



**AUTONOMOUS GRADING SYSTEM FOR  
SINHALA LANGUAGE ESSAYS OF  
GRADE 5 STUDENTS**

P.W.Maddumage

B.Sc. (Hons) Degree in Information Technology Specializing in  
Data Science

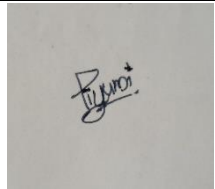
Department of Information Technology

Sri Lanka Institute of Information Technology  
Sri Lanka

February 2024

## DECLARATION

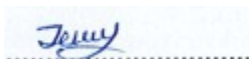
We declare that this is my own work, and this proposal does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other university or Institute of higher learning and to the best of our knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Name	Student ID	Signature
Maddumage P. W.	IT21007538	

The above candidates are carrying out research for the undergraduate Dissertation under my supervision.

Signature of the Supervisor:

Date:






Ms. Jenny Krishara

Signature of the Co-Supervisor:

Date:





Ms. Wishalya Thisara

## **ABSTRACT**

In the context of modern education, the demand for efficient and objective assessment methods is ever-growing. This research focuses on the development of an autonomous grading system tailored for Sinhala language essays written by Grade 5 students. The objective is to alleviate the time-consuming task of manual grading while maintaining a high level of accuracy and fairness in evaluating students' writing skills.

The proposed system employs natural language processing (NLP) techniques to analyze and assess various linguistic and structural aspects of Sinhala essays. It incorporates advanced machine learning algorithms to identify and evaluate components such as grammar, vocabulary, coherence, and overall content quality. To ensure cultural relevance and language-specific nuances, the system is specifically trained on a dataset comprising Sinhala language essays from Grade 5 students.

A significant sub-topic of this research involves the identification of native words within the essays and their replacement with generic words. This process not only enhances the system's ability to assess the fundamental language skills of students but also contributes to a standardized evaluation framework. By identifying native words, the system can address regional variations and promote a more universally applicable grading methodology.

The research addresses the challenges associated with Sinhala language processing and introduces innovative solutions to overcome them. The system aims to provide constructive feedback to students, aiding in their language development, and allowing educators to focus on personalized guidance rather than routine grading tasks.

The evaluation of the autonomous grading system involves a comprehensive comparison with manually graded essays to validate its effectiveness and reliability. The research contributes to the ongoing discourse on the integration of technology in education and highlights the potential of autonomous grading systems to enhance the efficiency and objectivity of assessment processes, particularly in linguistically diverse contexts like Sinhala language education.

**Key Words** - Essay grading · Natural language processing · Sinhala

## TABLE OF CONTENT

Declaration.....	i
Abstract .....	ii
Table of content.....	iii
List of Figures .....	iv
List of Tables.....	v
1. Introduction .....	1
2. Background and Literature Survey.....	3
3. Research Gap.....	6
4. Research Problem.....	7
5. Objectives .....	8
5.1 Primary Objectives .....	8
5.2 Specific Objectives .....	9
6. Methodology.....	10
6.1 System Architecture .....	13
6.1.1 Overall System .....	13
6.2 Data Collecting Techniques.....	14
6.3 Tools and Technologies.....	15
6.4 Commercialization plan.....	17
6.5 Budget .....	18
7. Project Requirements.....	19
7.1 Non-functional Requirements.....	19
7.2 Functional Requirements .....	19
7.3 System Requirements .....	19
8. Work Breakdown Chart.....	20
9. Gantt Chart .....	21
10. References.....	22

## **LIST OF FIGURES**

Figure 2.1	NLP layers and tasks	5
Figure 6.1	Diagram of overall system	13
Figure 6.2	Sample Data Demonstration	14

## **LIST OF TABLES**

Table 3.1 - Comparisons between former research and the systems	6
Table 6.5    Budget table	18

# 1. INTRODUCTION

The research landscape in the domain of autonomous grading systems has witnessed substantial exploration, particularly in the realm of natural language processing (NLP) and machine learning (ML). Prior works have delved into diverse languages, but the specific intricacies of Sinhala language essays, especially at the Grade 5 level, demand a unique approach. Understanding the existing body of research provides a foundation for our project, ensuring that our methodology aligns with and advances the current state of the art.

