Temporary ID: TMP-22-102

Sri Lanka Institute of Information Technology

Project Topic Assessment – 2022 (Regular)

Topic

Sinhala-English code-switching conversational AI with enhanced NLP pipeline components, XAI, and Performance Evaluation.

Abstract (200 Words Max):

There is a considerable amount of research conducted focusing on high-resource languages. Because of that, there are only a handful of NLP tools and datasets for low-resource natural languages. Building NLP tools around low-resource languages tend to be challenging since they require a large, dialect-rich text corpus in the interested language. Moreover, the increment of code-switching (usage of a mix of multiple languages among monolingual speakers) has become another major issue in handling textual data in NLP as it requires processing two or more languages simultaneously. This research focuses on developing NLP tools from scratch, enhancing, and optimizing them for Sinhala-English code-switched textual data where the number of resources available is considerably low. The developed models and tools will then get attached as NLU pipeline components to a domain-specific conversational AI designed using RASA with Explainable AI techniques and performance evaluation visualizations for a higher level of human interpretability.

Research Group/Area: Select the area by referring to the document uploaded to the Course Web

Knowledge Inspired Computing (KIC)

Natural Language Processing (NLP)

Supervisor should fill this part

Supervisor and Co-Supervisor endorse the proposed project, and hence, guide the students to acquire required knowledge skills pertaining to above sub domains of their specializations.

Supervisor: **Dr. Lakmini Abeywardhana** Signature

Continuation of Previous Year Project? ☐

If yes, state the Project ID

and year

Co-Supervisor: **Ms. Dinuka Wijendra** Signature

External Supervisor Name

Team Members:

Student Name	Student ID	Specialization
Leader: Dissanayake D. M. I. M.	IT19069432	DS
Member 2: Hameed M. S.	IT19064932	SE
Member 3: Jayasinghe D.T.	IT19075754	DS
Member 4: Sakalasooriya S. A. H. A.	IT19051208	DS

Research Problem:

Below are the main research problems identified,

1. NLP (Natural Language Processing) tools and models for processing Sinhala and Sinhala-English code-mixed textual data and feature engineering are low. [1] Developing deep learning-based models such as named entity recognizers from scratch requires manual data annotating, which is time-consuming and repetitive. Building multilingual models that can handle hundreds of languages may not be suitable for attaching to a conversational AI since those models can be heavy in size and require a lot of training data not in one language but many.

2. Generally, users use a physical keyboard that has an English-specific layout when interacting with the computing devices. If they need to type in Sinhala or else code-switch between English and Sinhala, they will have to use a service like Helakuru to type in the Sinhala words they want and then copy and paste it into the conversational AI, which is a cumbersome task, and most users would not even bother to go through the trouble for it. Due to that reason, having a keyboard interface that can handle code-switching is a must [2][3].

3. Maintaining AI assistants requires knowing terminal commands, running Python scripts, and configuring YAML files to update a single word in the training data by retraining the assistant. It is not ideal for the average user or developer, thus maintaining and improving an AI assistant can be almost impossible for them, which should not be the case since conversational AIs undergo domain changes over time due to constant usage.

4. It is impossible to see how well the AI assistant is performing and whether changes made helped to improve the overall performance. It requires knowing machine learning concepts, a deep understanding of the RASA framework/CLI and terminal commands. Even non-machine learning experts should have the ability to maintain conversational AI by evaluating the model as performing well, overfitting or underfitting.

5. Explainability and human-interpretability of the black-box models is a popular ongoing research area under explainable ai (XAI). DIET Classifier [4] is one of the intent classifiers used in Rasa and is also a black-box model that uses an attention mechanism. Hence, the explainability of the model is low [5]. Although research including LIME [6], SHAP [7], and others provide either local or global explanations, they are independently applied. There is only a handful of research that has tried to blend them [8].

6. Sinhala-English code-mixed datasets are hard to come by. Even a large-enough pure Sinhalese text corpus is hard to find to train NLP models and tools [1]. Generating data from scratch can be highly biased towards a single dialect, conversation pattern, or sentence structure.

7. Although Rasa Open-Source conversational AI framework [9] has a built-in customizable NLU pipeline with modular pipeline components, there is no default mechanism to pre-process or add language-specific dense features to feed the machine learning models. Thus, it will be hard to comprehend that “ඩිග්‍රිය”, “degree”, and “පාඨමාලාව” have similar meanings in the domain-specific code-mixed training data with the limited amount of initial training data and may require data-heavy machine learning models such as language models and word embeddings models.

