|  |  |
| --- | --- |
| Internship Project Title | TCS ion RIO-125: Automate Detection and Recognition of Grammatical Errors |
| Name of the Company | TCS iON |
| Name of the Industry Mentor | Vikash Singh |
| Name of the Institute | P.B. College of Engineering |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Start Date | End Date | Total Effort (hrs.) | Project Environment | Tools used |
| 29-03-2024 | 13-05-2024 | 35 | Google Collab | Python3 |

**TABLE OF CONTENT**

|  |  |
| --- | --- |
| **S.NO** | **CONTENT** |
| 1. | Acknowledgements |
| 2. | Objective |
| 3. | Introduction / Description of Internship |
| 4. | Internship Activities |
| 5. | Approach / Methodology |
| 6. | Model |
| 7. | Exceptions / Exclusions |
| 8. | Charts, Table, Diagrams |
| 9. | Algorithms |
| 10. | Challenges & Opportunities |
| 11. | Risk Vs Reward |
| 12. | Reflections on the Internship |
| 13. | Recommendations |
| 14. | Outcome / Conclusion |
| 15. | Enhancement Scope |
| 16. | Link to code and executable file |
| 17. | Research questions and responses |

**1. ACKNOWLEDGEMENTS**

I am conveying my sincere gratitude towards my Industry Mentor, Vikash Singh helping me throughout this project till now and providing me this wonderful platform to complete this project. I am thankful for answering my queries at every phase of the project. I also want to thank all my friends who helped me with valuable suggestions during this project.

**2. OBJECTIVE**

The objective of this internship was to develop machine learning and deep learning algorithms for automating the detection and recognition of grammatical errors in text data. The aim was to preprocess input text with or without grammatical errors, detect these errors, and then correct them according to the grammatical rules of the language.

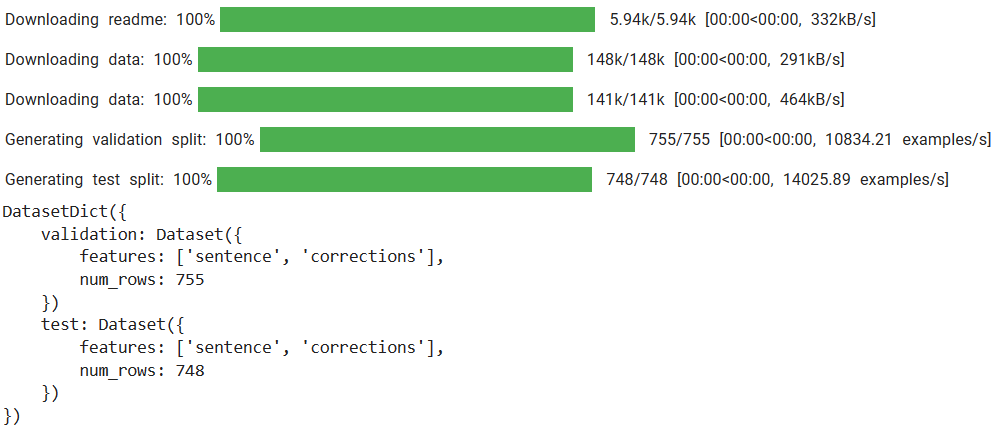
**3. INTRODUCTION / DESCRIPTION OF INTERNSHIP**

The internship focused on leveraging natural language processing techniques and deep learning models to build a system capable of identifying and correcting grammatical errors in text. This involved data collection, preprocessing, model training, evaluation, and the development of a user-friendly interface for real-time correction.

**4. INTERNSHIP ACTIVITIES**

The internship activities included:

* Researching and understanding existing grammatical error correction systems and datasets.
* Data collection using the JFLEG dataset, which provides correct and incorrect sentence pairs.



* Preprocessing the data by tokenizing paragraphs into sentences and removing extra spaces.
* Exploratory data analysis (EDA) to understand the distribution of sentence lengths, parts of speech, and frequent words in correct and incorrect sentences.
* Training a T5-based transformer model using the fine-tuning approach to learn grammatical error correction patterns.
* Developing utility functions for sentence correction, highlighting errors, and creating a graphical user interface (GUI) for interactive correction.
* Conducting model validation and evaluation using a separate validation dataset.
* Creating a comprehensive project report documenting the entire process, challenges faced, and recommendations for future improvements.