Relevant studies have investigated various aspects of NLP and ML in educational settings. Brown and Yule [1] emphasized the importance of conversational English analysis, a precursor to the nuanced understanding required for essay evaluation. Jurafsky and Martin [2] outlined the fundamentals of speech and language processing, offering insights into the complexities inherent in linguistic analysis. Additionally, Seneviratne and Adikari [3] conducted a comprehensive review of Sinhala language processing, highlighting the unique challenges posed by the language.

Mastering techniques in these areas is crucial for our project. Techniques such as syntactic and semantic analysis, sentiment analysis, and feature extraction play pivotal roles in comprehensively evaluating essays. Our system will draw on these techniques, adapting and enhancing them to suit the specific nuances of Sinhala language essays. Moreover, incorporating the identification and replacement of native words with generic words, as a sub-topic, addresses regional variations and contributes to the development of a more universally applicable grading methodology.

The state of the art in autonomous grading systems today involves a fusion of advanced NLP and ML algorithms to evaluate essays across various languages. However, the existing solutions may not be optimized for the complexities of Sinhala language structure and cultural nuances. [4] Our research aims to bridge this gap by developing a system that is specifically tailored for Grade 5 Sinhala language essays, ensuring a more accurate and culturally sensitive evaluation.

Previous attempts to solve similar problems have often been language-agnostic or limited to widely studied languages. Our approach distinguishes itself by being language-specific, focusing on Sinhala, and incorporating a sub-topic that identifies and replaces native words with generic ones. This not only enhances the system's accuracy but also contributes to a more standardized and universally applicable assessment process, addressing the limitations of existing approaches.

In summary, our research builds upon the wealth of knowledge in autonomous grading systems, leveraging relevant studies and techniques in NLP and ML. By focusing on Sinhala language essays of Grade 5 students and incorporating a sub-topic for native word identification, we aim to contribute a nuanced and culturally aware autonomous grading system that advances the state of the art in educational technology.



## **2. BACKGROUND AND LITERATURE SURVEY**

### **Background:**

The educational landscape is evolving with advancements in technology, prompting the exploration of innovative approaches to assessment and grading. Autonomous grading systems, driven by natural language processing (NLP) and machine learning (ML), have gained prominence in the assessment of written content. The need for efficient and objective evaluation is particularly pressing in linguistically diverse environments. Sinhala, as an example of such a language, poses unique challenges due to its distinct linguistic characteristics.

Educational researchers have long recognized the limitations of manual grading, including subjectivity, time consumption, and potential bias. The advent of autonomous grading systems seeks to address these challenges by leveraging computational linguistics to evaluate essays in a more standardized and objective manner. The focus on Grade 5 students is significant, as this age group marks a critical stage in language development, and early interventions can have lasting impacts.

### **Literature Survey:**

#### **1. NLP in Education:**

- Research by Brown and Yule (1983) [1] laid the foundation for understanding conversational English, emphasizing the importance of linguistic analysis. This work is fundamental in the context of essay evaluation, providing insights into language structure and coherence.

## 2. ML Applications in Language Processing:

- Jurafsky and Martin's work (2020) [2] in "Speech and Language Processing" provides a comprehensive guide to the application of ML in language-related tasks. Understanding these techniques is vital for developing an autonomous grading system that goes beyond mere syntax and considers semantic aspects of Sinhala essays.

## 3. Sinhala Language Processing Challenges:

- Seneviratne and Adikari (2016) [3] conducted a detailed review of Sinhala language processing, identifying challenges specific to the language. This review serves as a cornerstone for our project, guiding us in overcoming language-specific hurdles in the development of the grading system.

### Component: Identifying Native Words and Replacement:

#### 1. Cultural and Linguistic Sensitivity:

- The Sinhala Official Language Act No. 33 of 1956 provides a legal and cultural backdrop, emphasizing the importance of preserving the Sinhala language. Our sub-topic of identifying native words and replacing them with generic ones aligns with the cultural sensitivity required in developing autonomous grading systems for Sinhala essays.

#### 2. Previous Work on Native Word Identification:

- While generic native word identification methodologies exist, adapting them to Sinhala requires a nuanced approach. Literature addressing native word identification in other languages, such as English or Hindi, serves as

a reference point. Building on these methodologies, our sub-topic aims to create a language-specific solution for Sinhala.

