AI-POWERED EDUCATIONAL ASSISTANT FOR SINHALA RESOURCES

25-26J-448

Project Proposal Report

Ranaweera P.H.K

IT22234452

B.Sc. (Hons) in IT Specialized in Software Engineering

Department of Computer Science and Software Engineering

Sri Lanka Institute of Information Technology

Sri Lanka

August 2025

AI-POWERED EDUCATIONAL ASSISTANT FOR SINHALA RESOURCES

25-26J-448

Project Proposal Report

Ranaweera P.H.K
IT22234452

B.Sc. (Hons) in IT Specialized in Software Engineering

Department of Computer Science and Software Engineering

Sri Lanka Institute of Information Technology

Sri Lanka

August 2025

DECLARATION

We declare that this is our 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
Ranaweera P.H.K.	IT22234452	

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

Signature of the supervisor: Date:

(Prof. Dilshan De Silva)

……………………………. ……………………..

ABSTRACT

This research proposes the development of a Sinhala Document Processing and Embedding framework as a core component of an AI-powered educational assistant for Sri Lankan schools. The system addresses the lack of robust digital linguistic resources in Sinhala by enabling the digitization of both printed and handwritten educational materials. Using a combination of Optical Character Recognition (OCR), Natural Language Processing (NLP), and embedding generation techniques, the framework will transform Sinhala documents into machine-readable semantic representations.

The methodology consists of five phases: data collection, preprocessing and OCR, embedding generation, integration, and evaluation. Printed and handwritten Sinhala resources will be processed using open-source OCR engines such as Tesseract and EasyOCR, alongside deep learning-based CNN/CRNN models for handwriting recognition. Extracted text will be normalized and embedded using models such as FastText, Word2Vec, and multilingual BERT, producing semantic vectors suitable for intelligent retrieval and downstream educational applications. The framework will support both online (cloud-hosted) and offline (lightweight local) modes, ensuring accessibility across urban and rural schools.

Evaluation will be performed using accuracy metrics for OCR (CER, WER), similarity tasks for embeddings, and usability tests in real classroom environments. By relying on free and open-source tools, university resources, and minimal additional costs, the project remains cost-effective while delivering professional-grade outcomes.

This research bridges a critical gap in Sinhala Natural Language Processing by integrating OCR and semantic embeddings into a unified system. It lays the foundation for advanced educational applications such as semantic search, automated answer evaluation, and resource-based question answering, contributing to equitable access to AI-powered education in Sri Lanka.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Full Form
AI	Artificial Intelligence
NLP	Natural Language Processing
OCR	Optical Character Recognition
CNN	Convolutional Neural Network
CRNN	Convolutional Recurrent Neural Network
mBERT	Multilingual Bidirectional Encoder Representations from Transformers
SBERT	Sentence-BERT
CER	Character Error Rate
WER	Word Error Rate
GPU	Graphics Processing Unit
API	Application Programming Interface
SQL	Structured Query Language
SQLite	Structured Query Language -Lightweight Edition
FAISS	Facebook AI Similarity Search
AWS	Amazon Web Services
S3	Simple Storage Service
UI	User Interface
DB	Database

LIST OF APPENDICES

1. INTRODUCTION

1.1. Background

AI integration into the education sector has brought about major changes in how teaching, learning, and assessment are carried out across the world.AI-powered systems such as automated grading, intelligent tutoring, and resource-based question answering have reduced educators’ workload, ensured consistency in evaluation, and provided learners with personalized feedback and support. However, most of these innovations have been directed towards languages such as English [5], [6], [9], resulting in limited progress for low-resource languages like Sinhala, which serves as the principal language of education in Sri Lanka.

One of the fundamental challenges in extending AI applications to Sinhala education lies in the limited availability of digital linguistic resources. Unlike English, which benefits from extensive datasets, corpora, and pre-trained models, Sinhala has relatively few structured digital resources to support advanced Natural Language Processing (NLP) [1], [2]. This scarcity not only constrains the development of downstream applications such as automated grading, and question answering, but also hinders the digitization of handwritten materials, which are still widely used in Sri Lankan schools. The lack of robust Sinhala handwriting recognition systems further complicates the creation of comprehensive digital resources, as students’ handwritten notes, exam scripts, and other educational documents cannot be easily transformed into machine-readable formats. Since these applications rely heavily on rich textual representations and embeddings, the absence of both typed and handwritten Sinhala datasets significantly limits the effectiveness of AI-driven educational tools.

