**ENHANCING CONVERSATIONAL AI MODEL PERFORMANCE AND EXPLAINABILITY FOR SINHALA-ENGLISH BILINGUAL SPEAKERS**

2022-056

Project Proposal Report

Dissanayake D.M.I.M.

(Hameed M.S., Jayasinghe D.T., Sakalasooriya S.A.H.A.)

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

Department of Information Technology

Sri Lanka Institute of Information Technology

Sri Lanka

January 2022

**ENHANCING CONVERSATIONAL AI MODEL PERFORMANCE AND EXPLAINABILITY FOR SINHALA-ENGLISH BILINGUAL SPEAKERS**

2022-056

Project Proposal Report

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

Department of Information Technology

Sri Lanka Institute of Information Technology

Sri Lanka

January 2022

# Declaration, Copyright Statement and The Statement of The Supervisors

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** | **A picture containing text  Description automatically generatedSignature** |
| Dissanayake D.M.I.M. | IT19069432 | A picture containing diagram  Description automatically generated |
| Hameed M.S. | IT19064932 |  |
| Jayasinghe D.T. | IT19075754 | A picture containing text  Description automatically generated |
| Sakalasooriya S.A.H.A. | IT19051208 | A picture containing insect  Description automatically generated |

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

Name of the supervisor: Dr. Lakmini Abeywardhana

Signature of the supervisor: Date: 11/02/2022

Name of the co-supervisor: Ms. Dinuka Wijendra

Signature of the co-supervisor: Date: 11/02/2022

# Abstract

Machine learning model explainability and interpretability is an emerging research interest among Machine Learning and Artificial Intelligence researchers, and this field is known as Explainable Artificial Intelligence (XAI). It refers to the ability to explain how machine learning models make decisions. With the advancement of machine learning, it is often ambiguous to humans how state-of-the-art machine learning models give predictions. For instance, most machine learning models are black-box models, and how models like transformers generate predictions is often puzzling and even confusing. Explainability of the model predictions is considered crucial for various reasons, including identifying if the models have any biases, increasing the trustworthiness of the models, and finetuning or debugging the models. Explainable artificial intelligence has made its way towards the domain of natural language processing with the recent research conducted such as “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. This research component focuses on developing an explainable artificial intelligence algorithm called “DIME: dual interpretable model-agnostic explanations” by enhancing existing explainable AI techniques and utilizing the developed method to explain the predictions given by the “DIET Classifier”: the transformer-based deep learning text classification model used in Rasa conversational AI assistants. DIME algorithm focuses on using global explanations of the model to derive local model explanations, hence the word “dual” in DIME.

Keywords: Natural Language Processing, Explainable AI, DIET Classifier, DIME, Conversational AI

# **Table of Contents**

[Declaration, Copyright Statement and The Statement of The Supervisors i](#_Toc95507082)

[Abstract ii](#_Toc95507083)

[Table of Contents iii](#_Toc95507084)

[List of Figures v](#_Toc95507085)

[List of Tables vi](#_Toc95507086)

[List of abbreviations vii](#_Toc95507087)

[1. Introduction 1](#_Toc95507088)

[1.1 Background & Literature survey 1](#_Toc95507089)

[1.2 Research Gap 4](#_Toc95507090)

[1.3 Research Problem 6](#_Toc95507091)

[2. Objectives 10](#_Toc95507092)

[2.1 Main Objectives 10](#_Toc95507093)

[2.2 Specific Objectives 10](#_Toc95507094)

[3. Methodology 12](#_Toc95507095)

[3.1 Requirements Gathering and Analysis 12](#_Toc95507096)

[3.1.1 Functional requirements 12](#_Toc95507097)

[3.1.2 Non-functional requirements 13](#_Toc95507098)

[3.2 Feasibility Study 13](#_Toc95507099)

[3.2.1 Technical feasibility 13](#_Toc95507100)

[3.2.2 Financial feasibility 13](#_Toc95507101)

[3.2.3 Legal feasibility 14](#_Toc95507102)

[3.2.4 Operational feasibility 14](#_Toc95507103)

[3.2.5 Scheduling feasibility 14](#_Toc95507104)

[3.3 Preparation of Datasets 14](#_Toc95507105)