By integrating these insights from the background and literature survey, our research aims to contribute a sophisticated autonomous grading system for Sinhala language essays of Grade 5 students. The focus on both general NLP and ML principles, as well as the intricacies of Sinhala language processing, ensures a holistic and culturally relevant approach to automated essay grading.

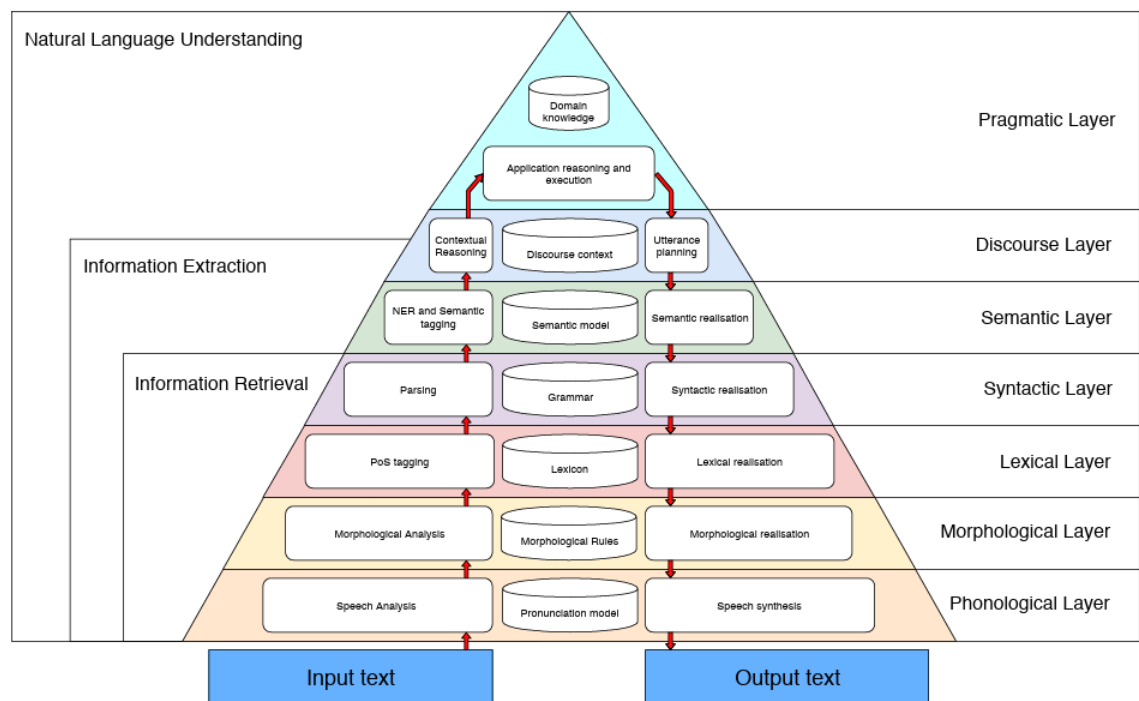


Figure 2.1 – NLP layers and tasks [9]

Source - Survey on Publicly Available Sinhala Natural Language Processing Tools and Research

### 3. RESEARCH GAP

	[1]“Named entity recognition for sinhala language”	[2] “Analysis of sinhala using natural language processing techniques”	[3]“An automated essay evaluation system using natural language processing and sentiment analysis”	[4]” Automated Essay Grading System using NLP Techniques”	“Dhara” Android Application
Integrate with mobile app	NO	NO	NO	NO	YES
Named Entity recognition	YES	YES	YES	YES	YES
Sentiment Analysis	YES	YES	YES	YES	YES
Targeting school students scope	NO	NO	NO	YES	YES
Using only Sinhala Language	YES	YES	NO	NO	YES

Table 3.1 - Comparisons between former research and the systems

Named Entity Recognition involves detecting and categorizing important information in text known as named entities. Named entities refer to the key subjects of a piece of text, such as names, locations, companies, events and products, as well as themes, topics, times, monetary values and percentages. Sentiment analysis is the process of analyzing digital text to determine if the emotional tone of the message is positive, negative, or neutral.