To address this gap, the AI-Powered Educational Assistant for Sinhala Resources project introduces a core component: Sinhala Document Processing and Embedding. This component is responsible for digitizing Sinhala educational materials-including textbooks, notes, and exam resources-and transforming them into machine-readable embeddings that capture semantic meaning. By doing so, it enables downstream tasks such as question answering, knowledge retrieval, and automated evaluation to move beyond keyword matching and instead operate on conceptual understanding.

By supporting both online and offline functionality, the system remains accessible to schools in rural regions with poor connectivity while also serving urban schools with stable internet infrastructure. Through document processing and embedding, Sinhala educational content can be indexed, searched, and understood by AI systems in ways that were previously restricted to resource-rich languages. This not only lays the foundation for more advanced educational tools-such as semantic search, voice/text

Q&A, and automated answer evaluation-but also contributes to bridging the digital divide in Sri Lankan education.

1.2. Literature Survey

The advancement of Natural Language Processing (NLP) and Optical Character Recognition (OCR) has significantly influenced the development of intelligent educational systems across many languages. However, Sinhala, despite being the primary medium of instruction in Sri Lanka, remains underrepresented in this domain. Existing research has made important contributions to Sinhala digitization, OCR, and embedding, which directly inform the design of a document processing and embedding module for AI-powered educational systems.

Early work such as *Adapting the Tesseract Open-Source OCR Engine for Tamil and Sinhala Legacy Fonts* [1] demonstrated the feasibility of extending open-source OCR engines to recognize under-resourced languages. By creating a parallel corpus and adapting Tesseract to handle Sinhala and Tamil legacy fonts, this study provided one of the first systematic approaches to digitizing Sinhala print material. While effective for printed text, the approach highlights the need for more advanced models to address handwriting and diverse font variations.

Building on this, *SinhaLearn: NLP, CNN, and OCR Based Data Driven Approach* [2] introduced a more comprehensive framework that combined deep learning methods with traditional NLP. This research integrated convolutional neural networks (CNNs) for Sinhala handwriting recognition, alongside grammar and spelling correction modules. Such a system showcases the potential of combining OCR with downstream NLP processes, directly aligning with the requirements of Sinhala document digitization for educational resources.

Another critical area of exploration is the representation of Sinhala in embedding spaces. *Bilingual Lexical Induction for Sinhala-English using Cross-Lingual Embedding Spaces* [3] tackled the challenge of aligning Sinhala with resource-rich languages like English. By demonstrating cross-lingual embedding alignment, the study highlighted how semantic representations can bridge Sinhala resources with global NLP models. This directly supports the embedding component of the proposed system, enabling semantic search, retrieval, and document understanding in both monolingual and bilingual contexts.

In addition to Sinhala-specific research, comparable efforts in related contexts also provide valuable insights. For instance, *Intelligent Image Text Reader using Easy OCR, NRCLex & NLTK* [4] proposed an integrated OCR-NLP pipeline capable of text recognition, translation, and emotion detection across multiple languages such as Tamil and Hindi. Although not Sinhala-focused, this work demonstrates the feasibility of coupling OCR with higher-level NLP tasks, including stop-word removal, named entity recognition, and text-to-speech output [4], [5]. The methodology reinforces the importance of embedding OCR outputs into semantic and application-driven pipelines [6], [9], [10], a principle highly relevant to the Sinhala educational domain.

Together, these studies establish a strong foundation for developing Sinhala document processing and embedding systems [1]-[4]. They emphasize the importance of combining OCR for digitization [1], [5], embeddings for semantic understanding [7]-[10], and NLP pipelines for practical applications [4]. However, they also reveal a research gap: the lack of a unified, Sinhala-specific solution capable of processing both printed and handwritten materials [1], [2], while generating embeddings that support intelligent educational applications [3], [7], [9]. Addressing this gap is central to the proposed component of the AI-Powered Educational Assistant for Sinhala Resources.