References

[1]. N. de Silva, "Survey on publicly available Sinhala Natural Language Processing tools and research," Jun. 2019. [Online]. Available: https://arxiv.org/abs/1906.02358

[2]. A. Bawa, P. Khadpe, P. Joshi, K. Bali, en M. Choudhury, "Do Multilingual Users Prefer Chat-Bots That Code-Mix? Let's Nudge and Find Out!," *Proc. ACM Hum. -Comput. Interact.*, vol 4, no CSCW1, pp. 1-23, May 2020, doi: 10.1145/3392846.

[3]. J. Emond, B. Ramabhadran, B. Roark, P. Moreno and M. Ma, "Transliteration Based Approaches to Improve Code-Switched Speech Recognition Performance," 2018 IEEE Spoken Language Technology Workshop (SLT), 2018, pp. 448-455, doi: 10.1109/SLT.2018.8639699.

[4]. T. Bunk, D. Varshneya, V. Vlasov, A. Nichol, "DIET: Lightweight Language Understanding for Dialogue Systems," 2020. [Online]. Available: https://arxiv.org/abs/2004.09936

[5]. H2O.ai, *Interpretable Machine Learning Using LIME Framework - Kasia Kulma (PhD), Data Scientist, Aviva.* (Dec. 19, 2017). Accessed: Dec. 25, 2021. [Online Video]. Available: https://www.youtube.com/watch?v=CY3t11vuuOM&list=TLPQMjYxMjIwMjHjkkneYrcmyg&index=2

[6]. M. T. Ribeiro, S. Singh, C. Guestrin, "'why should I trust you?': Explaining the predictions of any classifier," 16 Feb 2016. [Online]. Available: https://arxiv.org/abs/1602.04938

[7]. S. Lundberg, S. Lee, "A Unified Approach to Interpreting Model Predictions," 2017. [Online]. Available: https://arxiv.org/abs/1705.07874

[8]. M. Setzu, R. Guidotti, A. Monreale, F. Turini, D. Pedreschi, F. Giannotti, "GLocalX - from local to global explanations of black box AI models," *Artif. Intell.*, vol 294, no 103457, bl 103457, May 2021, doi: 10.1016/j.artint.2021.103457.

[9]. T. Bocklisch, J. Faulkner, N. Pawlowski, A. Nichol, "Rasa: Open Source Language Understanding and Dialogue Management," 2017. [Online]. Available: https://arxiv.org/abs/1712.05181

Solution proposed:

The proposed solutions for the above-identified problems include,

1. Building NLP tools that can highly reduce data annotating time for custom entity tagging-related tasks, building lightweight monolingual deep learning word embedding models that can efficiently handle Sinhala-English code-switched textual data which can be used for sparse and dense feature engineering, handling out-of-vocabulary (OOV) tokens, domain-specific Custom Entity Recognizing and efficient training data annotating.

2. Creating a conversational AI with a keyboard interface that makes it possible for the user to interact in both English and Sinhala by switching between the languages (support for code-switching) for words with easy shortcuts. This will allow the users to directly communicate with conversational AI without the use of any third-party tools and keyboard interfaces.

3. Simplify the model improvement for the machine learning models used in the conversational AI assistant by adding more training data related to the domain and then training it with the click of a button without having to explicitly know the framework specifics. Adding training data can include data that can be manually added. Such as new ways of responding to questions, new questions along with their relevant intent, adding the questions that caused fallbacks along with their relevant intent and corresponding answer.

4. Create an algorithm to determine whether models are overfitting, underfitting, or performing well under the current training and validation data and add the results as indicators that can be easily understood by the users without having to make decisions by analyzing training and validation accuracy/loss curves manually, which require in-depth knowledge in machine-learning. It makes it possible to understand possible improvements that can be done to the models and as a performance evaluation method.

5. Create a post hoc model-agnostic algorithm using the black-box approach that can give both local and global explanations which can be used to explain how DIET (Dual Intent Entity Transformer) classifier classifies user intents when a specific question or a phrase is given to the conversational AI. Allow local explanations to consider global feature importance when calculating the local feature importance as a weight to blend local and global explanations.

6. According to the research components and specific tasks, the need for two separate datasets was identified. (1). A slightly larger dataset with domain-specific data to be used to train models from scratch and to fine-tune models when domain adaptation tasks are being performed. (2). A comparably small or medium-sized dataset to train the conversational AI assistant to be able to handle possible FAQs related to SLIIT with a high level of accuracy.