Type	Examples
Material	පුටු, ගෙය
Agentive	දුටින්තා, බීම
Common	ගොවියා, මිනිසා
Abstract	සුදු, උස
Proper	කොළඹ, අමර
Compound	කිරිබත් (කිරි+බත්), සුදුමල් (සුදු+මල්)

Table 3.2 Noun categorization

Form	Case	Singular		Plural
		Definite	Indefinite	
1	Nominative	පොත	පොතක්	පොත්
2	Accusative	පොත	පොතක්	පොත්
4	Dative	පොතට	පොතකට	පොත්වලට
5	Genitive	පොතේ	පොතක	පොත්වල
6	Locative	පොතේ	පොතක	පොත්වල
7	Instrumental	පොතෙන්	පොතකින්	පොත්වලින්
8	Ablative	පොතින්	පොතකින්	පොත්වලින්

Table 3.3 Examples for inflection of nouns

Source - Survey on Publicly Available Sinhala Natural Language Processing Tools and Research

## **4. RESEARCH PROBLEM**

Identifying words specific to different regions of Sri Lanka poses a multifaceted research challenge for the autonomous grading system. The linguistic diversity within the Sinhala language manifests in regional variations, resulting in the usage of words, phrases, or expressions unique to specific geographical areas. Addressing this challenge requires the system to undergo a rigorous training process that incorporates a diverse dataset reflecting the linguistic nuances prevalent in different regions of Sri Lanka. The system must be equipped with the ability to recognize not only formal language structures but also colloquialisms and regional dialects. Developing a robust algorithm for regional word identification demands a nuanced understanding of the sociolinguistic context, ensuring that the grading system can accurately distinguish and catalog region-specific lexicon within Grade 5 Sinhala language essays.

Once identified, the next research problem involves the systematic replacement of region-specific words with more general equivalents to facilitate a standardized evaluation process. This task requires the grading system to be equipped with a comprehensive repository of general words that can serve as substitutes for region-specific terms without compromising the essence or meaning of the content. The challenge lies in preserving the contextual relevance and coherence of the essays while adopting a more universally applicable vocabulary. The system needs to balance linguistic sensitivity to regional variations with the necessity for a standardized assessment framework. Developing an efficient and culturally respectful algorithm for word replacement constitutes a pivotal aspect of enhancing the objectivity and fairness of the autonomous grading system in evaluating Grade 5 Sinhala language essays across diverse regional contexts in Sri Lanka.

## **5. OBJECTIVES**

### **5.1 Primary Objectives**

- Create an essay grading application for Sinhala Language essays of grade 5 students. The system will be powered in following key sectors
- Identifying native words and replace them with more generic words
- Identifying difficult words which are using more complex letters
- Grammar rules check of sentences
- Evaluate a generalized answer

### **5.2 Specific Objectives**

- Identifying native words and replace them with generic words
- Conduct a comprehensive linguistic analysis to identify words that are specific to a different regions of Sri Lanka
- Create a database of native words by sourcing linguistic references, cultural dictionaries, and corpus data specific to Sinhala Language
- Named Entity Recognition (NER) refers to the process of identifying and classifying named entities, which are specific objects, individuals, locations, organizations, or any other entity with a proper name, in a given text. In the context of the research topic, "Name Entity Recognition of different native words," the focus is on identifying and classifying native words within Sinhala language essays.

- The objective of Name Entity Recognition is to automatically locate and categorize native words that hold significance within the context of Sinhala language and culture. Native words can include region-specific terms, cultural references, or words that might not have direct equivalents in other languages. The NER process involves analyzing the linguistic features of the text, such as syntax, semantics, and context, to determine which words fall into the category of native entities.
- For example, if the essay mentions a specific cultural practice, a traditional dish, or a local festival, the NER system would recognize these as native entities. The identification of such entities contributes to a deeper understanding of the cultural richness embedded in the essays. This recognition is crucial for subsequent analysis and processing, including the replacement of these native words with more generic terms, as discussed in the research problems earlier.
- In summary, Name Entity Recognition for different native words in the context of Sinhala language essays involves the automated identification and classification of specific entities that are integral to the cultural and linguistic fabric of the Sinhala language. This process is foundational to the development of an autonomous grading system that can account for and adapt to the diverse linguistic nuances present in Grade 5 Sinhala language essays

## 6. METHODOLOGY

### Corpus Collection:

Begin by assembling a comprehensive and diverse corpus of Grade 5 Sinhala language essays. This corpus should encapsulate the linguistic variations found in different regions of Sri Lanka, including colloquialisms, regional dialects, and culturally specific terms.

### Preprocessing:

Conduct thorough text preprocessing, including tokenization, stemming, and lemmatization, to standardize the format and structure of the essays. This step facilitates a more accurate analysis of linguistic features.