[3.3.1 General dataset for machine learning model training 15](#_Toc95507106)

[3.3.2 Domain specific dataset for conversational AI training 15](#_Toc95507107)

[3.4 Individual Component Architecture 16](#_Toc95507108)

[3.5 Machine Learning Model Training and Testing 18](#_Toc95507109)

[3.6 Generating Model Explanations 18](#_Toc95507110)

[3.6.1 Calculating global feature importance 18](#_Toc95507111)

[3.6.2 Generating local model explanations 21](#_Toc95507112)

[3.7 User Interfaces and Text Visualizations 22](#_Toc95507113)

[3.8 Tools and Technologies 23](#_Toc95507114)

[4. Description of Personal and Facilities 25](#_Toc95507115)

[5. Gantt Chart 26](#_Toc95507116)

[6. Commercialisation Plan 27](#_Toc95507117)

[7. Budget and Budget Justification 31](#_Toc95507118)

[References 32](#_Toc95507119)

[Appendices 35](#_Toc95507120)

[Appendix A: Survey Form 35](#_Toc95507121)

[Appendix B: Supervision Confirmation Emails 46](#_Toc95507122)

# List of Figures

[Figure 1. 1: Summarized Explainable AI approaches 3](https://mysliit-my.sharepoint.com/personal/cdap_sliit_lk/Documents/2022%20REGULAR/01%20Topic%20Evaluation/Student%20Submissions/Proposal%20Draft/TMP-22-102/IT19069432.docx#_Toc95256728)

[Figure 1. 2: Summarized types of model explanations 3](https://mysliit-my.sharepoint.com/personal/cdap_sliit_lk/Documents/2022%20REGULAR/01%20Topic%20Evaluation/Student%20Submissions/Proposal%20Draft/TMP-22-102/IT19069432.docx#_Toc95256729)

[Figure 1.3: Summary of survey responses received for the question "Do you know how AI-based chatbots make decisions?” 7](https://mysliit-my.sharepoint.com/personal/cdap_sliit_lk/Documents/2022%20REGULAR/01%20Topic%20Evaluation/Student%20Submissions/Proposal%20Draft/TMP-22-102/IT19069432.docx#_Toc95256730)

[Figure 1.4: Summary of the responses received for the survey question "Are you interested to know how machine learning/ deep learning models make decisions?" 8](https://mysliit-my.sharepoint.com/personal/cdap_sliit_lk/Documents/2022%20REGULAR/01%20Topic%20Evaluation/Student%20Submissions/Proposal%20Draft/TMP-22-102/IT19069432.docx#_Toc95256731)

[Figure 1.5: Summary of the responses received for the survey question "Do you know what model explainability or model interpretability of machine learning models is?" 9](https://mysliit-my.sharepoint.com/personal/cdap_sliit_lk/Documents/2022%20REGULAR/01%20Topic%20Evaluation/Student%20Submissions/Proposal%20Draft/TMP-22-102/IT19069432.docx#_Toc95256732)

[Figure 1.6: Summary of the responses received for the survey question "Have you used model explainability tools” 9](https://mysliit-my.sharepoint.com/personal/cdap_sliit_lk/Documents/2022%20REGULAR/01%20Topic%20Evaluation/Student%20Submissions/Proposal%20Draft/TMP-22-102/IT19069432.docx#_Toc95256733)

[Figure 3.1: High-level architectural diagram of DIME 16](https://mysliit-my.sharepoint.com/personal/cdap_sliit_lk/Documents/2022%20REGULAR/01%20Topic%20Evaluation/Student%20Submissions/Proposal%20Draft/TMP-22-102/IT19069432.docx#_Toc95256745)

[Figure 3.2: High-level architectural diagram of the purposed solution with all research components integrated 17](https://mysliit-my.sharepoint.com/personal/cdap_sliit_lk/Documents/2022%20REGULAR/01%20Topic%20Evaluation/Student%20Submissions/Proposal%20Draft/TMP-22-102/IT19069432.docx#_Toc95256746)

