GradeMate: Automated Exam Grading System

Final Year Project Proposal By

Muhammad Tayyab BSDSF21M007

Hamza Safwan BSDSF21M027

Ahmar BSDSF21M043

Muhammad Hussain BSDSF21M052

Advisor: Dr. Zubair Nawaz

Department of Data Science University of the Punjab

Lahore, Pakistan (2024)

GradeMate: Automated Exam Grading System

**Abstract**

This project proposes the development of an AI-powered system capable of accurately grading both multiple-choice questions (MCQs) and subjective questions, including open-ended questions (OEQs) and essays, in exams. The system will leverage natural language processing (NLP) techniques to understand and analyze the text content of subjective questions, enabling accurate grading. Machine learning algorithms will be employed to train the AI model on a diverse dataset of graded exams, improving its ability to evaluate both MCQ and subjective questions.

The expected outcomes of this project include accurate grading of both MCQs and subjective questions, improved efficiency in the grading process, enhanced objectivity and consistency in evaluation, and scalability to handle large volumes of exams. By automating the grading process, the system will reduce the time and effort required by human graders, allowing for faster feedback to students. Additionally, the AI-based grading system will help minimize human bias and ensure consistent evaluation standards, promoting fairness and transparency in the assessment process.

**1. INTRODUCTION**

**Background:**

The field of artificial intelligence (AI) & machine learning (ML) has witnessed remarkable advancements in recent years, with deep learning (DL) and its sub-domains Natural Language Processing (NLP) and Computer Vision (CV) emerging as powerful tools for a wide range of applications. In the realm of education, AI-based solutions have the potential to revolutionize traditional practices, including the assessment of student learning.

**Historical Perspective:**

* **Manual evaluation:** For centuries, exams have been evaluated manually by human examiners. This process is time-consuming, subjective, and prone to errors.
* **Computer-assisted evaluation:** While early attempts at computer-assisted evaluation focused on objective questions, the complexity of grading open-ended responses has remained a significant challenge.

**Problem Statement:**

The manual grading of exams in educational institutions is a time-consuming, labor-intensive, and subjective process. This often leads to inconsistent grading, increased workload for educators, and delayed feedback for students. Additionally, the subjective nature of human grading can introduce bias and hinder the accuracy of assessments.

**Significance of the Problem:**

The problem of inefficient and subjective exam grading is significant due to several factors:

* **Increased workload for educators:** Manual grading consumes a substantial amount of time, diverting educators’ attention from teaching and research activities.
* **Delayed feedback for students:** The time taken for manual grading can lead to delays in feedback, hindering students' learning and progress.
* **Inconsistent grading:** Human graders may have varying interpretations of assessment criteria, resulting in inconsistent scores and potential unfairness.
* **Subjectivity and bias:** Personal biases and preferences can influence grading decisions, compromising the objectivity of assessments.

**Target Customers:**

* **Educational institutions:** Universities, colleges, and schools can benefit from an automated grading system to reduce workload, improve efficiency, and ensure consistent assessments.
* **Students:** Students can benefit from faster feedback and more accurate assessments, leading to improved learning outcomes.
* **Educators:** Educators can free up their time to focus on teaching, research, and personalized instruction.

**Market Size:**

The potential market for AI-powered exam grading is significant, considering the global prevalence of educational institutions and the growing demand for efficient and equitable assessment practices. While quantifying the exact market size is challenging, it is evident that there is a substantial need for such solutions across various educational levels and regions.

**Objectives:**

* **Automate the grading process:** Develop an AI-based system capable of accurately grading both objective and subjective exam questions.
* **Improve grading efficiency:** Reduce the time and effort required for grading, allowing educators to focus on other important tasks.
* **Ensure grading consistency:** Minimize the impact of human subjectivity and bias on grading outcomes.
* **Provide timely feedback to students:** Enable faster return of graded exams, facilitating timely learning and improvement.
* **Enhance the overall assessment process:** Improve the quality and efficiency of assessment practices in educational institutions.

**Scope and Limitations:**

The project will focus on developing an autonomous exam grading system for a specific set of educational subjects and question formats. The system will be trained on a large dataset of graded exams to learn to accurately assess student responses. However, the system may have limitations in handling complex or ambiguous questions, or in adapting to new subject areas or question formats.

**2. LITERATURE REVIEW**

**Related Work:**

Automated grading systems have gained substantial attention in recent years, with several approaches aimed at reducing the labor-intensive process of manual grading while ensuring consistency and fairness. Early studies on automated essay scoring systems focused on addressing limitations such as coherence and content relevance (Jagadamba & Shree G., 2020) tackled the issue of subjective question grading, which remains a challenge in existing online exam systems. Their AI-based verifier assesses student responses by comparing them with predefined model answers, focusing on keywords, grammar, and core concepts. Although this system reduces manual workload, the inability to evaluate complex sentence structures and diverse responses highlights the limitations of current technologies in handling more sophisticated answers​.