1.3. Research Gap

Despite notable progress in the areas of Optical Character Recognition (OCR) and Natural Language Processing (NLP) for Sinhala, several limitations persist that hinder the development of comprehensive solutions for education-focused applications. Early studies, such as the adaptation of the Tesseract OCR engine for Sinhala fonts [1], concentrated mainly on printed text and offered limited support for handwritten documents or noisy real-world data. Subsequent efforts, including frameworks like SinhaLearn [2], introduced handwriting recognition and grammar correction, but these approaches were restricted by small-scale datasets and did not extend to generating robust semantic embeddings required for advanced downstream applications such as semantic search, automated grading, or intelligent question answering [9], [10].

Research into embeddings for Sinhala has primarily focused on bilingual alignment with English [3], emphasizing cross-lingual applications rather than the development of rich monolingual embeddings that capture Sinhala’s unique linguistic features, dialectal diversity, and cultural context. Furthermore, OCR-NLP pipelines demonstrated in other regional languages, such as Tamil and Hindi [4], highlight the feasibility of integrating digitization with higher-level NLP tasks [5], [6]. However, Sinhala still lacks a unified framework that combines OCR-based digitization [1], embedding generation [7]-[10], and semantic representation, particularly with an emphasis on accessibility in both online and offline environments.

Therefore, the research gap lies in the absence of a comprehensive Sinhala document processing and embedding solution that can:

- Reliably handle both printed and handwritten Sinhala text.
- Generate high-quality semantic embeddings tailored to the Sinhala language.
- Provide a foundation for educational applications such as automated answer evaluation, intelligent search, and resource-based Q&A.

- Function effectively in both urban schools with stable internet access and rural schools with limited connectivity.

Addressing this gap is critical to enabling AI-powered educational systems that are inclusive, linguistically sensitive, and adaptable to the specific needs of Sri Lankan students and educators.

1.4. Research Problem

Although Artificial Intelligence (AI) has been successfully applied to education in many languages, the Sinhala language still lacks a robust digital infrastructure to support advanced AI-driven applications. Existing OCR solutions for Sinhala are largely limited to printed text recognition and perform poorly with handwritten content, varied fonts, and noisy documents commonly encountered in real educational settings. Furthermore, the scarcity of large-scale Sinhala corpora and pre-trained models has constrained the development of reliable semantic embeddings capable of capturing the richness of Sinhala grammar, vocabulary, and regional dialects.

As a result, current systems are unable to provide accurate, context-aware digitization and representation of Sinhala educational materials, which in turn restricts the effectiveness of downstream applications such as automated grading, resource-based Q&A, and intelligent search. The lack of an integrated framework that combines OCR with high-quality Sinhala embeddings, while being accessible in both online and offline settings, creates a significant barrier to adopting AI-powered educational tools in Sri Lankan schools, particularly in rural areas.

Thus, the core research problem is:

How can Sinhala documents - both printed and handwritten-be effectively digitized and transformed into meaningful semantic embeddings that support downstream AI-driven educational applications, while ensuring accessibility across diverse technological and infrastructural contexts in Sri Lanka?

2. OBJECTIVES

2.1.Main Objective

To develop an AI-powered Sinhala document processing and embedding framework capable of digitizing both printed and handwritten text, and generating semantic representations that enable advanced educational applications in Sri Lankan schools.

2.2. Specific Objectives

- To design and implement a Sinhala OCR pipeline capable of accurately recognizing both printed and handwritten text with support for diverse fonts, styles, and document qualities.
- To build a Sinhala embedding model that captures semantic meaning, grammar, and regional variations of the language for effective downstream tasks.
- To integrate OCR and embeddings into a unified framework that supports document digitization, intelligent search, and semantic representation of Sinhala educational resources.
- To evaluate the system’s performance against benchmarks using accuracy, precision, recall, and semantic similarity metrics for both printed and handwritten documents.
- To ensure accessibility in diverse contexts by enabling both online (cloud-based) and offline (local device) modes, making the solution practical for schools with varying levels of internet connectivity.
- To provide a foundation for downstream educational applications, such as automated answer evaluation, resource-based Q&A, and personalized learning tools, through enriched Sinhala embeddings.

3. METHODOLOGY

3.1. Workflow Description

The proposed methodology for Sinhala Document Processing and Embedding consists of four major phases: data collection, preprocessing and OCR, embedding generation, and system evaluation.

1. Data Collection

- Printed Sinhala documents will be gathered from textbooks, examination papers, and academic notes used in Sri Lankan schools.
- Handwritten documents will be collected from student answer sheets, classroom notes, and controlled handwriting samples.
- A balanced dataset will be curated to include variations in font styles, handwriting patterns, and document quality (e.g., scanned copies, images with noise).

