Automated Grading and Feedback System for Assignments using NLP

Neel Singh
MS CS
Purdue University
singn03@pfw.edu

Ajinkya More
MS CS
Purdue University
morea04@pfw.edu

Arjun Malik
MS CS
Purdue University
malia01@pfw.edu

Abstract

In college departments, grading assignments can be a time-consuming and labor-intensive process for instructors. This project aims to develop an automated grading and feedback system using Document AI techniques to streamline the grading process, provide timely feedback to students, and improve the efficiency of assessment workflows.

1 Introduction

Grading assignments in college departments can be a labor-intensive and time-consuming task for instructors. The proposed Automated Grading and Feedback System aims to address this issue by automating the grading process, thereby reducing grading time and providing students with immediate feedback on their assignments. This system leverages Document AI techniques and will be powered by the Hewlett Foundation AES Dataset.

2 Motivation

The motivation behind this project stems from the need to enhance the learning experience for students and reduce the burden on instructors within our college department. Manual grading often leads to delayed feedback, which can hinder students' progress and cause anxiety. By automating the grading process, we can provide students with rapid and constructive feedback, ultimately improving their learning outcomes.

3 Problem Statement

The primary problem to address is the inefficiency and time consumption of manual assignment grading. Students often experience anxiety while waiting for their assignments to be graded, affecting their overall learning experience. The proposed system aims to provide a solution by automating the grading process using the Hewlett Foundation AES Dataset, enabling students to receive rapid feedback and improving instructors' workflow.

4 Model/Algorithm to Address Problem

The system will consist of several key components:

4.1 Data Collection

The primary dataset for this project is the Hewlett Foundation AES Dataset. This dataset includes essays written by students and their corresponding scores, making it ideal for training and testing the automated grading and feedback generation components of our system.

4.2 Text Extraction and Analysis

Optical Character Recognition (OCR) and text extraction techniques will be implemented to convert scanned or image-based assignments into machine-readable text.

4.3 Automated Grading

A grading model will be developed using supervised learning techniques, using the Hewlett Foundation AES Dataset where essays are manually graded.

4.4 Feedback Generation

The system will generate detailed feedback for students based on the grading model's assessment, using scoring information from the Hewlett Foundation AES Dataset. Feedback will include comments, suggestions for improvement, and grades.

4.5 Plagiarism Detection

Algorithms for plagiarism detection will be implemented using the Hewlett Foundation AES Dataset to identify potential instances of academic dishonesty in student assignments.

5 Datasets

The primary dataset for this project is the Hewlett Foundation AES Dataset.

6 Measurement of Success

The success of this project will be measured based on a set of key performance metrics relevant to the problem:

- Accuracy of Automated Grading: We will assess the accuracy of the automated grading model by comparing its grades to the ground truth grades provided in the Hewlett Foundation AES Dataset. This evaluation will include metrics such as accuracy, precision, recall, and the F1 score, providing a comprehensive understanding of the grading model's performance.

- Quality and Helpful Feedback: The generated feedback's quality and helpfulness will be evaluated through qualitative analysis by both students and instructors. Additionally, we may employ natural language processing metrics, including BLEU, ROUGE, and readability scores, to quantify the quality of the feedback.

- Effectiveness of Plagiarism Detection: Plagiarism detection algorithms will be assessed for their effectiveness in identifying potential instances of academic dishonesty. Metrics such as true positive rate, false positive rate, and the F1 score will be used to evaluate the performance of the plagiarism detection.

7 Automated Grading and Feedback System for Assignments: Project Update 1

7.1 Current Results

We started our project by developing three models, which were Random Forest (RF), Logistic Regression (LR) and Support Vector Machine. After preprocessing our features using TFIDF vectorizer, we were able to reach F1 score of nearly 50 % which could be partly attributed to difficulties inherent in using traditional machine learning models for essay grading.

7.2 Upcoming Results

Given our initial results, we are now exploring more in-depth preprocessing techniques. We believe more sophisticated data preprocessing can significantly increase the models' ability to comprehend and grade essays accurately. Furthermore, we may switch over from traditional machine learning models to deep learning based models like Roberta and GPT2, which have their own tokenizers, embeddings and model and can offer more granular understandings of textual data for improved grading accuracy.

7.3 Analysis Done

We processed the essays using basic preprocessing, such as converting to lowercase and removing non-alphabetic characters, before employing the TF-IDF vectorizer for feature extraction. Although our models were rigorously trained, their F1 score was not too high, leading us to speculate that training features might not have been rhobust enough to significantly explain model variance.

Performance Metrics:

Model	F1 Score
Random Forest	0.496
Logistic Regression	0.445
SVM	0.480

7.4 Upcoming Analysis

Now that we have gained valuable insights, we plan to:

- Explore advanced preprocessing techniques like lemmatization, stopword removal and n-gram feature extraction further. - Neural networks like GPT2 and tiny Roberta will be tested to see how significantly they improve performance metrics. - We will also focus on hyperparameter tuning both for potential neural networks as well as our feature extraction methods in order to guarantee optimal performance. - Given the nature of our dataset and task, metrics other than accuracy will be explored, including Root Mean Squared Error and Mean Absolute Error for more comprehensive evaluation.

7.5 Problems Encountered

Our initial models proved less-than-successful at accurately predicting results than anticipated, leading us to re-evaluate preprocessing and modeling strategies and attempt to balance out samples as some had less samples than anticipated leading to potential bias in predictions. - Balancing was also

a concern since some scores may contain less samples leading to potential errors in predictions. - Lack of diversity within our dataset and too simple model (underfitting) could pose an obstacle to models generalization power.

7.6 Conclusion

This project has been an adventure of discovery. Early setbacks in model performance only encouraged us to dive deeper and discover more advanced techniques. Now with preprocessing and neural networks as key focus areas, we remain optimistic of reaching our goal of an automated grading system.