(Ramesh & Sanampudi, 2022) reviewed traditional natural language processing (NLP) models, revealing challenges like surface-level analysis without adequately assessing deeper content comprehension. They concluded that while deep learning models have shown promise, there is still a need for advancements in grading essays based on content and conceptual understanding.

A more advanced approach to automated grading is introduced by (Gobrecht et al., 2024) who developed an AI-driven short-answer grading system leveraging fine-tuned transformer models. This system outperformed human graders by reducing subjectivity and error, especially in unseen questions and courses. The model was found to be more consistent than human re-graders, significantly reducing the absolute error rate by 44%. However, their research also identified a decline in accuracy with increasingly complex questions, which poses a challenge for scaling such systems across diverse academic disciplines​. In a more recent study, (Lee & Song, 2024) experimented with large language models (LLMs) like GPT-3.5, GPT-4.0, and Gemini-pro in their AI grader system for college exams. GPT-4.0 demonstrated the most reliable and consistent results compared to human graders. Their work highlights the potential of LLMs in producing accurate and objective grading, reducing human subjectivity and overwork. However, the system's dependence on well-structured rubrics suggests that AI systems must be paired with clear guidelines to maximize effectiveness​.

**Gap Analysis:**

Despite the significant advancements in autonomous grading systems, there are several gaps that remain unaddressed, which our project aims to fill:

* **Complexity of Subjective Answer Evaluation:**

Current autonomous grading systems primarily focus on keyword matching and grammatical accuracy. However, these systems struggle with more nuanced aspects such as understanding complex sentence structures, evaluating mathematical formulas, or handling context-rich answers that do not strictly follow a template. Our project will address this by incorporating advanced natural language processing (NLP) techniques to better assess the depth of reasoning and conceptual understanding in subjective answers.

* **Handling Variations in Student Responses:**

Many existing automated grading systems focus on consistency but may falter when faced with a wide range of acceptable answers that vary in expression but convey the same meaning. Our project will introduce semantic analysis techniques that allow the system to recognize and score rephrased answers, ensuring that students are not penalized for minor variations in language or structure.

* **Scalability Across Disciplines:**

Some systems demonstrated success in reducing human error in specific subjects, their model's performance dropped when dealing with more complex or interdisciplinary questions Similarly, many found that their GPT-based grading system worked well in technical subjects but left open questions about its application in fields like humanities or social sciences, where subjective interpretation plays a larger role. Our project aims to create a more flexible grading model capable of handling both technical and non-technical subjects by fine-tuning models for interdisciplinary use.

* **Limited Focus on Feedback Quality:**

Although some systems provide accurate grading, the quality of feedback given to students remains an underexplored area. Current models either do not provide feedback or offer minimal insights into how students can improve. Our project will integrate feedback generation capabilities, using AI to not only grade but also explain the reasoning behind the grades and offer suggestions for improvement, thereby enhancing the learning experience.

**3. PROJECT OVERVIEW/GOAL**

**Project Title**: GradeMate: Automated Exam Grading System

**Group Leader:** Hamza Safwan

**Project Members:**

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Registration# | Email Address | Signature |
| Muhammad Tayyab | BSDSF21M007 | bsdsf21m007@pucit.edu.pk |  |
| Hamza Safwan | BSDSF21M027 | bsdsf21m027@pucit.edu.pk |  |
| Ahmar | BSDSF21M043 | bsdsf21m043@pucit.edu.pk |  |
| Muhammad Hussain | BSDSF21M052 | bsdsf21m052@pucit.edu.pk |  |
|  |  |  |  |

**Project Goal:** To develop a robust AI-powered grading system that accurately evaluates both MCQ and subjective questions in exams, ensuring efficiency, objectivity, and scalability while providing valuable feedback to students.

**Objectives:**

1. Automate the grading process, reducing need for manual grading.

2. Provide quick and accurate grading for both MCQs & OEQs.

3. Build scalable system that can handle grading for large institutions

4. Deliver personalized feedback to students, helping them understand

their mistakes & also build trust in the autonomous grading system.

**Project Success criteria:** The automated grading system will be

considered successful if it accurately evaluates exams, is efficient, scalable, and provides meaningful feedback.

**Assumptions:**

User have a good quality visible exam script which is to grade.

User have a standard question formats in exam script.