**5. APPROACH / METHODOLOGY**

The methodology followed a systematic approach:

* Data Collection: Utilizing the JFLEG dataset for training and validation.
* Data Preprocessing: Removing extra spaces, and performing EDA.
* Model Training: Fine-tuning the T5 transformer model for grammatical error correction.
* Model Evaluation: Validating the model's performance on a separate validation dataset.
* User Interface Development: Creating a GUI using Gradio for real-time correction.
* Project Report: Documenting the entire process, challenges, outcomes, and recommendations.

**6. MODEL**

The T5 (Text-to-Text Transfer Transformer) model is a powerful natural language processing (NLP) model developed by Google Research. The "base" version refers to a specific size variant of the T5 model architecture, which is smaller compared to larger variants like "large" or "3B" (3 billion parameters). Here are some key points about the T5-base model:

* **Architecture:** T5-base is based on the transformer architecture, specifically the encoder-decoder transformer architecture. This architecture consists of encoder layers that process input text and decoder layers that generate output text.
* **Parameter Size:** The T5-base model typically has around 220 million parameters. This size is considered moderate compared to larger variants like T5-large (770 million parameters) or T5-3B (3 billion parameters).
* **Pre-training:** T5-base is pre-trained on a large corpus of text data using unsupervised learning objectives, such as denoising autoencoding, masked language modeling, and sequence-to-sequence prediction tasks. This pre-training phase helps the model learn general language understanding.
* **Fine-tuning:** After pre-training, the T5-base model can be fine-tuned on specific downstream tasks, such as text summarization, translation, question answering, and grammatical error correction. Fine-tuning adapts the model to perform well on these specific tasks by adjusting its parameters.
* **Text-to-Text Framework:** One of the distinctive features of T5 is its "text-to-text" framework, where both input and output are treated as text. This allows for a unified approach to various NLP tasks, making it easier to apply the same model architecture across different tasks.
* **Versatility:** Due to its text-to-text framework and fine-tuning capabilities, T5-base is versatile and can be applied to a wide range of NLP tasks. It has achieved state-of-the-art results on benchmarks like the GLUE benchmark for general language understanding.
* **Usage:** Researchers and practitioners often use T5-base as a starting point for NLP projects, experimenting with fine-tuning on specific datasets and tasks to achieve customized performance.
* **Availability:** T5-base is available as part of the Hugging Face Transformers library, making it accessible for developers and researchers to use in their NLP projects.

**7. EXCEPTIONS / EXCLUSIONS**

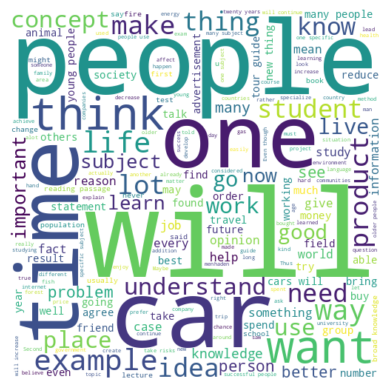
Some exceptions or exclusions from the internship scope may include:

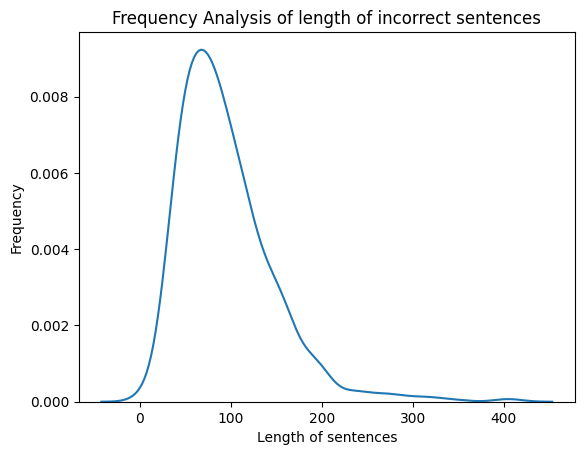
* Advanced error types beyond grammatical errors (semantic errors, contextual errors, etc.).
* Limitations in the model's ability to handle complex sentence structures or domain-specific language.
* Performance variations based on input data quality and diversity.

