**SAHIL**

**21113135**

**CIVIL,3Y**

* Utilized the Stochastic Gradient Descent (SGD) Classifier model in conjunction with TF-IDF vectorization to identify artificially generated text.
* Collected a diverse dataset, sourced from Kaggle datasets ([source1](https://www.kaggle.com/datasets/nbroad/daigt-data-llama-70b-and-falcon180b), [source2](https://www.kaggle.com/datasets/thedrcat/daigt-v2-train-dataset)), comprising both human-written and AI-generated text for comprehensive training and evaluation.
* Analyzed and compared the most common words in essays written by students and AI-generated text, visualizing the distinctions.
* Plotted and compared the essay length of AI-generated and student-written essays to observe potential patterns.
* Split the training data into 15 parts, designating one part (12th) as the validation dataset for model evaluation.
* Constructed a vocabulary using tokenization on the training dataset to facilitate TF-IDF vectorization.
* Applied TF-IDF vectorization to convert tokenized text into numerical features, highlighting term importance based on their frequency within a document and rarity across the entire dataset.
* Trained an SGDClassifier model on the TF-IDF transformed data to discern patterns and characteristics distinguishing between human and AI-generated text.
* Evaluated the model's performance on both training and validation sets, assessing metrics including false negatives, false positives, and accuracy.