Capstone Project Weekly Report

Date: 13/10/2024

Project Details:

1. Sponsor Company: Prepshala.Pvt.LTd

2. Project Title: An adaptive screening tool using Generative AI and LLMs for the assessment of language and personality traits

Note: All the fields in the form are required.

Project Milestones:

Progress made in the current week and contribution from individual team members:

During the current week, significant progress was made in refining our assessment models based on feedback received from both the sponsor and faculty mentors. We initiated the week by presenting the models to the sponsor mentor, who commended the team for the work done but noted that the Root Mean Squared Error (RMSE) was higher than expected. The sponsor suggested re-evaluating the model using the original essay texts rather than the summarized versions.

To address this, we utilized the full dataset from Hugging Face, consisting of 9,806 records, and performed thorough Exploratory Data Analysis (EDA) to understand more about dataset. This included examining the lengths of the essays and the distribution of band scores. We discovered that most essays were between 200 and 400 words in length, while the summarization model condensed these to around 100-150 words. Following the EDA, we cleaned the dataset for further processing.

Taking the sponsor's feedback into consideration, we developed four distinct models:

1. Model 1 trained on summarized essays.
2. Model 2 trained on full-length essays.
3. Model 3 trained on full-length essays with rounded band values.
4. Model 4 trained on summarized essays with rounded band values.

To enhance model performance, we implemented hyperparameter tuning techniques, including cross-fold validation, regularization, and dropout. we were able to **significantly reduce the Mean Absolute Error (MAE) and RMSE** across all models, achieving much better performance metrics.

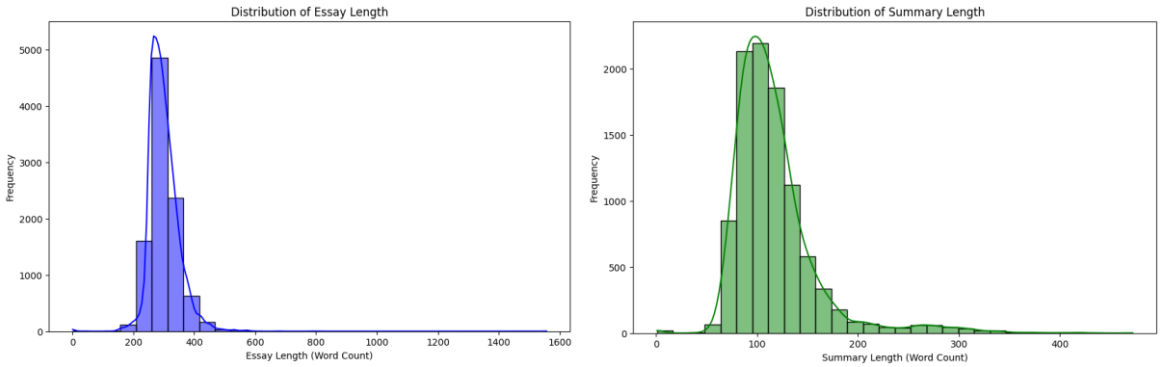
Following are the key results of the models:

- **Model 1:** Final Test Loss: 0.8061, Final Test MAE: 0.5308, Final Test RMSE: 0.6828
- **Model 2:** Final Test Loss: 3.6535, Final Test MAE: 1.5020, Final Test RMSE: 1.8137
- **Model 3:** Final Test Loss: 3.0842, Final Test MAE: 1.3587, Final Test RMSE: 1.6687
- **Model 4:** Final Test Loss: 0.4499, Final Test MAE: 0.3080, Final Test RMSE: 0.4429

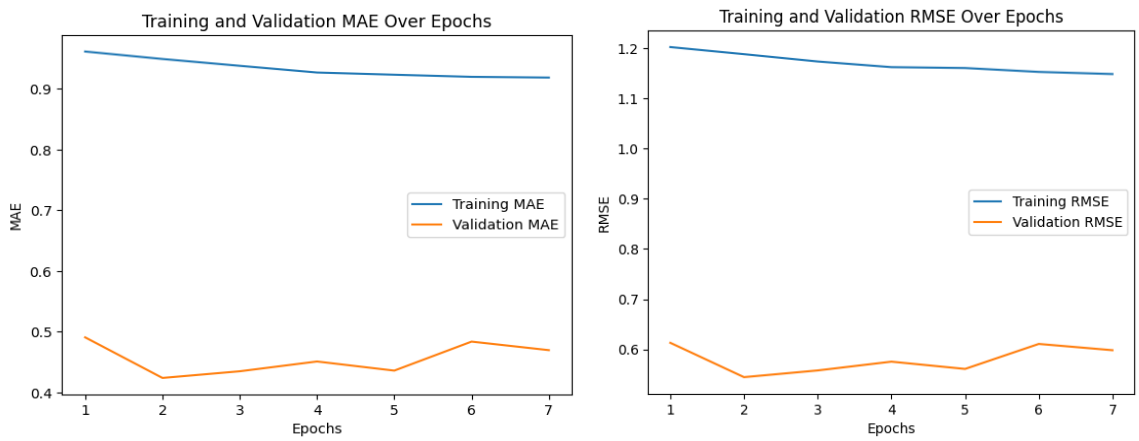
We split the dataset into 3 sets training, validation, and test sets. While the essay and summarization models showed only minor differences in MAE and RMSE on the training and validation sets, the summarization models consistently outperformed the essay-based models on the test set, demonstrating better performance in predicting IELTS band scores.