2. Preprocessing and OCR Development

- Image Preprocessing: Apply noise removal, binarization, skew correction, and segmentation to improve recognition accuracy.
- OCR for Printed Text: Adapt open-source OCR tools (e.g., Tesseract with Sinhala-trained data) to handle modern Sinhala fonts.
- OCR for Handwriting: Investigate deep learning-based models (CNN/CRNN architectures) trained or fine-tuned on Sinhala handwriting datasets.
- Text Normalization: Standardize Unicode encodings, resolve ligature inconsistencies, and normalize spacing for downstream processing.

3. Sinhala Embedding Generation

- Corpus Development: Construct a digital Sinhala text corpus using collected OCR outputs and supplementary online resources (educational websites, e-books, etc.).
- Embedding Models: Train or fine-tune pre-trained language models (e.g., FastText, Word2Vec, or transformer-based embeddings like multilingual BERT) for Sinhala.
- Semantic Representation: Evaluate embeddings on tasks such as synonym detection, sentence similarity, and cross-dialect understanding.

4. Integration Framework

- Develop a unified processing pipeline that links OCR outputs to embeddings, creating searchable and semantically meaningful representations of Sinhala documents.
- Online Mode: Cloud-hosted for scalability and faster processing.
- Offline Mode: Lightweight local deployment with compressed models for low-resource school environments.

5. System Evaluation

- OCR Evaluation: Use accuracy, character error rate (CER), and word error rate (WER) metrics on both printed and handwritten datasets.
- Embedding Evaluation: Assess using cosine similarity, intrinsic word similarity tasks, and extrinsic evaluations on question answering or document retrieval.
- User Testing: Pilot in selected schools (urban and rural) to test usability, accessibility, and performance under real classroom conditions.

The overall workflow of the proposed Sinhala Document Processing and Embedding methodology is illustrated in Figure 1.

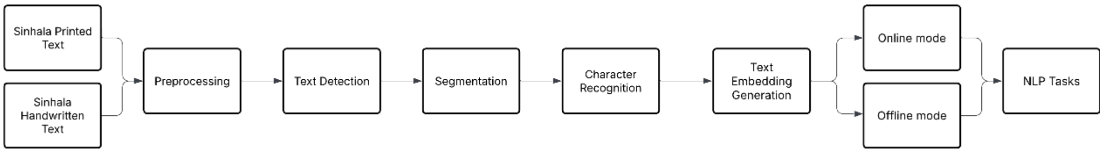


Figure 1 Workflow of the Sinhala Document Processing and Embedding methodology

3.2. Tools and Technologies

The development of the Sinhala Document Processing and Embedding framework will leverage a combination of OCR libraries, NLP toolkits, deep learning frameworks, and storage solutions to ensure accurate digitization, semantic representation, and integration with downstream educational applications.

3.2.1. OCR and Image Processing

- Tesseract OCR -For printed Sinhala text recognition.
- OpenCV -For preprocessing tasks such as noise removal, binarization, skew correction, and segmentation.
- Custom CNN/CRNN Models (PyTorch/TensorFlow) -For recognizing handwritten Sinhala text.
- EasyOCR -As a lightweight alternative OCR engine supporting Sinhala, useful for rapid prototyping.

3.2.2. Natural Language Processing (NLP)

- NLTK / SpaCy -For text preprocessing (tokenization, stop-word removal, POS tagging).
- Indic NLP Library -For handling script-specific challenges such as Unicode normalization, compound words, and ligatures in Sinhala.
- Sinhala WordNet (if available) -For synonym and semantic relationship mapping.

3.2.3. Embedding Models

- Word2Vec / FastText -To generate word-level embeddings for Sinhala documents.
- Multilingual BERT (mBERT) -To capture semantic meaning and contextual variations in Sinhala text.
- Sentence Transformers (SBERT) -For sentence-level embeddings to support semantic search and similarity scoring.

3.2.4. Machine Learning & Deep Learning Frameworks

- TensorFlow and PyTorch -For training and fine-tuning models for OCR, embeddings, and downstream NLP tasks.
- Hugging Face Transformers -For integration with pre-trained multilingual language models such as mBERT and XLM-R.