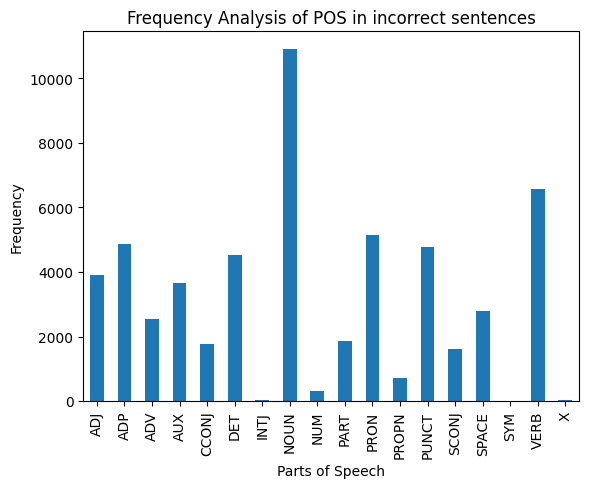
**8. CHARTS, TABLE, DIAGRAMS**

Charts, tables, and diagrams were used extensively during the internship:

Frequency analysis charts for sentence lengths, parts of speech, and word clouds.







**9**. **ALGORITHMS**

The key algorithms utilized were:

* Tokenization and fine-tuning of the T5 transformer model.
* Sentence preprocessing algorithms for data cleaning and normalization.
* Utility functions for error highlighting and correction.

**10. CHALLENGES & OPPORTUNITIES**

Challenges encountered during the internship included:

* Understanding and preprocessing the dataset effectively.
* Fine-tuning the T5 model with optimal hyperparameters.
* Developing an intuitive and efficient user interface for real-time correction.

Opportunities identified for future enhancements:

* Incorporating more diverse datasets to improve model generalization.
* Exploring advanced transformer models for better grammatical error correction.
* Integrating additional features such as contextual suggestions and error explanations.

**11. RISK VS REWARD**

The risk involved in the project included model performance limitations, data biases, and user interface usability issues. However, the rewards were significant, including the development of a functional grammatical error correction system, hands-on experience with deep learning and NLP techniques, and contributions to the research community.

**12. REFLECTIONS ON THE INTERNSHIP**

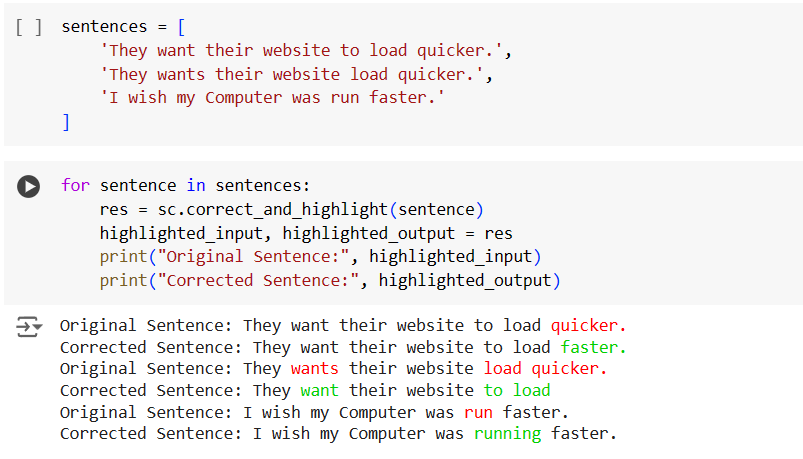
* The internship provided valuable insights into the field of natural language processing and machine learning.
* It enhanced my technical skills, problem-solving abilities, and project management capabilities.
* The collaborative environment and mentorship contributed to a rewarding learning experience.

