# Automated Short Answer Grading (ASAG)

## Mini Project Presentation

Lana Anvar (80522012)
Neha A S (80522016)

M.Sc. (Five Year Integrated) in
Computer Science (Artificial Intelligence & Data Science)

29 April 2025

COCHIN UNIVERSITY OF SCIENCE AND TECHNOLOGY

# Agenda

# Problem Statement

- **Core Issue**:Automating the grading of short textual answers to reduce manual effort, minimize bias, and enable scalable assessment.

# Motivation

- Manual grading of short answers is time-consuming, subjective, and inconsistent.
- The rise of MOOCs demands scalable and fair assessment.
- NLP + ML = automated, objective, real-time evaluation.

# Literature Review: Word2Vec + PySpark Approach

**Reference:** Automated Short Answer Grading with Word Embedding-Based Semantic Similarity Using PySpark' by Akhilesh P. et al., 2024

- **Technique:** Word2Vec embeddings + Cosine Similarity to measure semantic similarity between student and reference answers.
- **Scalability:** Used PySpark for distributed processing, enabling high-volume grading.
- **Outcome:** Regression-based evaluation yielded:
  - MSE = 0.2727
  - MAE = 0.4644
  - $R^2$ = 0.67
- **Strengths:**
  - Language-model-based grading.
  - Fair, unbiased scoring using diverse training data.
  - Good semantic understanding over keyword matching.

# Literature Review: Gaps Identified

**Limitations in the Word2Vec + PySpark ASAG Model:**

- **Over-reliance on Word2Vec Static Embeddings:** Word2Vec generates static word representations and does not capture context variability (e.g., polysemy, syntax).
- **Minimal Feature Diversity and Shallow Modeling:**There was no use of advanced learning algorithms beyond basic similarity calculations."
- **Lack of Interpretability and Feature Transparency:** The use of only word embeddings and cosine similarity provides little transparency into what linguistic elements influenced the grading.
- **Dataset Size:** Relied heavily on preprocessed columns like student_modified and ref_modified, and used a relatively small, specific dataset.

# Objectives

1. **Develop an Automated Short Answer Grading (ASAG) System**: Design a system to automatically evaluate and grade student responses.
2. **Extract Multi-Dimensional Textual Features**: TF-IDF similarity, BLEU score, ROUGE-L F1 score, Jaccard similarity, word overlaps, etc.
3. **Evaluate using Regression Models**: Linear Regression, Random Forest, SVR, and Gradient Boosting using MAE, MSE, and $R^2$.

# Dataset Used: ASAG Dataset

The ASAG dataset supports research in the automatic evaluation of student responses based on model answers.

**Content Highlights:**

- 646 total entries
- Domain-specific questions paired with reference answers
- Student responses in free-text form
- Cosine similarity scores between student and reference answers
- Human-assigned grades reflecting answer relevance/correctness

# Dataset Overview and Column Selection I

**Dataset Overview:**

- 646 student responses.
- Columns available:
  - Student answers
  - Reference model answers
  - Grades
  - Preprocessed versions (modified/demoted forms)
  - Precomputed embeddings (Word2Vec-based)
  - Precalculated similarity scores and alignments
  - Question text and IDs

**Columns Used:**

| Column Name | Purpose |
|---|---|
| `student_answer` | Raw student response text |
| `ref_answer` | Corresponding reference model answer |
| `grades_round` | Human-assigned score (target variable) |

# Dataset Overview and Column Selection II

**Columns Not Used and Justification:**

- **Preprocessed Forms** (`student_modified`, `ref_modified`, etc.):
- **Embedding Columns** (`embed_stud`, `embed_ref_demoted`, etc.):
- **Similarity/Alignment Columns** (`cos_similarity`, `aligned_score`, etc.):
- **Question Metadata** (`question`, `question_id`):

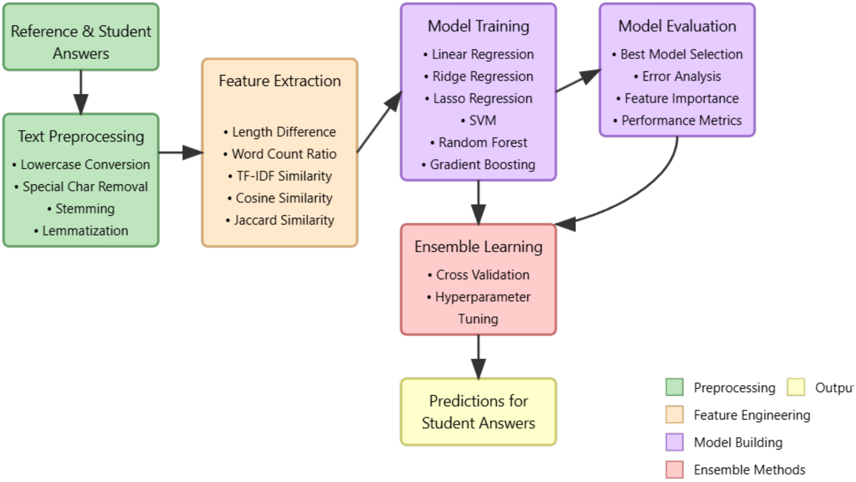
*Focus on raw inputs and ground truth labels ensured flexibility, transparency, and control over feature extraction.*