[Figure 3.3: A sample of predictions given by a fully trained Rasa model, including intent, intent ranking, and entities predicted using DIET classifier explored with Postman. 18](https://mysliit-my.sharepoint.com/personal/cdap_sliit_lk/Documents/2022%20REGULAR/01%20Topic%20Evaluation/Student%20Submissions/Proposal%20Draft/TMP-22-102/IT19069432.docx#_Toc95256747)

[Figure 3.4: Providing the original dataset and recording the model accuracy 20](https://mysliit-my.sharepoint.com/personal/cdap_sliit_lk/Documents/2022%20REGULAR/01%20Topic%20Evaluation/Student%20Submissions/Proposal%20Draft/TMP-22-102/IT19069432.docx#_Toc95256748)

[Figure 3.5: Providing the altered dataset by removing “IT” and recording the model accuracy 20](https://mysliit-my.sharepoint.com/personal/cdap_sliit_lk/Documents/2022%20REGULAR/01%20Topic%20Evaluation/Student%20Submissions/Proposal%20Draft/TMP-22-102/IT19069432.docx#_Toc95256749)

[Figure 3.6: Generating local explanations using global feature importance 22](https://mysliit-my.sharepoint.com/personal/cdap_sliit_lk/Documents/2022%20REGULAR/01%20Topic%20Evaluation/Student%20Submissions/Proposal%20Draft/TMP-22-102/IT19069432.docx#_Toc95256750)

[Figure 3.7: Proposed web interface wireframe of the DIME local server for model explanation visualizations. 23](https://mysliit-my.sharepoint.com/personal/cdap_sliit_lk/Documents/2022%20REGULAR/01%20Topic%20Evaluation/Student%20Submissions/Proposal%20Draft/TMP-22-102/IT19069432.docx#_Toc95256751)

[Figure 4.1: Work breakdown structure of the individual research component 25](https://mysliit-my.sharepoint.com/personal/cdap_sliit_lk/Documents/2022%20REGULAR/01%20Topic%20Evaluation/Student%20Submissions/Proposal%20Draft/TMP-22-102/IT19069432.docx#_Toc95256833)

[Figure 5. 1: Gantt chart for overall project and the individual component 26](https://mysliit-my.sharepoint.com/personal/cdap_sliit_lk/Documents/2022%20REGULAR/01%20Topic%20Evaluation/Student%20Submissions/Proposal%20Draft/TMP-22-102/IT19069432.docx#_Toc95256855)

[Figure 6.1: proposed commercialization plan 30](#_Toc95497340)

# List of Tables

[Table 1.1: Comparison of XAI research done that applies to the domain of NLP and methodology used in contrast to this research. 5](#_Toc95256952)

[Table 3.1: Sample training data and predictions 19](#_Toc95256964)

[Table 3.2: Sample altered training data and respective predictions 19](#_Toc95256965)

[Table 3.3: Summary of Tools and Technologies to be used according to the tasks 24](#_Toc95256966)

[Table 6.1: Feature variations and cost difference of packages proposed by the commercialisation plan 29](#_Toc95507153)

[Table 7.1: Budget justification for the overall research project 31](#_Toc95497390)