   It was decided to scrape domain-specific textual data from websites related to SLIIT and from news articles. (https://support.sliit.lk/, https://sliitinternational.lk/, and https://www.sliit.lk/ are few such websites). Moreover, publicly available documents such as PDFs and Docx files in the above-mentioned sites will be taken as they are both "public" and "official".

   To capture as many tokens and sentence patterns in code-mixed data as possible, the gathered data in the form of the formal-written format will be converted into four different code-mixed sentences in the form of spoken format with different patterns and at the end, the deduplication step will be followed to eliminate any redundant data. Even though the same sentences are repeated, notice that the patterns and tokens (words) of each sentence will be different from others. This can also be seen as a data augmentation approach to overcome the low-resource text data issue.

7. The issue of assistants not being able to learn sparse and dense features well for non-English languages will be handled by attaching the models and algorithms designed in the first three research components to the Rasa NLU pipeline. Then, the assistant will be able to extract features from those models while training. The last research component applies to the conversational AI after the model has been trained to properly test and evaluate it, which is a crucial part of any machine-learning or AI-related product in general.

System Overview Diagram for the solution proposed. Recommended to draw using draw.io. Note: This is not an activity/flow (UML) diagram

1. **Main components including the data sources, stakeholders, interaction among the stakeholders, etc.**
2. **Interconnection among the components**
3. **Major SW and HW components**