3.2.5. Storage & Indexing

- SQLite -Lightweight local database for storing embeddings in offline mode.
- FAISS / Annoy -Vector similarity search libraries for efficient retrieval of embeddings.
- Firebase / AWS S3 -Cloud storage and synchronization in online mode.

3.2.6. Development & Deployment Tools

- Python -Primary programming language for OCR, NLP, and embeddings.
- Jupyter Notebook / Google Colab -For experimentation and prototyping.
- Docker -For containerized deployment ensuring consistency across environments.
- Flutter (Mobile App Integration) -For embedding retrieval and access on mobile devices.

4. PROJECT REQUIREMENTS

The successful implementation of the Sinhala Document Processing and Embedding framework requires a clear definition of functional and non-functional requirements, as well as expected test cases that verify system behavior under real-world conditions.

4.1. Functional Requirements

The system will deliver the following essential functionalities:

1. Accept Sinhala educational resources in both printed and handwritten formats (PDFs, images, and scanned student answer sheets).
2. Perform document preprocessing operations such as noise removal, skew correction, and binarization to enhance OCR accuracy.
3. Recognize Sinhala text through Optical Character Recognition (OCR) for both printed and handwritten inputs.
4. Normalize extracted text into a standardized Sinhala Unicode representation to ensure consistency.
5. Generate semantic embeddings using approaches such as Word2Vec, FastText, and multilingual BERT (mBERT).
6. Store embeddings in SQLite for offline mode and in Firebase/FAISS for cloud-based retrieval in online mode.
7. Provide search and retrieval capabilities based on semantic similarity rather than exact keyword matches.
8. Operate seamlessly in both online (cloud-hosted) and offline (local device) environments.
9. Support integration with the AI-Powered Educational Assistant platform to enable downstream tasks such as intelligent search and automated evaluation.

4.2. Non-Functional Requirements

In addition to core functionality, the system must satisfy the following quality attributes:

1. **Accuracy** - OCR accuracy of at least 85% for printed Sinhala and 75% for handwritten Sinhala.
2. **Performance** - Document processing should complete within 5 seconds per printed page and 10 seconds per handwritten page on mid-range devices.

3. **Scalability** - Cloud mode should support batch processing of textbooks containing more than 100 pages.
4. **Portability** - The system should operate on laptops, desktops, and Android mobile devices.
5. **Reliability** - Offline mode must work without internet connectivity, ensuring previously processed data remains usable.
6. **Usability** - Provide a simple, user-friendly interface suitable for educators and students with minimal technical expertise.
7. **Security** - Data stored in SQLite (offline) must remain private, while cloud storage (Firebase/AWS) must enforce secure authentication.

4.3. Expected Test Cases

To validate the requirements, the expected test cases are summarized in Table 1.

Table 1: Expected Test Cases for the Sinhala Document Processing and Embedding Component

Test Case	**Input**	**Expected Output**	**Requirement Verified**
TC1	Upload a printed Sinhala PDF	Extracted text with ≥ 85% accuracy	Functional Req. 1, 3
TC2	Upload a handwritten Sinhala answer sheet	Extracted text with ≥ 75% accuracy	Functional Req. 3
TC3	Upload a low-quality/noisy scan	Preprocessed image with normalized Sinhala text	Functional Req. 2, 4
TC4	Enter search query "ගුරු" (teacher)	Retrieve semantically related documents (e.g., "ගුරුවරයා")	Functional Req. 5, 7
TC5	Run the system in offline mode	Embeddings stored in SQLite and searchable locally	Functional Req. 8
TC6	Run the system in online mode	Embeddings stored in Firebase/FAISS with fast retrieval	Functional Req. 6, 8
TC7	Process a 100-page	Completed within 5 minutes	Non-Functional

	Sinhala textbook in cloud mode		Req. 3
TC8	Process one handwritten page on a mid-range phone	Completed within 10 seconds	Non-Functional Req. 2
TC9	Retrieve previously stored embeddings	Consistent retrieval in both offline and online modes	Functional Req. 6, 7

5. DESCRIPTION OF PERSONNEL AND FACILITIES

5.1. Personnel

This component of the research is conducted individually by Hasindu Koshitha Ranaweera (Student ID: IT22234452), an undergraduate student in the B.Sc. (Hons) in Information Technology program at the Sri Lanka Institute of Information Technology (SLIIT).