Additionally, we connected internally as a team to review the progress and demo of the developed models. We also discussed the approach for the next module, which focuses on the listening module. Identified a few datasets and approaches.

All team members actively contributed to model development, EDA, and planning for the next phase of the project.



EDA: Distribution of Essay and Summaries.



Model 1: Results.

```
Prompt (Not Evaluated): The ocean is ____. (Please fill in the blank with **one word**.)
Your response: blue.
Prompt: Nowadays people use social media to keep in touch with others and be aware of the news. Do the advantages of this outweigh the disadvantages? (Answer in
Your response: Social media has become an integral part of modern life, allowing people to stay connected with friends and family, as well as to keep up with the
WARNING:absl:Compiled the loaded model, but the compiled metrics have yet to be built. `model.compile_metrics` will be empty until you train or evaluate the mode
WARNING:absl:Compiled the loaded model, but the compiled metrics have yet to be built. `model.compile_metrics` will be empty until you train or evaluate the mode
WARNING:absl:Compiled the loaded model, but the compiled metrics have yet to be built. `model.compile_metrics` will be empty until you train or evaluate the mode
WARNING:absl:Compiled the loaded model, but the compiled metrics have yet to be built. `model.compile_metrics` will be empty until you train or evaluate the mode
1/1 ━━━━━━━━━━━━━━━━━━━━ 2s 2s/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 179ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 174ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 186ms/step
Model 1 - Response 1 (Dataset Question) predicted band score: 5.5
Model 2 - Response 1 (Dataset Question) predicted band score: 5.0
Model 3 - Response 1 (Dataset Question) predicted band score: 5.0
Model 4 - Response 1 (Dataset Question) predicted band score: 6.0
```

Evaluation on user's response text.

Tasks to finish in next week:

1. **Model Demonstration and Feedback:** Present the current model to the project sponsor for review. Based on their feedback, make the necessary adjustments to optimize the model's performance. Discuss and finalize the approach for the listening module.
2. **Consultation with Project Mentor:** Arrange a meeting with the project mentor to showcase the progress made so far. Gather their insights for further refinement and finalize the approach for the listening module.
3. **Develop the Listening Module:** Create and implement a model for the listening module.
4. **Enhance the Writing Module:** Improve the writing module based on feedback from both the professor and the sponsor mentor. Plan to add relevancy feature to the module.

Updates/MoM from Sponsor and Faculty Mentor:

Note: **It is expected that you have at least one weekly connect with the faculty mentor and sponsor.** If you were not able to schedule meetings with the sponsor or faculty mentor in the current week, please mention the reason for your inability to meet with the Sponsor or Faculty Mentor.

Updates/MoM from Sponsor:

We had a 45-minute call with the sponsor on 7th October 2024, during which we demonstrated the models developed for the writing module. Below are the key points discussed:

1. **Feedback on RMSE:** The sponsor mentor highlighted the higher RMSE observed in the current model and advised that summarization might alter the tone and phrasing, potentially affecting accuracy. Now, We have successfully achieved **RMSE of 0.4.**
2. **Comparison of Models:** The mentor suggested applying the models directly to the essays (without summarization) to predict bands based on essay length, then applying them to the summarized version. The results from both approaches should be compared to assess the model's performance differences.
3. **Rounded Bands and Metrics:** It was recommended to round off the predicted bands rather than using continuous values, and then recalculate RMSE and MAE for better accuracy.
4. **Classification Model:** There was a discussion on whether it would be beneficial to apply a classification model instead of a regression-based approach. The mentor encouraged reducing the RMSE further, regardless of the approach used.
5. **Listening Module Approach:** We also discussed the potential approach for developing the listening module, including key considerations and strategies.

Updates/MoM from Faculty Mentor:

We had a 45-minute call with the faculty mentor on 8th October 2024, during which we demonstrated the current model and shared the feedback received from the sponsor mentor. Below are the key takeaways:

1. **Model Performance:** The faculty mentor appreciated the results achieved with the current model, particularly the MAE and RMSE metrics. The professor mentioned that an RMSE of 1.5 is acceptable and not a significant concern. However, We have now achieved a RMSE of 0.4
2. **Summarization Discussion:** We discussed whether summarization should be applied. The professor noted that since the dataset contains essays of varying lengths, the model would naturally be trained on shorter lengths as well. If length is a concern, the model will adapt accordingly.
3. **Summarization Model Review:** Upon reviewing the summarization model, the professor found that our model maintained the original tone and summarized the content effectively. The professor suggested applying models to the full essays and then comparing the results after summarization. We have already implemented this approach.
4. **Listening Module:** We discussed various datasets and approaches for developing the listening module. The professor recommended starting with the data we currently have while continuing to explore additional datasets. The professor also offered to assist in finding relevant datasets for this module.

Challenges:

Mention any technical and non-technical challenges that you faced during the current week that hindered your project progress. Enter "NA" if you didn't face any challenges.

Technical Challenges:

1. One of the primary challenges we encountered was that all available datasets contained full-length IELTS standard essays, which were much longer than the required 5-10 lines. Its challenging to find the dataset for 5-10 essays dataset.
2. Ensuring that summarization maintains the original context and tone without significant loss of meaning has been challenging.
3. Difficulty in finding high-quality IELTS-level datasets for each module, particularly for the listening and speaking modules.
4. Successfully addressed model overfitting and managed to reduce the RMSE.

Non-Technical Challenges:

NA

Mention any other queries/challenges regarding the project that you want to highlight:

NA