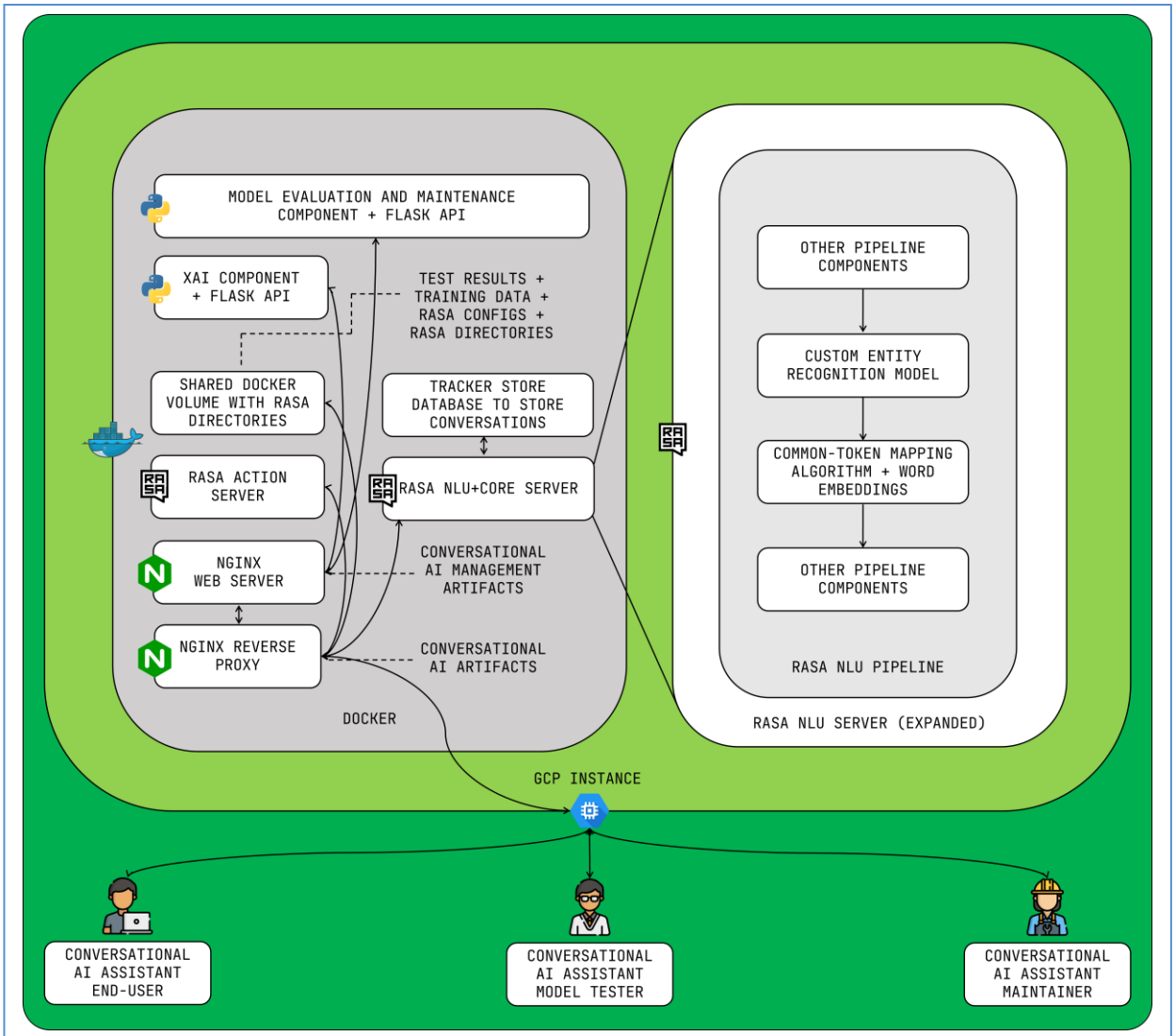

Figure 1. High-level architectural diagram of the purposed solution after all research components integrated

Objectives (1 main objective and 4 sub objectives):

Main Objective:

Developing NLP tools for text preprocessing, feature engineering, training data/model improvements, model performance evaluation, and generating explanations on intent classifications; that can be attached to a domain-specific conversational AI assistant designed for Sinhala-English code-switched text corpus.

Sub Objective 1:

Use explainable AI (XAI) techniques and develop algorithms to generate explanations for black-box intent classifiers both locally and globally.

Sub Objective 2:

Enable codeless model improvement and develop an algorithm to generate model performance evaluation for non-machine-learning experts.

Sub Objective 3:

Developing rule-based algorithms to effectively handle Sinhala-English code-mixed textual data through “token mapping” to make the word embedding models lightweight and to handle out-of-vocabulary (OOV) tokens effectively.

Sub Objective 4:

Develop an efficient way to annotate entities and Recognize domain-specific entities using custom entity recognition in code-mixed textual data.

Task divided among the members

Member 1: IT19069432 - Dissanayake D.M.I.M.

Developing algorithms using explainable AI (XAI) techniques to offer both global and local model explanations for black-box intent classifiers used in conversational Ais. The main tasks of the research component have been described below.

1. Developing an algorithm to generate local explanations based on the global explanations using it as a weighting term, DIME (Dual Interpretable Model-agonastic Explanations).
2. Developing a strategy to calculate feature importance for individual tokens in the training data relevant to a specific intent (class) or relevant to all the intents.
3. Allowing to explore the model explanations for any user input and visualize the model explanations in a human-interpretable way.

Member 2: IT19064932 - Hameed M.S.

Developing an efficient and code-less approach to improving training data of the conversational AI assistant via eliminating the need to interact with backend and implementing efficient model testing/evaluation technique with the zero-coding approach to allow non-machine-learning experts to easily improve the conversational AI assistant models. The main tasks of the research component are as follows.

1. Finding the best possible approach to allow training data improvements to be done without any coding knowledge or manually interacting with the backend.
2. Providing a way to efficiently re-train and deploy new machine learning models of the conversational AI assistant for non-technical users.
3. Evaluating machine learning models and designing an algorithm to automatically identify any overfitting or underfitting scenarios in evaluation reports generated by RASA and indicating them in the frontend.

Member 3: IT19075754 - Jayasinghe D.T.

Developing rule-based algorithms to handle code-switched textual data efficiently and handle out-of-vocabulary tokens (OOV) as described below.

1. Developing an algorithm that uses character-mapping to enable End-users to interact with the conversational AI assistant using code-switching (a rule-based keystroke mapper)
2. Developing an algorithm using “token mapping” which can be used to handle out-of-vocabulary (OOV) tokens in code-mixed training data due to less amount of training data available.
3. Demonstrating how lightweight monolingual word embeddings models that need less amount of training data can handle code-mixed training data tokens.
4. Attaching the algorithm to the conversational AI assistant’s NLU pipeline before the word embeddings components.

Member 4: IT19051208 - Sakalasooriya S.A.H.A.

Building a Custom Entity Recognizer for Sinhala-English code-mixed corpus based on different techniques. The main tasks of this component are as follows.