**13. RECOMMENDATIONS**

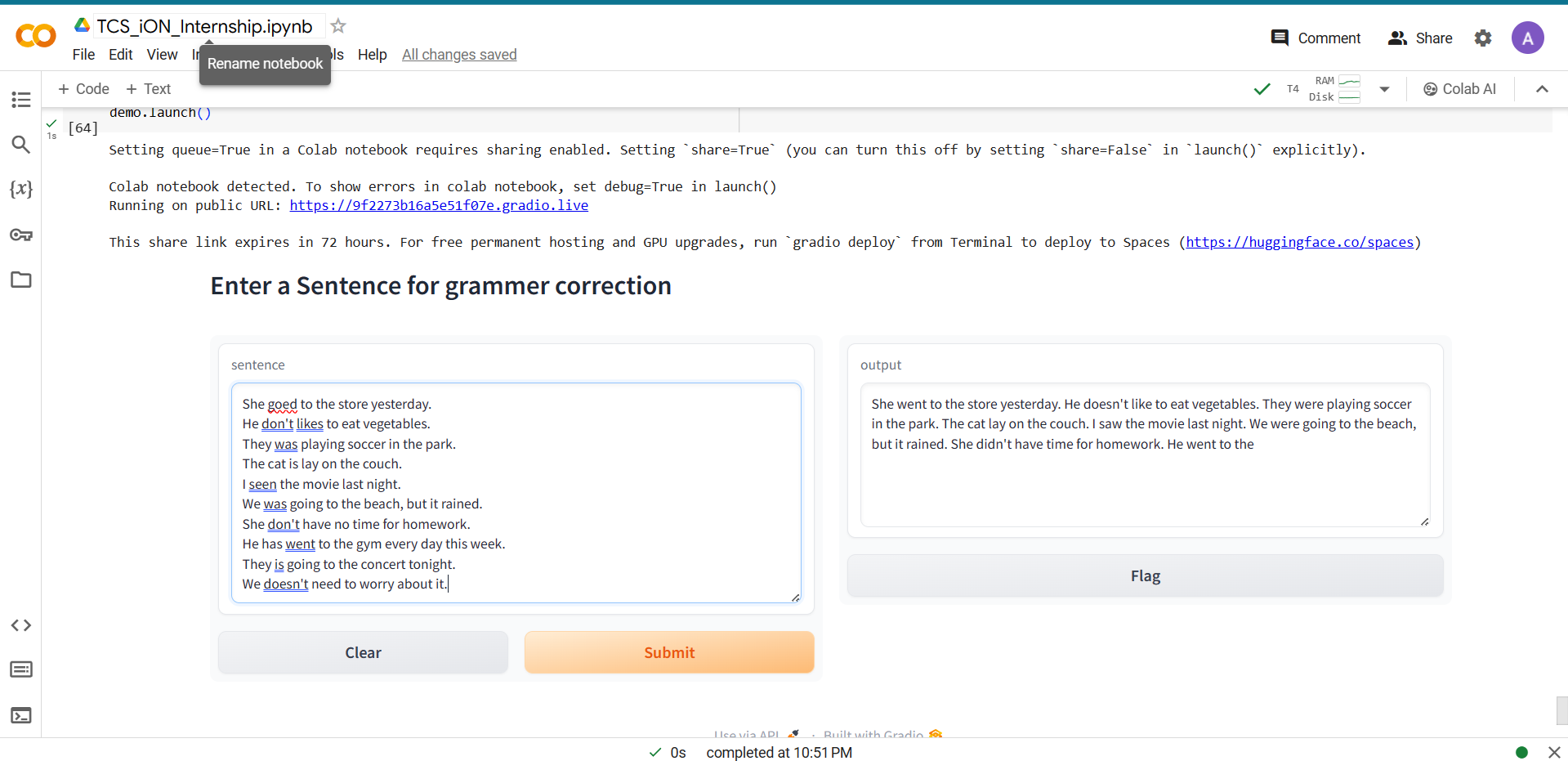
Based on the internship experience, I recommend: Continuous learning and exploration of advanced NLP and deep learning concepts. Collaboration with domain experts to enhance model accuracy and domain-specific corrections. Sharing insights and findings through research papers and community forums for knowledge dissemination.

**14. OUTCOME / CONCLUSION**

The outcome of the internship was the successful development of a grammatical error detection and correction system using the T5 transformer model. The system demonstrated promising results in correcting various grammatical errors in real-time text inputs.



Gradio provides a seamless and interactive platform for showcasing the grammar correction project, allowing users to input sentences and instantly receive corrected outputs with highlighted corrections, enhancing the project's accessibility and user experience.



**15. ENHANCEMENT SCOPE**

Future enhancements and areas for improvement include: Integration of contextual and semantic error detection capabilities. Fine-tuning the model with larger and more diverse datasets for improved accuracy. Enhancing the user interface with additional features such as error explanations and feedback mechanisms.

**16. LINK TO CODE AND EXECUTABLE FILE**

The code repository for the project can be accessed at

GitHub Link : [ArunK192002/Automate-detection-and-recognition-of-grammatical-errors (github.com)](https://github.com/ArunK192002/Automate-detection-and-recognition-of-grammatical-errors)

**17. RESEARCH QUESTIONS AND RESPONSES**

1. **Research Question:** What are the key challenges in developing a grammatical error correction system?

**Response**: The key challenges include dataset quality, model generalization, handling diverse error types, and real-time correction efficiency.

1. **Research Question:** How effective is the T5 transformer model in grammatical error correction?

**Response**: The T5 model shows promising results in correcting grammatical errors, but continuous fine-tuning and dataset enhancements are essential for optimal performance.

1. **Research Question:** What are the potential applications of automated grammatical error correction systems?

**Response**: Automated grammatical error correction systems have applications in education, content creation, language translation, and improving overall text quality in various domains.