# Annotation Process in ASAG Dataset

**Annotation Process:**
Each student answer is evaluated and scored numerically, likely using expert rubrics. These scores reflect:

- Semantic similarity to model answers
- Factual accuracy
- Appropriateness in context

# System Architecture



# Feature Extraction

**Linguistic and Statistical Features Extracted:**

- **TF-IDF Cosine Similarity**
- **BLEU Score (N-gram precision)**
- **ROUGE-L F1 Score (Longest Common Subsequence)**
- **Jaccard Similarity**
- **Character  Word Count Ratios:**
  - Character Length Ratio and Difference
  - Word Count Ratio and Difference
- **Token Overlap Ratio**

## Models Used

**Regression Algorithms Implemented:**

- **Linear Regression:**
- **Ridge and Lasso Regression:**
- **ElasticNet**
- **Support Vector Regression (SVR):** Performed best with MAE = 0.393, capturing non-linear patterns effectively.
- **Random Forest Regressor**
- **Gradient Boosting Regressor**

**Ensemble Learning:**

- **Voting Regressor:** Combined predictions from top models (SVR, RF, GB) to boost robustness.

# Model Evaluation

| Model | MSE | MAE | $R^2$ |
|---|---|---|---|
| Linear Regression | 0.281 | 0.453 | 0.378 |
| Ridge Regression | 0.281 | 0.453 | 0.377 |
| Lasso Regression | 0.452 | 0.606 | -0.001 |
| ElasticNet | 0.452 | 0.606 | -0.001 |
| SVR | 0.266 | 0.393 | 0.412 |
| Random Forest | 0.297 | 0.454 | 0.343 |
| Gradient Boosting | 0.306 | 0.470 | 0.322 |
| Ensemble Model | 0.272 | 0.431 | 0.399 |

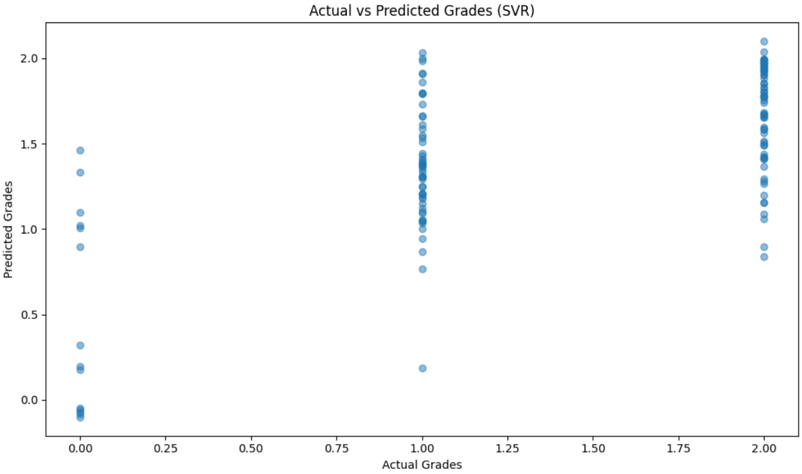
Table: Model Evaluation Metrics

# Comparison of Results: Best Model vs Paper

| Metric | SVR | Results from Paper |
|---|---|---|
| MSE | 0.266 | 0.2727 |
| MAE | 0.393 | 0.4644 |
| $R^2$ | 0.412 | 0.67 |

Table: Comparison of key metrics between the best model and results reported in the paper.

# Visualizations




Prediction vs Actual Grades

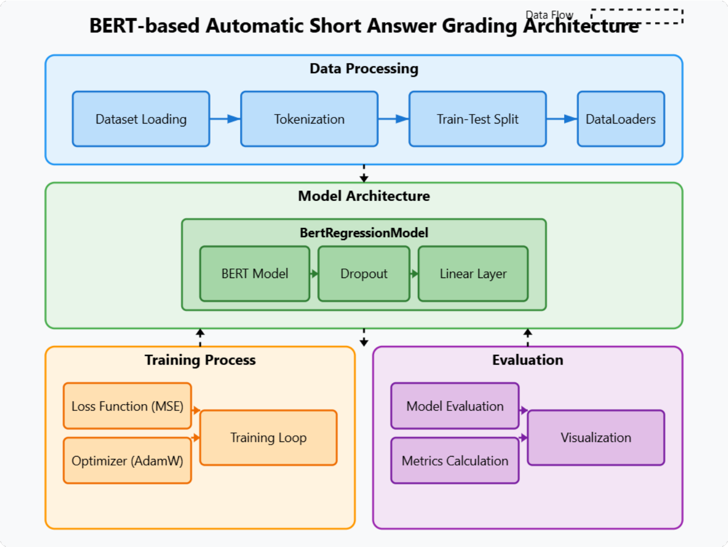
# Objectives (BERT Methodology)