1. Building a data annotating tool to annotate entities easily without having to tag each entity in the training dataset manually.
2. Exploring different approaches to build the tool such as n-grams, reverse-stemming, word-wise cosine similarity algorithms to effectively find different variations of the same token that share the same base form.
3. Evaluating the performance of the annotating tool for each different approach mentioned above.
4. Designing a deep learning custom entity recognition model using the spaCy library to detect domain-specific custom entities in Sinhala-English code-mixed dataset and attaching it to the conversational AI.

Technologies to be used:

For Custom Entity training data tagging tool and Model:
Python 3, spaCy, React, Flask, Rasa SDK, Google Colab notebooks

For Explanation AI algorithm development:
Python 3, sklearn, Rasa SDK, Flask, Google Colab notebooks

For Token mapping tool and Keyboard Interface:
Python 3, Flask, NLTK, Gensim, Rasa SDK, Google Colab notebooks

For Conversational AI Training and Testing Interface:
Python 3, React, Flask, TensorFlow, Pandas

For Conversational AI Development and Final product deployment:
RASA Open Source 2.8.12, Docker, NGINX, Flask, React JS, Chart JS, Python 3, MongoDB

If supervisor States that this year is a continuation of previous work, state the further work the students should do compared to the previous years.
(NOTE: This part has to be filled by the supervisor)

Not a continuation of a previous work.

Appendix

Appendix 1: Supervision and Co-supervision confirmation emails.

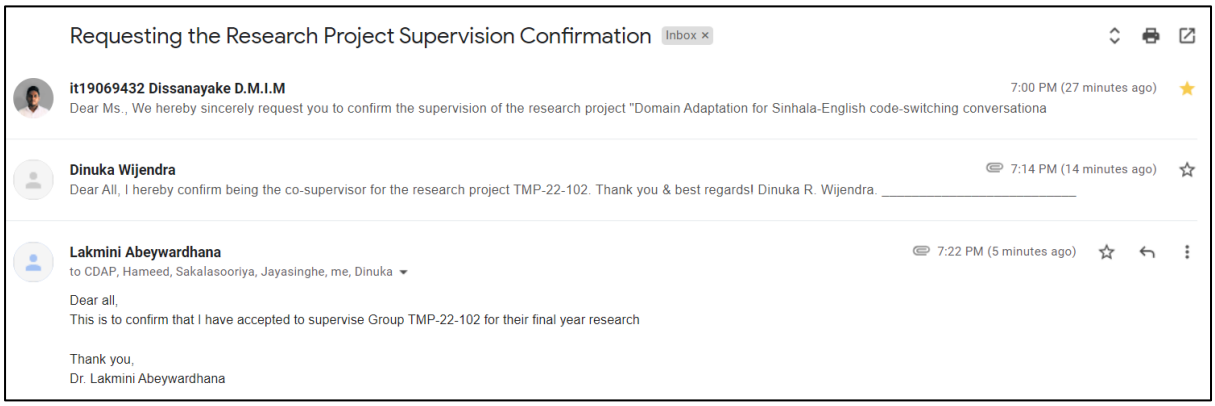

Figure 2. Research project supervision confirmation email

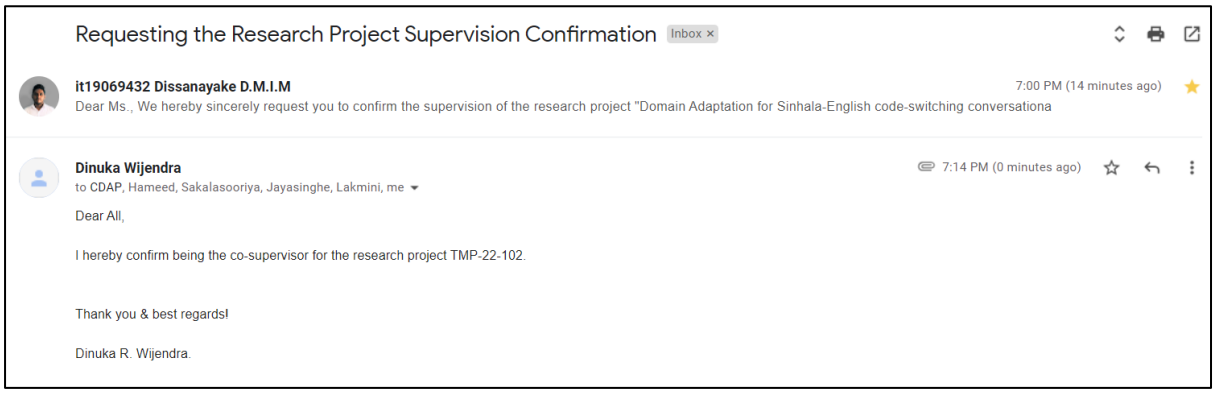

Figure 3. Research project co-supervision confirmation email

Appendix 2: Research component alternation approval confirmation emails.

