The Machine Learning Canvas

Designed for: Auto grading IELTS Essay | Designed by: Group 7 | Date: Aug 18, 2023 | Iteration: 3

Decisions

Identifying Decision Points:
Providing instant feedback to candidates after they submit their writing responses.

Assisting instructors in allocating personalized guidance to candidates.

Helping educational institutions track overall proficiency trends.

Real-time Predictions:
Ensuring that the predictions generated by the system are delivered in a timely manner. Candidates and educators rely on instant or near-real-time feedback for effective decision-making.

End-User Value Creation:
For Candidates: Instant feedback on strengths and areas for improvement in their writing responses.

For Instructors: Efficient allocation of guidance based on identified weaknesses in candidates' writing.
For Institutions: Insightful proficiency trends to inform curriculum adjustments.

Value-Added Steps:
Candidate Feedback: Develop a user-friendly interface to deliver feedback with suggestions for improvement.
Instructor Support: Provide instructors with clear reports on candidates' strengths and areas for focus.
Institutional Insights: Generate periodic reports to highlight proficiency trends and inform curriculum adjustments.

Interpretability and Explanation:
Ensure the system provides explanations for its predictions. Transparency and interpretability are essential for users to understand the basis of the grading decisions and trust the system.

Feedback Loop:
Establish a feedback mechanism for users to provide input on the system's predictions and recommendations.

ML task

Input: IELTS writing responses in textual format.

Associated prompts or topics for each writing response (to guide the grading process).

Output to Predict: Grade labels representing different proficiency levels (e.g., 0-9, where 0 indicates low proficiency and 9 indicates high proficiency).

Type of Problem: Multi-class classification task.

In this project, the aim is to leverage Large Language Models (LLMs) to automatically grade IELTS writing responses based on their quality and proficiency level. LLMs are pre-trained on vast amounts of text data, allowing them to capture intricate language features and nuances.

The input consists of the written responses and the corresponding prompts/topics. Instead of engineering traditional features, LLMs automatically generate contextualized embeddings for the input text. These embeddings encapsulate semantic information, grammatical structures, vocabulary usage, and coherence.

The problem remains a multi-class classification task since the goal is to predict grade labels from a range of possible values. However, LLMs bring a unique advantage by inherently understanding the linguistic context and subtleties present in the text.

Value Propositions

Who: The end users of the predictive system are IELTS candidates, language instructors, and educational institutions.

What: We are striving to provide an automated IELTS grading solution that offers accurate and consistent evaluation of writing responses for IELTS candidates. This system aims to alleviate the manual grading burden on instructors and provide candidates with instant and unbiased feedback on their writing skills.

Why: The objectives we are serving include:

- Efficiency and Time Savings
- Instant Feedback
- Consistency and Fairness
- Scalability
- Personalized Learning

Data Sources

Open Datasets: We can use datasets from platforms like Kaggle. There might be publicly available datasets that contain IELTS writing responses and their grades.

Educational Websites: Some educational websites provide practice writing prompts and sample responses for IELTS. We might be able to scrape this data or request permission to use it for the project.

Collecting Data

Data Sources: The primary source of data for this project will likely be external datasets that contain IELTS writing responses along with their corresponding grades. These datasets can be collected from platforms like Kaggle or other educational resources that provide sample IELTS responses.

Collecting Data: The project will involve collecting a diverse range of IELTS writing responses. The inputs will consist of the written responses to various IELTS writing prompts, while the outputs will include the corresponding grades assigned to each response. The collected data should cover a broad spectrum of writing styles, topics, and proficiency levels to ensure the model's effectiveness in grading different types of responses accurately.

The data collection process might involve steps such as:

Data Exploration: Research and identify suitable datasets that contain IELTS writing responses and grades.

Data Cleaning: Preprocessing the data to remove any irrelevant information, formatting inconsistencies, or errors that could affect the quality of the dataset.

Labeling: Ensuring that each writing response is associated with the correct grading label. This might involve manual labeling or using existing labeled datasets.

Balancing: Aiming to maintain a balanced representation of different grades and writing styles within the dataset to avoid bias in the model's training.