The researcher is responsible for every stage of the Sinhala Document Processing and Embedding module, including:

- Designing the document ingestion pipeline to handle both scanned Sinhala educational resources and handwritten materials.
- Implementing OCR solutions (Tesseract and CNN/CRNN models) for printed and handwritten Sinhala text recognition.
- Preprocessing extracted text using normalization, tokenization, and stop-word removal techniques tailored for the Sinhala language.
- Developing semantic embedding strategies (Word2Vec, FastText, and mBERT) to create searchable vector representations of Sinhala text.
- Designing and deploying an embedding storage and retrieval system using SQLite for offline mode and FAISS/Firebase for online mode.
- Ensuring smooth integration of the Sinhala Document Processing pipeline into the AI-Powered Educational Assistant platform.

The project is supervised by Prof. Dilshan De Silva (Supervisor) and Ms. Chamali Pabasara (Co-Supervisor), who provide expertise in OCR, low-resource NLP, and document embedding strategies.

5.2. Facilities

The research utilizes a combination of personal and institutional facilities to simulate both development and deployment in real-world educational environments.

- **Computing Devices** - A personal laptop with sufficient memory and processing power is used as the primary development environment. Additional devices (a secondary laptop and an Android smartphone) are used to test OCR accuracy, embedding retrieval, and offline functionality in mobile contexts.
- **Internet Connectivity** - Development and training make use of home and university Wi-Fi connections for accessing online datasets, running cloud-based experiments, and synchronizing embeddings. Offline scenarios are simulated to ensure system functionality in rural schools with poor connectivity.
- **Software and Development Tools**
    - **OCR**: Tesseract OCR, EasyOCR, OpenCV.
    - **NLP & Embeddings**: Hugging Face Transformers, Indic NLP Library, FastText, mBERT, Sentence-BERT
    - **Deep Learning Frameworks**: PyTorch and TensorFlow for model training and fine-tuning.
    - **Storage & Indexing**: SQLite for offline embedding storage; FAISS and Firebase for cloud-based retrieval.
    - **Annotation & Preprocessing**: Label Studio for dataset annotation, Python-based text preprocessing scripts.
- **Collaboration and Version Control -** A private GitHub repository is used for code versioning and collaboration. Google Colab Pro provides GPU resources for OCR model experimentation and embedding training. Microsoft Teams and Google Drive are used for communication and progress sharing with supervisors.

By combining these resources with the researcher’s technical expertise, this project ensures the development of a robust Sinhala Document Processing and Embedding system that operates effectively in both online and offline modes, thereby supporting practical deployment in diverse Sri Lankan educational settings.

6. BUDGET AND JUSTIFICATION

The budget for this component is minimal since most tools are open-source and existing personal/university devices are adequate.

6.1. Estimated Budget

Table 2. Estimated Budget for the Sinhala Document Processing and Embedding Component

Item	Description	Estimated Cost (LKR)
Google Colab Pro	GPU access for OCR fine-tuning and embedding experiments (USD 9.99 × 12 months)	39,600
Software & Tools	Free and open-source: Tesseract OCR, EasyOCR, Hugging Face Transformers, PyTorch, ONNX Runtime, Scikit-learn, SQLite, Label Studio	0
Hardware	No new devices purchased. Personal laptop and smartphone used for testing Sinhala OCR models	0
Internet	Covered by existing home and university Wi-Fi connections	0
Contingency	Minor costs (printing, data storage, dataset annotation support)	5,000

Total Estimated Cost : LKR 44,600 (≈ USD 150 for one year)

6.2. Justification Summary

- GPU Access: Sinhala OCR and embeddings require computational resources. Google Colab Pro will provide affordable GPU support for training/fine-tuning. University lab GPUs will be used for heavy experiments.
- Software & Tools: Open-source OCR frameworks (Tesseract, EasyOCR), deep learning libraries (PyTorch, Hugging Face Transformers), and database tools (SQLite) ensure zero licensing cost.
- Hardware: Personal laptop and university lab PCs are sufficient for training and testing. Mobile device used for field-level testing.

- Data Annotation: Sinhala document dataset annotation is necessary for OCR accuracy and embedding quality. Most annotation will be done by the researcher; a small contingency budget is kept for outsourcing quality checks.
- Cost Efficiency: By leveraging free/open-source tools and existing devices, this component maintains a very low cost while still supporting advanced OCR + embedding research.