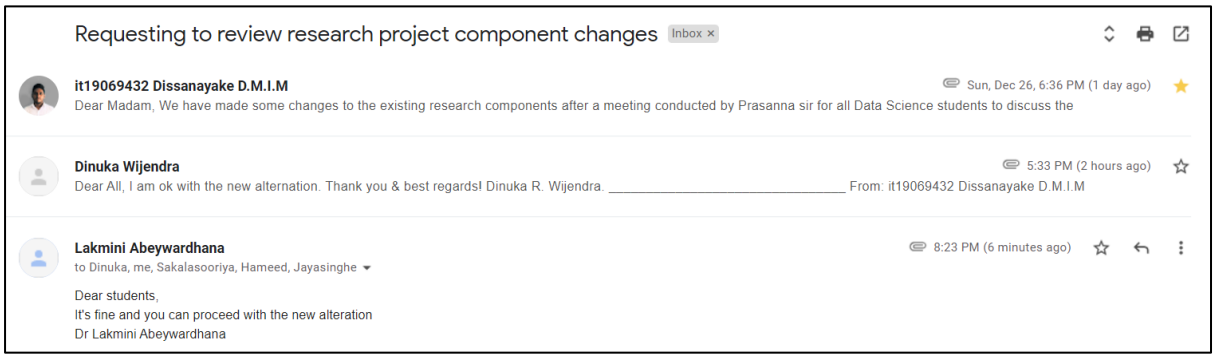

Figure 4. Research component alternation approval confirmation email sent by the supervisor

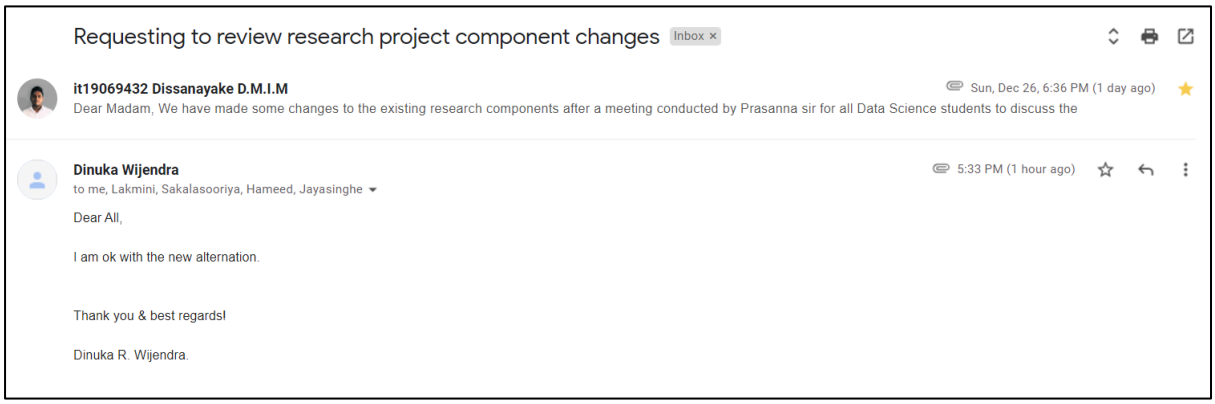

Figure 5. Research component alternation approval confirmation email sent by the co-supervisor

<u>This part will be filled by the Topic Screening Panel members</u>

Acceptable: Mark/select as necessary

Acceptance/ Rejection	Correction State	
	Minor Correction	Major Corrections
Accepted	☒	☐
Resubmit	☐	☐
Rejected	☐	

Corrections (if necessary)

Rephrase the title since it is not clear

Major changes proposed:

_____

_____

_____

_____

Any other Comments:

_____

_____

_____

Approved by the review panel:

Member's Name	**Signature**
Dr. Anuradha Karunnasena	
Ms. Sanjeevi Chandrasiri	
Ms. Lokesha Weerasinghe	
Ms. Thamali Dassanayake	

Important:

1. According to the comments given by the panel, do the necessary modifications and get the approval by the **same panel**.
2. If the project topic is rejected, find out a new topic and inform the CDAP Group for a new topic pre-assessment.
3. A form approved by the panel must be attached to the **Project Charter Form**.