1. **Develop a BERT-based ASAG System**: Fine-tune a pre-trained BERT model for grading short student responses.
2. **Leverage Contextual Embeddings**: Use BERT's [CLS] token output to capture the full meaning of student and reference answers.
3. **Build a Regression Head**: Add a linear regression layer on top of BERT to predict continuous grades.
4. **Optimize and Fine-tune the Model**: Train using MSE loss, AdamW optimizer, and gradient clipping for stability.
5. **Evaluate with Strong Metrics**: Assess model performance with MSE, MAE, and $R^2$ scores.

# Why BERT for ASAG?

- Contextualized Word Representations:BERT processes both the student and reference answers in context, enhancing semantic understanding.
- Bidirectional Understanding: Reads text both left-to-right and right-to-left, providing better contextual knowledge.
- Fine-Tuned for ASAG Task:Pretrained on a large corpus, then fine-tuned for predicting grades, adapting to the nuances of student answers.

# BERT Regression Architecture



# BERT Tokenization & Embeddings

- Jointly tokenize `[CLS] ref [SEP] student [SEP]` via `BertTokenizer`.
- Pad / truncate to 256 tokens.
- Feed into `bert-base-uncased` to get **[CLS]** pooled output.

# BERT Regression Model Architecture

**Model Components:**

- **BERT Encoder:**
  - Pretrained `bert-base-uncased` model to extract deep semantic features from the combined reference and student answers.
- **Dropout Layer:**
  - Defined with 30% probability to prevent overfitting
- **Linear Regression Layer:**
  - A fully connected layer that maps BERT's output (768 dimensions) to a single scalar (predicted grade).

**Forward Pass Logic:**

1. Input token IDs and attention masks are passed to BERT.
2. Extract the `[CLS]` pooled output representing the entire input.
3. Apply dropout to the pooled output.
4. Pass the output through a Linear layer to predict the grade.

# BERT Model Training

- Optimizer: `AdamW`, LR $= 2 \times 10^{-5}$
- Loss: MSE
- Epochs: 10, Batch size: 16
- Gradient clipping (max_norm=1.0), dropout ($p = 0.3$)

# BERT Model Evaluation

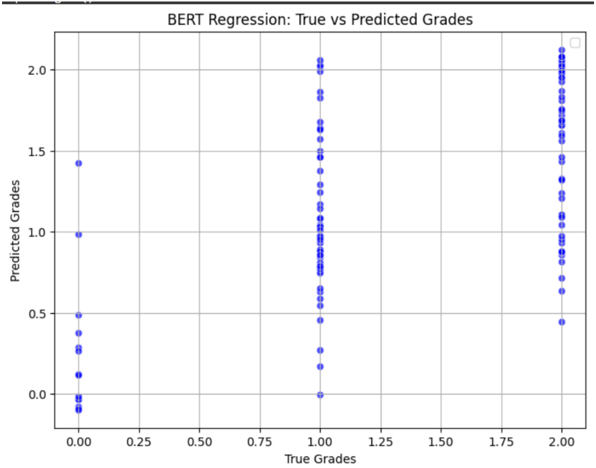
- **MSE:** 0.189 **MAE:** 0.300 **$R^2$:** 0.581

# Comparison of Results: Best Model vs Paper

| Metric | BERT | Results from Paper |
|---|---|---|
| MSE | 0.189 | 0.2727 |
| MAE | 0.300 | 0.4644 |
| $R^2$ | 0.581 | 0.67 |

Table: Comparison of key metrics between the best model and results reported in the paper.

# BERT Model Visualizations




(a) True vs. Predicted Grades

# Conclusion

- ASAG system delivers fair, consistent, and scalable grading.
- Machine Learning models and BERT both improved prediction accuracy.
- BERT further enhanced semantic understanding beyond surface-level features.
- Results show lower error rates and strong correlation between true and predicted grades.

# Future Enhancements

- Expand dataset size and diversity for better model generalization.
- Explore more advanced models like RoBERTa, DeBERTa, and GPT-based architectures.
- Improve preprocessing to handle grammar errors, spelling mistakes, and informal language.
- Focus on explainable AI techniques for transparent and fair grading.
- Implement active learning to continuously improve model based on human feedback.

# Thank You!

Questions?

Let's talk about grading smarter.