# List of abbreviations

Abbreviation Description

AI Artificial Intelligence

API Application Programming Interface

CLI Command Line Interface

CPU Central Processing Unit

DIET Dual Intent Entity Transformer

DIME Dual Interpretable Model-Agnostic Explanations

GDPR General Data Protection Regulation

GFI Global Feature Importance

GUI Graphical User Interface

GPU Graphics Processing Unit

IO Input/Output

IT Information Technology

LIME Local Interpretable Model-Agnostic Explanations

LFC Local Feature Contribution

NLP Natural Language Processing

NLU Natural Language Understanding

NLG Natural Language Generation

PDP Partial Dependency Plot

RAM Random Access Memory

SLIIT Sri Lanka Institute of Information Technology

SHAP Shapley Additive Explanations

TPU Tensor Processing Unit

UI User Interface

XAI Explainable Artificial Intelligence

CDD Conversation Driven Development

SaaS Software-as-a-service

CaaS Conversational AI-as-a-Service

# Introduction

## Background & Literature survey

Many top-tier applications and devices utilize Natural Language Processing (NLP) in this era of Artificial Intelligence. Almost all domains have adopted NLP in ample ways to perform Natural Language Understanding (NLU), Natural Language Generation (NLG), or both. One such popular application of NLP is chatbots, and its name has gradually become Conversational Artificial Intelligence (Conversational AI) after chatbots started to utilize artificial intelligence to generate highly natural human-like dialogs and conversations instead of executing hard-coded rules and pattern matching. Artificial Intelligence chatbots or conversational AIs use advanced machine learning models that train on human conversation data, and hence they are capable of handling complex queries presented to them and possess the ability to have a more natural conversation [3]. Traditional NLP depended on inherently explainable techniques, also known as white-box techniques, such as rules-based approaches and decision tree classifiers. However, with the rapid advancement of machine learning and deep learning models, their human interpretability has decreased over time due to their complicated nature. It is known as the trade-off between model explainability and accuracy [4]. These less-interpretable models and techniques are known as black-box models/black-box methods, such as word embeddings, encoder-decoder architecture, and other deep learning-based approaches [5].

Interpretability and explainability are two terms in Explainable AI that are considered interchangeable but can often be confusing since they have two distinct definitions [6] [7]. An interpretation is the process of mapping abstract concepts to a domain that humans can comprehend, while an explanation is the set of features that have contributed towards a specific decision made [4][8]. Explainability is the term mainly considered in Explainable AI since it focuses on generating explanations for the predictions made by non-transparent excessively complicated models to identify the set of features that contributed the most to the given prediction. Attempting to interpret such models without XAI might not always be practical. Although explanations are a part of interpretations, explanations alone do not always provide consistent model interpretations. As mentioned in [6] and [8], transparency of white-box models is a property of interpretability and not explainability so, the two terms must not be confused with one another. In general, less complicated models are well interpretable, while complex models require explanations generated to provide an acceptable interpretation of how the model gives predictions.

There are a few ways of implementing XAI. The main two approaches are known as intrinsic/self-explaining and post hoc. Intrinsic/self-explaining models are white-box models that are interpretable by design, while post hoc refers to generating explanations for existing black-box models [4]. Moreover, XAI approaches can be model-specific, model-agnostic, local interpretable, or global-interpretable [9]. Model-specific XAI techniques apply to a specific family of machine learning models, while model-agnostic XAI approaches are suitable for various types of models available. Recent research conducted in the XAI domain based on NLP has developed model-agnostic tools such as LIME [1] and SHAP [10] that have proven their value in contrast to the model-specific approaches. Different machine learning model types use these tools to generate model explanations, although their internal structure is diverse. Local interpretable explanation generation focuses on explaining a single data instance. Especially in NLP, for a data instance such as a sentence, paragraph, or question from a text corpus used to train models such as text classifiers can get explanations only considering the data instance itself but discarding how the model behaves for other data. This approach generates more data instances close to the original data instance based on distance matrices such as cosine similarity and trains an interpretable surrogate model that acts as a white-box model using those newly obtained data and text classifier model predictions for those data instances [1][11]. Shapley values [15] are significant when calculating the contribution of each feature while deriving explanations. The research paper [10] explains how Shapley values assist in calculating the individual contribution of features. In the global interpretable explanation generation approach, all data points the model trained on are considered. In other words, in global interpretable explanations generation, the behavior of the whole model is considered [9]. Feature importance is one of the properties that can interpret a model globally. It uncovers how significant a given feature is to a fully trained model considering all predictions done. There are numerous ways to calculate the global feature importance including, permutation feature importance and partial dependency plot-based (PDP) feature importance [16].

Text, application

Description automatically generated

Figure 1.1: Summarized Explainable AI approaches

Graphical user interface, text, application, chat or text message

Description automatically generated

Figure 1.2: Summarized types of model explanations

There are several reasons why Explainable AI should be a requirement rather than an optional tool in the AI domain. One of the reasons is to get a clear idea of how a machine learning model makes predictions. The generated explanations can aid in constructing feature importance visualizations and model interpretations, and it is crucial to enhance the human understandability of why the model predicted what it predicted. In [4], the research concludes how trusting black-box machine learning models has led to a pressing issue, why human interpretability for machine learning models is essential, and how XAI can help bridge the gap between interpretability and accuracy. New regulations like GDPR have introduced a right to explanation due to the rise of opaque and complex algorithms to enforce trust through acceptable and clear explanations. Another reason to generate model explanations is to validate complex machine learning models. Since tools such as LIME [1] and SHAP [10] visualize the contribution of the features towards a prediction made by the model, model explanations make it possible to validate if the model has considered acceptable features to make predictions. Moreover, it can identify any biases the model contains, and this process makes it possible to debug machine learning models. While there can be many other reasons why complex models should be made explainable, the reasons mentioned above should be more than enough to notice why model explainability is a must.

