20	5	B-NAME_STUDENT	sindy	10
11	20	7	I-NAME_STUDENT	samaca	11
12	20	9	I-NAME_STUDENT	gitam	12
13	56	9	B-NAME_STUDENT	nadine	13
14	56	11	I-NAME_STUDENT	born	14
15	86	7	B-NAME_STUDENT	eladio	15
16	86	10	I-NAME_STUDENT	amaya	16
17	93	1	B-NAME_STUDENT	silvia	17
18	93	2	I-NAME_STUDENT	villalobos	18
19	104	7	B-NAME_STUDENT	sakir	19
20	104	9	I-NAME_STUDENT	ahmad	20
21	112	5	B-NAME_STUDENT	francisco	21
22	112	6	I-NAME_STUDENT	ferreira	22
23	123	29	B-NAME_STUDENT	stefano	23
24	123	30	I-NAME_STUDENT	lovato	24

References

[1] Poornima K., Cauvery N K [2021]. Personally Identifiable Information (PII) Detection in the Unstructured Large Text Corpus using Natural Language Processing and Unsupervised Learning Technique : International Journal of Advanced Computer Science and Applications 12(9) https://www.researchgate.net/publication/355085553_Personally_Identifiable_Information_PII_Detection_in_the_Unstructured_Large_Text_Corpus_using_Natural_Language_Processing_and_Unsupervised_Learning_Technique

[2] Soumit R., Mainak M. [2018]. Identification And Processing of PII Data, Applying Deep Learning

Models With Improved Accuracy And Efficiency : ResearchGate https://www.researchgate.net/publication/376078296_IDENTIFICATION_AND_PROCESSING_OF_PII_DATA_APPLYING_DEEP_LEARNING_MODELS_WITH_IMPROVED_ACCURACY_AND_EFFICIENCY

[3] Jaikishan J., Mohana [2023]. Privacy-Preserving Personal Identifiable Information (PII) Label Detection Using Machine Learning: 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT) https://www.researchgate.net/publication/375880997_Privacy-Preserving_Personal_Identifiable_Information_PII_Label_Detection_Using_Machine_Learning

[4] Paulo S., Carolina G., Nuno A., Marilia C., Bogdan W. [2022]. Privacy risk assessment and privacy-preserving data monitoring: Expert Systems with Applications Volume 200 https://doi.org/10.1016/j.eswa.2022.116867

[5] Brad P. [2020]. Privacy protection with AI: Survey of data-anonymization techniques https://bradpayne.ca/privacy-protection-with-ai-survey-of-data-anonymization-techniques/