### Named Entity Recognition (NER):

Implement a Named Entity Recognition system tailored to identify native words within the essays. Train the NER model on the preprocessed corpus, emphasizing the recognition of entities specific to Sinhala language and culture. Fine-tune the model to distinguish regional variations in native words.

### Creation of Native Word Dictionary:

Develop a comprehensive dictionary containing identified native words along with their respective regions or cultural contexts. This dictionary serves as a reference for the subsequent step of word replacement.



#### Generic Word Repository:

Establish a repository of generic words that can serve as replacements for the identified native words. These generic words should be carefully chosen to maintain the overall coherence and meaning of the essays while providing a more standardized vocabulary.

#### Word Replacement Algorithm:

Design an algorithm that systematically replaces native words with their corresponding generic counterparts. This algorithm should consider contextual factors, ensuring that the replacement does not compromise the semantic integrity of the essays. Fine-tune the algorithm based on linguistic analysis and feedback from language experts.

#### Validation and Testing:

Validate the effectiveness of the methodology through rigorous testing on a separate set of Grade 5 Sinhala language essays. Evaluate the system's accuracy in identifying native words and its ability to replace them with generic words without distorting the original meaning.

#### Refinement and Iteration:

Iterate and refine the methodology based on the results of the validation phase. Incorporate feedback from language experts and educators to enhance the system's linguistic sensitivity and cultural relevance.

#### Integration with Grading System:

Integrate the developed methodology seamlessly into the larger autonomous grading system for Grade 5 Sinhala language essays. Ensure that the identification and replacement of native words contribute to a more standardized and culturally neutral assessment process.

#### Evaluation and Benchmarking:

Evaluate the overall performance of the grading system, considering factors such as efficiency, objectivity, and cultural appropriateness. Benchmark the system against manually graded essays to validate its effectiveness in providing fair and accurate assessments.

This comprehensive methodology aims to address the specific challenges posed by regional variations in Sinhala language essays, ensuring that the autonomous grading system can identify and replace native words in a manner that enhances the objectivity and cultural sensitivity of the assessment process

## 6.1 System Architecture

### 6.1.1 Overall System

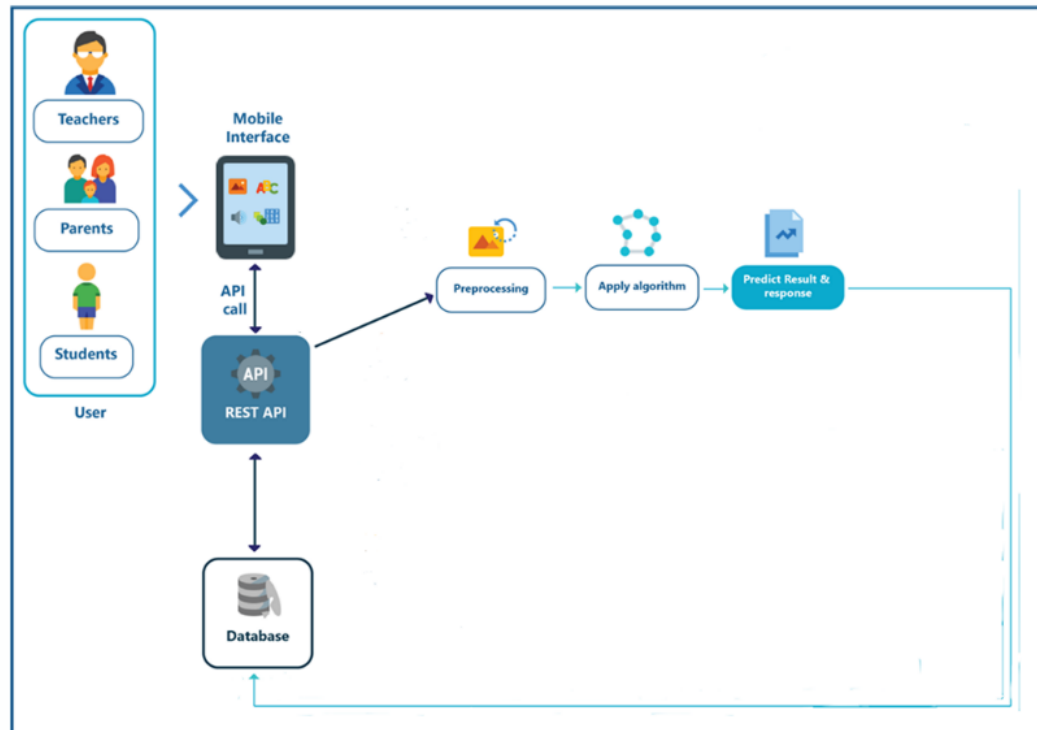


Figure 6.1 – Diagram of overall system

## 6.2 Data Collecting Techniques

Through several ways, data will be collected. These various ways are mentioned below.