**Risks and Obstacles:**

OCR failure (Unable to extract answers from exam due to poor quality exam script input)

Integration challenges (LMS, institutional infrastructure)

Human error (inaccurate input)

**Acceptance (**Resistance to AI-powered grading)

**Organization Address**: Department of Data Science, University of The Punjab, Lahore, Pakistan

**Target End users:** Educational Institutions, Educators, Teachers, Students

**Suggested Project Supervisor:** Dr. Zubair Nawaz

**Approved By:**

**Date:**

**4. Tools and technologies used with reasoning**

* **Frontend**
* Languages: HTML, CSS, JavaScript
* Libraries: ReactJS or AngularJS, ReactNative JS
* **Backend**
  + Languages: Python, C++
  + Frameworks: Django or FastAPI
* **Databases MySQL**/PostgreSQL
* **AI/ML Libraries:** TensorFlow, PyTorch, Hugging Face Transformers
* **Tools:** Visual Studio/ Visual Studio Code, Jupyter Notebook, Annaconda

**5. WORK DIVISION**

**Rotational Learning and Development Approach**

Each member will rotate responsibilities across the major areas of the project (Frontend, Backend, AI Model Development, Data Processing, OCR, and Testing) every 3–4 weeks. This allows every team member to experience each phase of the project.

**Phase 1: System Design & Project Planning (All Members)**

**Duration:** First 2 weeks

**Tasks:**

* **Collaborative Discussion:**

Design the system architecture.

Plan the project timeline, setting goals and milestones.

Decide on the technology stack to use (frontend, backend, AI models, etc.).

* **Learning Resources:**

Explore the basics of OCR, AI models for NLP, and grading systems.

Set up cloud services and databases.

* **Outcome:**

A detailed project plan and initial setup for the architecture.

* **Participation:** All members work together to ensure a uniform understanding of the overall system.

**Phase 2: OCR & Data Processing (Rotation 1)**

Each member will have hands-on experience in implementing and testing OCR.

**Duration:** 3–4 weeks

**Tasks:**

* Implement OCR using Paddle/EasyOCR/Tesseract.
* Fine-tune OCR for accurate extraction of text from exam scripts (including handwriting).
* Pre-process exam script images for improved accuracy (noise removal, segmentation, etc.).
* Connect the extracted data to a database.

**Participation:**

Member 1 & 2: Focus on OCR integration and text extraction.

Member 3 & 4: Handle database connectivity and processing of OCR results.

**After 2 weeks, rotate:**

Member 3 & 4: Work on OCR integration.

Member 1 & 2: Take over database connectivity and processing.

**Phase 3: Frontend & Backend Development (Rotation 2)**

All members will have a chance to contribute to both frontend and backend.

**Duration:** 4 weeks

**Frontend Tasks:**

* Build student and teacher dashboards for uploading exam scripts and viewing grades.
* Implement form validation, file upload mechanisms, and result visualization (using React/Angular).

**Backend Tasks:**

* Develop backend API using FastAPI/Django.
* Implement user authentication and connect frontend with backend.
* Build APIs for submitting exam scripts and returning results.

**Participation:**

* **Weeks 1–2:**

Member 1 & 3: Focus on frontend (React/Angular).

Member 2 & 4: Work on backend API and connecting to the database.

* **Weeks 3–4 (Rotate):**

Member 2 & 4: Switch to frontend.

Member 1 & 3: Switch to backend.

**Phase 4: AI Model Development & Grading Engine (Rotation 3)**

Each member will have the opportunity to work on the AI model for subjective grading and the rule-based engine for objective grading.

**Duration:** 4 weeks

**Tasks:**

* Build and fine-tune GPT-based models for subjective grading.
* Implement feedback generation for subjective questions.
* Develop the objective grading engine (multiple-choice, true/false) and integrate it with the backend.
* Test AI models with different data sets and evaluate performance.

**Participation:**

* **Weeks 1–2:**

Member 1 & 4: Work on AI models for subjective grading.

Member 2 & 3: Develop the rule-based engine for objective grading.

* **Weeks 3–4 (Rotate):**

Member 2 & 3: Work on AI models.

Member 1 & 4: Switch to objective grading engine.

**Phase 5: Testing, Integration, and Deployment (All Members)**

All members will participate in testing, quality assurance, and deployment of the system.

**Duration:** 4–6 weeks

**Tasks:**

* Perform unit and integration testing.
* Ensure the entire system works seamlessly, including frontend-backend integration, OCR, and AI models.
* Deploy the system on cloud platforms (AWS, Azure).
* Prepare for final review and documentation.

**Participation:**

All Members will perform testing on different modules and integrate everything together and will work on deployment and finalize the project for presentation.

**REFERENCES**

Gobrecht, A., Tuma, F., Möller, M., Zöller, T., Zakhvatkin, M., Wuttig, A., Sommerfeldt, H., & Schütt, S. (2024). Beyond human subjectivity and error: a novel AI grading system. *ArXiv Preprint ArXiv:2405.04323*.

Jagadamba, G., & Shree G., C. (2020). Online Subjective answer verifying system Using Artificial Intelligence. *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, 1023–1027. https://doi.org/10.1109/I-SMAC49090.2020.9243601

Lee, J. X., & Song, Y.-T. (2024). College Exam Grader using LLM AI models. *2024 IEEE/ACIS 27th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, 282–289. https://doi.org/10.1109/SNPD61259.2024.10673924

Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, *55*(3), 2495–2527. https://doi.org/10.1007/s10462-021-10068-2