Quality Control: Implementing quality control measures to ensure that the collected data is accurate, representative, and free from biases.

Ethical Considerations: Ensuring that the collected data respects privacy and complies with relevant data

This loop helps improve the system's effectiveness over time.

usage regulations.

Making Predictions

Prediction Frequency: Ideally, predictions should be available as soon as a candidate submits their response for grading.

Assessing the trade-off between prediction frequency and computational resources available. If real-time predictions are not feasible due to processing time, batch prediction updates at regular intervals.

Featurization Time: With the use of LLMs, featurization is often minimal or implicit, as LLMs handle text inputs directly. The model takes the raw text as input and generates predictions without a separate featurization step.

Prediction Latency: LLMs like GPT-3 can provide predictions in near real-time. The latency is typically in the order of seconds.

Feedback Loop: Collecting feedback from users, educators, and institutions regarding the usefulness and timeliness of the predictions. Continuous feedback helps in fine-tuning the prediction process and optimizing the prediction workflow.

Offline Evaluation

The goal here is to assess how accurately the automated essay grading system predicts human-assigned scores on unseen essays.

Data Preparation and Split.

Validation on Test Set.

Scoring Comparison: Comparing the predicted scores from the model with the human-assigned scores.

Testing relevant evaluation metrics specifically suited for grading tasks, such as weighted accuracy, quadratic weighted kappa, or even specialized metrics used in the educational assessment domain.

Interpreting the Model's Predictions: Analyze cases where the model's predictions diverge significantly from human scores. This analysis can lead to identifying potential challenges and areas of improvement.

Model Fine-Tuning and Iteration:

Addressing Overfitting and Generalization: Transformers are prone to overfitting due to their large number of parameters.

Ensuring that the model generalizes well to new essays by monitoring its performance on an unseen validation/test set.

Features

Instead of manual engineering features, we'll be working with the representations learned by the language model itself. The focus shifts towards prompt engineering, fine-tuning, and optimizing the decoding strategy to align the model's output with accurate grading.

Building Models

Model Building and Update Strategy:

Initial Model Building: Start by training the initial LLM-based grading model using the labeled dataset you've collected.

Fine-tune the LLM on the task of IELTS grading. This involves exposing the model to IELTS writing prompts and their corresponding grades during training.

Continuous Monitoring: Deploy the initial model and monitor its performance on new, real-world IELTS writing responses.

Collect feedback from users and educators regarding the model's accuracy and areas for improvement.

Incremental Updates: Depending on the frequency of new data availability and the rate of change in IELTS writing patterns, plan for incremental updates to the model.

Data Analysis: Regularly analyze the new data collected to understand any shifts in writing styles, grading criteria, or other relevant factors.

Identify any emerging trends or patterns that might warrant model adjustments.

Deciding Update Frequency: Determine the appropriate frequency of model updates based on the rate of change in IELTS writing patterns and the resources available for model training.

Live Evaluation and Monitoring

Tracking Metrics:

Tracking Metrics:
User Satisfaction: Monitor user satisfaction with the automated grading system. This can be measured through user surveys, feedback forms, or sentiment analysis of user reviews.
Accuracy Metrics: Continue tracking metrics that measure the accuracy of the system's predictions. For instance, you can monitor the correlation between the model's predicted scores and human-assigned scores for a subset of essays.
Feedback Collection: Gather feedback from users who interact with the system. Analyze user interactions, queries, and support requests to identify potential issues and areas for improvement.

Business Metrics:

- **Efficiency**: Measuring the efficiency gains of using the app over manual grading. Calculate the time saved for both teachers and students.
- **Scalability**: Evaluating the system's scalability by monitoring its performance as the number of essays to be graded increases.
- **Cost Reduction**: Quantifying the cost savings achieved by reducing the need for human graders.
- **User Engagement**: Tracking user engagement with the app. Are users using the system frequently? Are they achieving better results?

- Scheduling regular model updates to ensure that the system maintains its desired level of performance.
- Keeping an eye on potential biases and ethical concerns that may emerge in the live system. Continuously evaluating the system's fairness and potential for bias against different groups of users.
- Providing explanations about the grading process, how the system works, and what to expect.