- By contacting external supervisor (from a school) – Data will be collected from teachers of grade 5
- Online Survey – Datasets collected before for research purposes
- Online government websites – Download grade 5 syllabus and words
- Research papers – More local and international research papers will be studied while obtaining data
- Reading books – Recognized book will be read to understand the available data with them
- Meet resource persons – Lecturers, research officers, tuition teachers will be meet to gather data which available with them

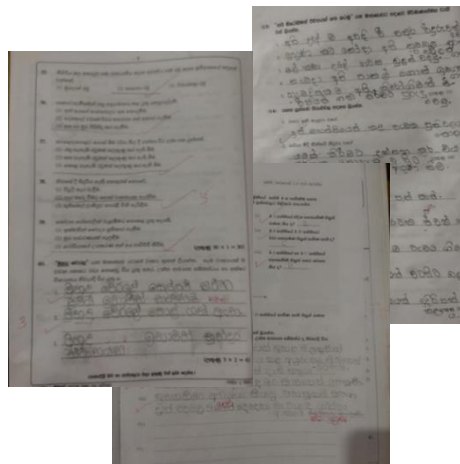


Figure 6.2 – Sample Data Demonstration

### 6.3 Tools and technologies

To identify native words and replace them with generic words in Sinhala language essays, you can leverage a combination of natural language processing (NLP) tools and technologies specifically adapted for the Sinhala language. Here are some tools and technologies that can be utilized:

Sinhala NLP Libraries:

Sinhala Language Processing Toolkit (SinhalaNLP): This toolkit provides various NLP tools for Sinhala, including tokenization, stemming, and part-of-speech tagging. It can aid in pre-processing and linguistic analysis.

Named Entity Recognition (NER) Tools:

spaCy for Sinhala (spacy-sinh): spaCy is a popular NLP library, and the spacy-sinh extension provides support for Sinhala. You can use it for named entity recognition to identify native words.

Custom NER Models: Train custom NER models using tools like the Stanford NER or spaCy's training capabilities, specifically tuned for Sinhala entities.

Word Embeddings:

FastText for Sinhala: FastText provides word embeddings for various languages, including Sinhala. Pre-trained word embeddings can be useful for capturing semantic relationships between words.

Machine Translation Models:

Sinhala-English Machine Translation Models: Use machine translation models to translate Sinhala essays into English, replace native words with English equivalents, and then translate them back to Sinhala. This approach may require a reliable machine translation system.

Custom Sinhala Dictionaries:

Sinhala-English Dictionaries: Build or use existing Sinhala-English dictionaries to map native words to their generic counterparts. Custom dictionaries can be created based on the specific context of Grade 5 Sinhala language essays.

#### Programming Languages:

Python: Python is widely used in NLP tasks. Libraries such as spaCy, NLTK, and scikit-learn can be beneficial for implementing the algorithms and models.

Java: If you prefer Java, libraries like Apache OpenNLP may provide tools for tokenization and other language processing tasks.

#### Text Editors or IDEs:

Visual Studio Code, PyCharm, or Jupyter Notebooks: These tools can be used for coding and implementing your algorithms.

#### Version Control System:

Git: Use Git for version control, especially if you are working collaboratively on the development of the system.

## 6.4 Commercialization plan

- **Social Media Commercialization:**

The proposed application aims to implement a sustainable business model through social media commercialization. Leveraging the popularity and reach of social media platforms, the application can generate revenue through advertisements, sponsored content, and partnerships with relevant educational organizations. By strategically integrating non-intrusive advertisements or sponsored educational resources, the platform can maintain a free or low-cost subscription model for users, ensuring accessibility while sustaining its operations.