7. CONCLUSION

The proposed Sinhala Document Processing and Embedding component is a crucial step toward bridging the digital divide in Sri Lanka's education system. By enabling the digitization of both printed and handwritten Sinhala materials and converting them into semantic embeddings, this research directly addresses the lack of linguistic resources that has historically limited AI applications in Sinhala education.

The integration of OCR with embedding generation ensures that educational content can be represented and retrieved not merely as text, but as meaningful semantic information. This foundation empowers downstream applications such as semantic search, question answering, and automated evaluation, which are essential to modern learning environments. Furthermore, the system's dual online-offline capability guarantees inclusivity by serving both resource-rich urban schools and rural institutions with limited connectivity.

The methodology's reliance on open-source tools, lightweight storage systems, and affordable compute resources ensures both academic feasibility and practical scalability. By combining technical innovation with a focus on accessibility, this component establishes a sustainable framework for Sinhala educational digitization.

Ultimately, the successful completion of this project will provide Sri Lankan educators and learners with tools previously limited to resource-rich languages, fostering equitable access to AI-powered educational technologies and setting the groundwork for future innovations in Sinhala Natural Language Processing

REFERENCES

[1] C. Vasantharajan, L. Tharmalingam, and U. Thayasivam, “Adapting the Tesseract Open Source OCR Engine for Tamil and Sinhala Legacy Fonts and Creating a Parallel Corpus for Tamil Sinhala English,” in Proc. 2022 Int. Conf. on Asian Language Processing (IALP), Singapore, Oct. 2022, pp. 143-149, doi: 10.1109/IALP57159.2022.9961304. [Online]. Available: https://ieeexplore.ieee.org/document/9961304

[2] J. V. Francis and Y. L. Bellanavithana, “SinhaLearn: NLP, CNN, and OCR Based Data Driven Approach for Enhancing Sinhala Proficiency of Grade 5 Scholarship Students,” in Proc. 2024 Moratuwa Engineering Research Conf. (MERCon), Moratuwa, Sri Lanka, Aug. 2024, pp. 536-541, doi: 10.1109/MERCon63886.2024.10689004. [Online]. Available: https://ieeexplore.ieee.org/document/10689004

[3] A. Liyanage, S. Ranathunga, and S. Jayasena, “Bilingual Lexical Induction for Sinhala English using Cross Lingual Embedding Spaces,” in Proc. 2021 Moratuwa Engineering Research Conf. (MERCon), Moratuwa, Sri Lanka, Jul. 2021, pp. 579-584, doi: 10.1109/MERCon52712.2021.9525667. [Online]. Available: https://ieeexplore.ieee.org/document/9525667

[4] C. Jeeva, T. Porselvi, B. Krithika, R. Shreya, G. S. Priyaa, and K. Sivasankari, “Intelligent Image Text Reader using EasyOCR, NRCLex & NLTK,” in Proc. 2022 Int. Conf. on Power, Energy, Control and Transmission Systems (ICPECTS), Chennai, India, 2022, pp. 1-6, doi: 10.1109/ICPECTS56089.2022.10047136. [Online]. Available: https://ieeexplore.ieee.org/document/10047136

[5] R. Smith, “An Overview of the Tesseract OCR Engine,” in Proc. Int. Conf. on Document Analysis and Recognition (ICDAR), Curitiba, Brazil, 2007, pp. 629-633. doi: 10.1109/ICDAR.2007.4376991. [Online]. Available: https://ieeexplore.ieee.org/document/4376991

[6] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, “Synthetic Data and Artificial Neural Networks for Natural Scene Text Recognition,” arXiv preprint arXiv:1406.2227, 2014. [Online]. Available: https://arxiv.org/abs/1406.2227

[7] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” arXiv preprint arXiv:1301.3781, 2013. [Online]. Available: https://arxiv.org/abs/1301.3781

