AI-Driven Educational Analytics System for Exam Question Difficulty Assessment

Progyan Sen, Roni Maity, Arpit Sarang, Ganesh Wayal
Team Name: prag

***Abstract*—This project presents an AI-driven educational analytics system designed to evaluate exam question quality and difficulty using machine learning and natural language processing techniques. The system analyzes textual features of questions and engagement-based metrics to classify difficulty into Easy, Medium, and Hard categories. A TF-IDF based feature extraction method combined with Logistic Regression is used for supervised classification. The system is deployed with a public user interface to provide analytical insights into question performance and difficulty patterns.**

***Index Terms*—Educational Analytics, NLP, TF-IDF, Logistic Regression, Question Difficulty Prediction, Machine Learning**

I. Introduction

Assessment design plays a crucial role in measuring student understanding. Poorly calibrated questions can distort evaluation outcomes. This project proposes a machine learning-based educational analytics system that predicts question difficulty using textual and engagement-based signals.

The system processes large-scale question data and classifies difficulty levels while providing statistical insights through an interactive dashboard.

II. Problem Statement

The objective is to design and implement a machine learning system that:

- Analyzes question text
- Uses engagement metrics (views, answers, score)
- Predicts question difficulty category
- Provides statistical and visual insights
- Is publicly deployed with a working UI

III. Dataset Description

The dataset is derived from the StackOverflow dataset available on Kaggle.

Initial dataset size: 686,252 questions.

Filtering conditions:

- At least one answer
- View count greater than 50
- Non-null title and body

Difficulty metric:

$$Difficulty = \frac{AnswerCount}{ViewCount}$$

Quantile-based thresholds (33%, 66%) classify questions into:

- Hard
- Medium
- Easy

Final class distribution:

- Easy: 231,903
- Medium: 227,844
- Hard: 226,505

IV. System Architecture

The system follows this pipeline:

1) Data cleaning and HTML removal
2) Text preprocessing
3) TF-IDF feature extraction
4) Numeric feature scaling
5) Feature concatenation
6) Logistic Regression training
7) Evaluation
8) UI deployment

V. Methodology

A. Text Processing

- HTML removal using BeautifulSoup
- URL removal
- Whitespace normalization
- Removal of code blocks

B. Feature Engineering

Text Features:

- TF-IDF
- ngram range = (1,2)
- max_features = 50,000
- min_df = 5

Numeric Features:

- Question length
- Score

Final feature matrix size:

$$(299,997 \times 50,002)$$

C. Model Used

Logistic Regression with:

- Solver: saga
- Max iterations: 5000
- Class weight: balanced

Train-test split:

- 80% Training
- 20% Testing

VI. Results

A. Training Dataset

Sampled dataset size: 300,000
Class distribution:

- Easy: 101,575
- Medium: 99,650
- Hard: 98,775

B. Model Performance

Overall Accuracy:

$$Accuracy = 0.4926$$

Class	Precision	Recall	F1-score
Easy	0.52	0.60	0.56
Hard	0.54	0.50	0.52
Medium	0.41	0.37	0.39

TABLE I
Classification Report

Macro Average F1-score: 0.49

C. Confusion Matrix

$$\begin{bmatrix} 12240 & 2893 & 5182 \\ 4402 & 9957 & 5396 \\ 7054 & 5516 & 7360 \end{bmatrix}$$

The model performs relatively better in distinguishing Easy questions, while Medium questions show significant overlap with both Easy and Hard categories.

VII. User Interface

The system includes a Streamlit-based publicly hosted dashboard.

Dashboard Features:

- CSV dataset upload
- Automatic difficulty prediction
- Display of predicted difficulty labels
- Class distribution visualization
- Confusion matrix visualization
- Question-level statistics (length, score, engagement)

The interface allows educators to analyze question quality interactively.

VIII. Discussion

The model achieves 49% accuracy on a balanced three-class dataset. While textual TF-IDF features combined with numeric metadata provide predictive signals, the overlap between Medium and adjacent difficulty classes reduces classification clarity.

The quantile-based difficulty labeling approach introduces boundary ambiguity, which contributes to confusion between classes.

IX. Conclusion

This project demonstrates the implementation of a scalable ML-based educational analytics system for question difficulty assessment. Although performance is moderate, the system successfully integrates NLP feature extraction, supervised classification, and public deployment into a unified framework. Future improvements include enhanced difficulty metrics, semantic embeddings, and advanced modeling techniques.

Repository

https://github.com/praggCode/Exam-Analytics-System