- **Free Subscription for Child Orphanages:**

In line with our commitment to social responsibility, the application intends to offer free subscriptions to child orphanages. This initiative seeks to provide underprivileged children with access to educational resources, fostering their academic growth and personal development. By eliminating subscription fees for child orphanages, the application contributes to bridging educational disparities and creating a positive impact on the lives of those who may have limited access to quality educational tools.

- **Low-Cost Subscription for Government Schools:**

Recognizing the budget constraints often faced by government schools, the proposed application plans to offer low-cost subscription plans tailored to meet their financial capacities. This approach ensures that even institutions with limited resources can benefit from the educational features and content provided by the platform. The aim is to support public education by providing affordable

access to a comprehensive learning environment, thus contributing to the improvement of educational outcomes in government schools.

- **Relatedly High-Cost Subscription for International Schools:**  
Catering to the specific needs and financial capabilities of international schools, the proposed application will offer a premium, albeit higher-cost subscription plan. This premium subscription can include additional features, advanced analytics, and personalized support to meet the sophisticated requirements of international educational institutions. The revenue generated from high-cost subscriptions contributes to sustaining the platform's operations and allows for continued improvement and expansion of services for all user categories

## 6.5 Budget

Item	Budget (LKR)
Designing Costs	9000.00
Document Printing	4500.00
Field Visits	5000.00
Server costs	3000.00
Total	21500.00

Table 6.5 – Budget table



## **7. PROJECT REQUIRMENTS**

### **7.1 Non-functional requirements**

Usability

Availability

Maintainability

Performance

Security

### **7.2 Functional requirements**

System should perform linguistic analysis on Sinhala language essays to identify native words, considering various linguistic features, including syntax, semantics, and context.

Implement a Named Entity Recognition (NER) module specifically trained for Sinhala to accurately identify native words within the essays.

Integrate machine learning models to enhance the system's ability to learn and adapt to new native words or linguistic variations over time.

### **7.3 System requirements**

The system ought to be applicable for all the people in Sri Lanka and grade 5 students are targeted.

The system is a mobile application that is able to works on any smart device.