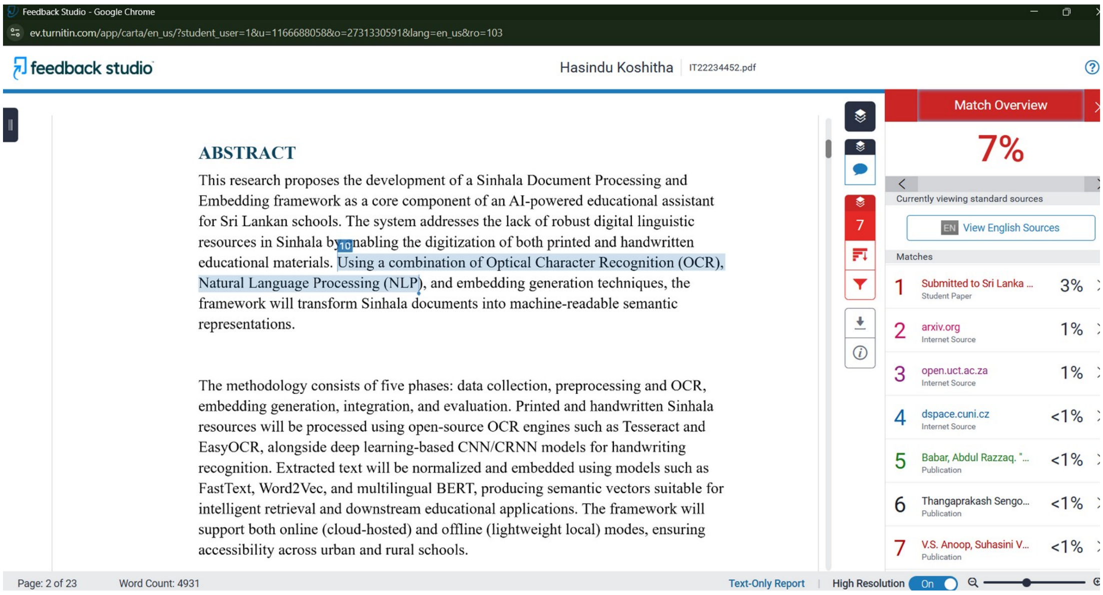
[8] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching Word Vectors with Subword Information,” Trans. Assoc. Comput. Linguistics (TACL), vol. 5, pp. 135-146, 2017. [Online]. Available: https://arxiv.org/abs/1607.04606

[9] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in Proc. Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Minneapolis, USA, Jun. 2019, pp. 4171-4186. [Online]. Available: https://arxiv.org/abs/1810.04805

[10] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” in Proc. Conf. on Empirical Methods in Natural Language Processing (EMNLP), Hong Kong, China, Nov. 2019, pp. 3982-3992. [Online]. Available: https://arxiv.org/abs/1908.10084

[11] J. Johnson, M. Douze, and H. Jégou, “Billion-Scale Similarity Search with GPUs,” IEEE Trans. Big Data, vol. 7, no. 3, pp. 535-547, 2021, doi: 10.1109/TBDATA.2019.2921572. [Online]. Available: https://ieeexplore.ieee.org/document/8733051

APPENDIX A – Turnitin Similarity Report



ev.turnitin.com/app/carta/en_us/?student_user=1&u=1166688058&o=2731330591&lang=en_us&ro=103

Hasindu Koshitha | IT22234452.pdf

ABSTRACT

This research proposes the development of a Sinhala Document Processing and Embedding framework as a core component of an AI-powered educational assistant for Sri Lankan schools. The system addresses the lack of robust digital linguistic resources in Sinhala by 10nabling the digitization of both printed and handwritten educational materials. Using a combination of Optical Character Recognition (OCR), Natural Language Processing (NLP), and embedding generation techniques, the framework will transform Sinhala documents into machine-readable semantic representations.

The methodology consists of five phases: data collection, preprocessing and OCR, embedding generation, integration, and evaluation. Printed and handwritten Sinhala resources will be processed using open-source OCR engines such as Tesseract and EasyOCR, alongside deep learning-based CNN/CRNN models for handwriting recognition. Extracted text will be normalized and embedded using models such as FastText, Word2Vec, and multilingual BERT, producing semantic vectors suitable for intelligent retrieval and downstream educational applications. The framework will support both online (cloud-hosted) and offline (lightweight local) modes, ensuring accessibility across urban and rural schools.

1	Submitted to Sri Lanka ... Student Paper	3%
2	arxiv.org Internet Source	1%
3	open.uct.ac.za Internet Source	1%
4	dspace.cuni.cz Internet Source	<1%
5	Babar, Abdul Razzaq. "... Publication	<1%
6	Thangaprakash Sengo... Publication	<1%
7	V.S. Anoop, Suhasini V... Publication	<1%

Page: 2 of 23 Word Count: 4931