## Research Gap

Researchers in the XAI domain usually select one or many XAI approaches (Figure 1. 1) and one or many out of types of explanations (Figure 1. 2). Numerous research papers seem to have focused on explaining black-box models. Every XAI research that has introduced a tool has also comprised visualization techniques to display the contributions of the features. Recently published literature survey papers such as [3]-[5], [8], and [9] are immensely helpful for a researcher to get an idea about the existing research gap.

Recent research carried out in the XAI domain that covers the area of NLP, including LIME [1][11] and SHAP[10], has incorporated various explainability techniques to calculate and visualize local and global explanations. Similar research papers to the above-specified research papers have been listed in Table 1.1 in contrast to the methods incorporated in this research component to define the research gap [12]-[14].

Table 1.1: Comparison of XAI research done that applies to the domain of NLP and methodology used in contrast to this research.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Research Name** | **Can be utilized for NLP** | **Intrinsic/ Post hoc** | **XAI Scope (Local/ Global)** | **Model-specificity** | **Visualization Technique** | **Feature contribution calculation** |
| LIME | Yes | Post hoc | Local | Agnostic | Highlighted text + Raw contribution value plots | Cosine Distance + Local linear surrogate model + Ridge Regression |
| SHAP | Yes | Post hoc | Local or Global | Agnostic | Highlighted text + Raw SHAP value plots | LIME, DeepLIFT, and other approaches with  SHAP values. |
| Deep  LIFT | Yes (Deep SHAP) | Post hoc | Local | Specific | Refer DeepSHAP [10] | Back-propagating contributions |
| Self-Explain | Yes | Self-explaining | Local | Specific | Highlighted features | Regularization with explanation specific losses |
| DIME *(This Research)* | Yes | Post hoc | **Local** with aid of **Global** feature importance | Agnostic | Highlighted features + Raw contribution score plots | **Global feature importance** + Shapley Values based on **model confidence** |

As demonstrated in the Table 1.1, there are similar research papers in contrast to this research in terms of applicability for NLP, Intrinsic/Post hoc, Model Specificity, and Visualization techniques. However, this research focuses on a scope of explanations that is notably different from the methods described in other research papers. Here, the local model explanations will consider the impact of global feature importance, both as a weighting factor and a feature selection criterion, when generating model explanations. The other specialty is that the method discussed in this research will consider model confidence given by the DIET classifier when calculating the feature contributions towards the predictions made by the models.

## Research Problem

Algorithms and machine learning models are getting complex every second. Even widely used applications such as chatbots have employed machine learning and NLP due to their high performance and accuracy compared to traditional approaches. As previously explained, these complex algorithms and advanced machine learning techniques, such as deep learning, are opaque by design, and humans are having a hard time interpreting how these work under the hood and making predictions [4]. Although many XAI-based explainers exist, they cannot generate explanations for all machine learning model types. It is sometimes impossible to uncover model explanations for application-specific machine learning models, even with model-agnostic explainers such as LIME [1] and SHAP [10]. The machine learning model inside the chatbot framework Rasa is an example of such a model. The models trained and stored within Rasa contain a cluster of machine learning models, NLP pipeline components, and metadata. Rasa NLU is responsible for extracting the model, loading the pipeline components, and processing text data. Generating model explanations for text classifiers such as the DIET classifier is tedious since Rasa does not provide a way to examine it. Extracting prediction probabilities from DIET is also not directly supported, which is required by many model-agnostic text explainers [2][17]. However, it is better if there is a way to interpret these application-specific machine learning models by its users to trust the predictions given by the models used within these widely used frameworks.