## 8.WORK BREAKDOWN CHART

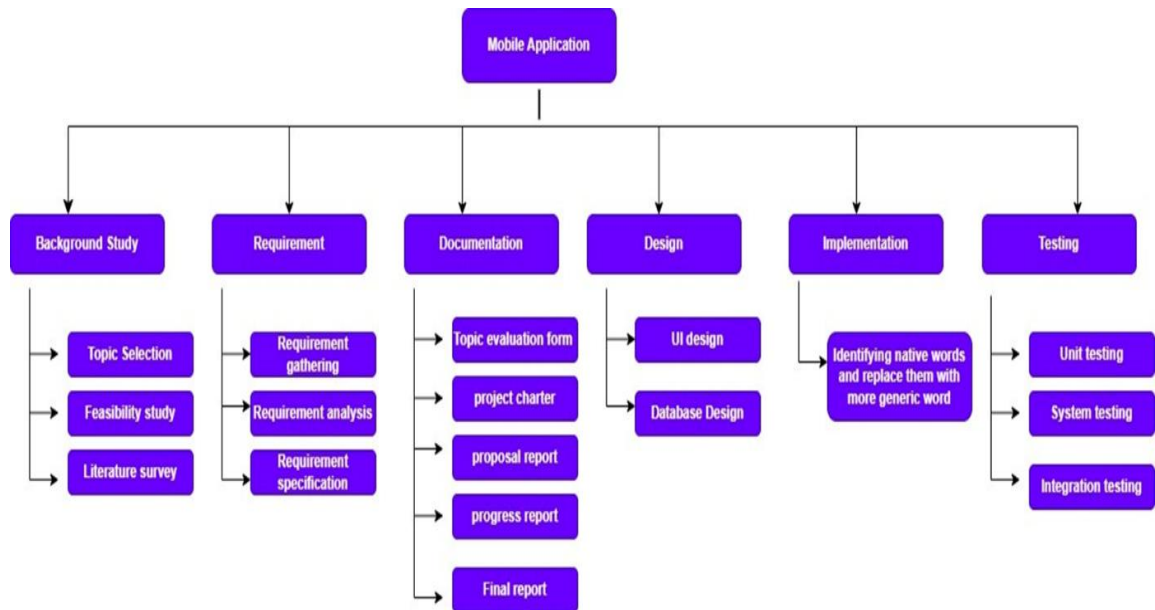


Figure 8.1 – Work breakdown chart

9. GANTT CHART

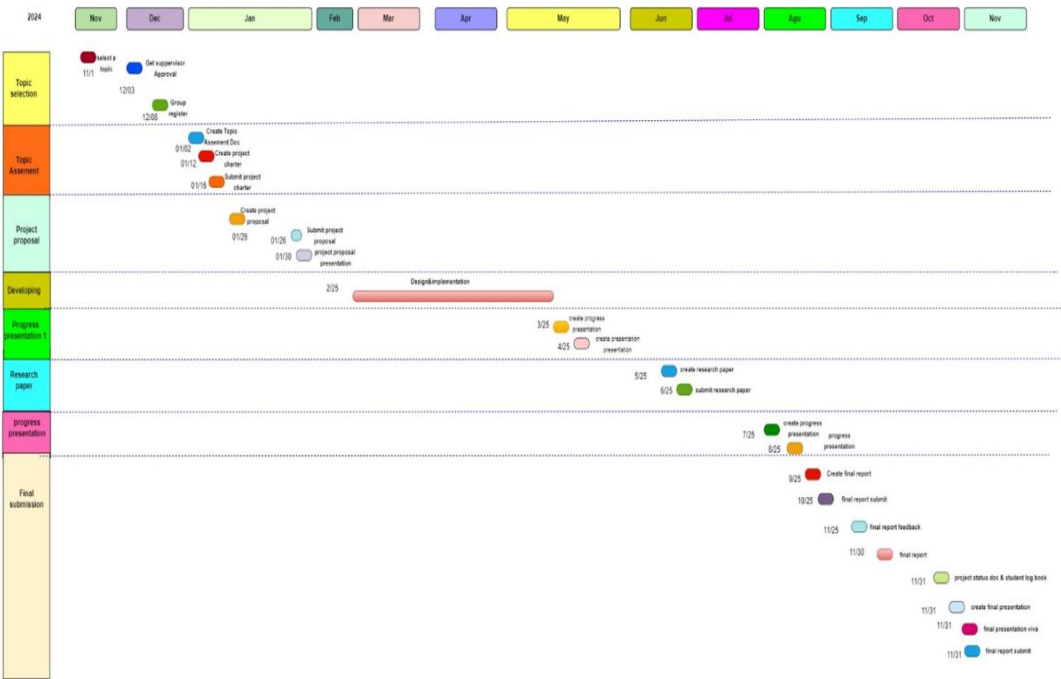


Figure 9.1 – Gantt Chart

## 10.REFERENCES

- [1] Brown, P., & Yule, G. (1983). Teaching the Spoken Language: An Approach Based on the Analysis of Conversational English. Cambridge: Cambridge University Press.
- [2] Jurafsky, D., & Martin, J. H. (2020). Speech and Language Processing (3rd ed.). Pearson.
- [3] Seneviratne, U., & Adikari, A. (2016). Sinhala Language Processing: A Review. In 2016 Moratuwa Engineering Research Conference (MERCon).
- [4] Sinhala Official Language Act No. 33 of 1956, Government of Sri Lanka.
- [5] J. K. Dahanayaka and A. R. Weerasinghe, “Named entity recognition for sinhala language,” in Advances in ICT for Emerging Regions (ICTer), 2014 International Conference on. IEEE, 2014, pp. 215–220
- [6] S. Gallege, “Analysis of sinhala using natural language processing techniques,” 2010.
- [7] Vijaya Shetty, SKadagathur Raghavendra, Rao GuruvyasPranav, Prashantha PatilShow ,Gunakimath Suryakanth, “An automated essay evaluation system using natural language processing and sentiment analysis,” December 2022
- [8] Kaushal Yadav, “Automated Essay Grading System using NLP Techniques,” 2020, International Journal of Engineering and Advanced Technology
- [9] Y. Wijeratne, N. de Silva, and Y. Shanmugarajah, “Natural Language Processing for Government: Problems and Potential,” LIRNEasia, 2019.
- [10] S. Herath, T. Ikeda, S. Yokoyama, H. Isahara, and S. Ishizaki, “Sinhalese morphological analysis: a step towards machine processing of sinhalese,” in [Proceedings 1989] IEEE International Workshop on Tools for Artificial Intelligence. IEEE, 1989, pp. 100–107.