Many text explainers in the XAI domain focuses on generating either local or global model explanations. Almost none of the research has considered that these two kinds of model explanations can be blended to derive accurate local model explanations [5]. There is a possibility that using global feature importance might improve the quality of local model explanations, though it remains undiscovered. This research concentrates on finding the impact of global feature importance on local model explanations.

A recent survey that studied the current trends in machine learning model interpretability and explainability confirms that the majority prefer having explanations for the machine learning model predictions. The population of interest of the survey was 110 Undergraduates of the Faculty of Computing of Sri Lanka Institute of Information Technology, and it was decided based on the technical nature of the survey questions. Refer to Appendix A: Survey Form to observe the complete list of survey questions and responses.

Chart, pie chart

Description automatically generated

Figure 1.3: Summary of survey responses received for the question "Do you know how AI-based chatbots make decisions?”

Responses in the Figure 1.3 clearly shows that more than three-fourths of the entire population have little to no understanding of the process of natural language understanding techniques used within chatbot frameworks. Figure 1.4 indicates that 91.8% of the whole population prefers having explanations for the predictions made by machine learning models. These results confirm that there is indeed a demand for model prediction explanations and humans prefer interpretable model explanations.

Chart, pie chart

Description automatically generated

Figure 1.4: Summary of the responses received for the survey question "Are you interested to know how machine learning/ deep learning models make decisions?"

Although 32.7% of the population has claimed (Figure 1. 5) that they are aware of the terms model interpretability and explainability, Figure 1.6 clearly illustrates that only 26.4% have used text explainer tools such as LIME and SHAP. 66.4% of the total population are unaware of model explainability and interpretability, and this is due to the following reasons.

1. Deriving explanations and achieving interpretability requires additional work. XAI tools are not a part of the machine learning toolkits of machine learning-based applications by design. For instance, machine learning-enabled chatbots do not contain any text explainers. Users must generate explanations by themselves manually, which requires machine learning expertise.
2. Machine learning models created by some modern AI-enabled applications are different from a usual machine learning model. For instance, in the conversational AI framework, Rasa, the machine learning model is a collection of machine learning models and components compressed into a single ".bin" file rather than a single model. Especially in Rasa, a user should know how to find the classification model files in the zipped model files and the classification model architecture to generate model explanations. Thus, it is not easy to use the existing model-agnostic text explainers. It is preferable if there is a universal way to get model explanations without manually inspecting the zipped model of machine learning-based frameworks such as Rasa.

Chart, pie chart

Description automatically generated

Figure 1.5: Summary of the responses received for the survey question "Do you know what model explainability or model interpretability of machine learning models is?"

Chart, pie chart

Description automatically generated

Figure 1.6: Summary of the responses received for the survey question "Have you used model explainability tools”

# Objectives

## Main Objectives

The main objective of the overall research project is to develop NLP tools for text preprocessing, feature engineering, training data and machine learning model improvements, model performance evaluation, and model explanation generation on DIET intent classifiers; that can be attached to a domain-specific Rasa conversational AI assistant designed for Sinhala-English code-switched text corpus.

The main objective of this individual research component is to develop an XAI algorithm, DIME (dual interpretable model-agnostic explanations), to deliver local model explanations with the help of global feature importance. As previously explained, currently, there is little to no strategy exists that tries to aid "global feature importance" to derive local model explanations. This implementation concentrates on deriving global feature importance for all tokens in the training data to investigate how important a word is to the fully trained model and utilizing these scores to derive local model-specific explanations.

## Specific Objectives

Specific objectives of this research component can be listed down as follows.

1. Develop a strategy to calculate the global feature importance for individual features and normalize the scores in a logically explainable manner.
2. Develop a strategy to incorporate global feature importance to derive local model-agnostic explanations for the DIET classifier in Rasa framework.
3. Integrate the above strategies as a modular python package that can be easily installed.
4. Develop a visualization technique to illustrate the local text explanations with the contribution scores towards the prediction in a human-interpretable way.
5. Build a Server with API endpoints that can be consumed by the frontend applications and integrate model-agnostic explanation visualizations with the conversational AI maintenance frontend.
6. Integrate the XAI approach developed with the Rasa framework seamlessly and allow Rasa users to get model explanations for the DIET classifier.

The Methodology chapter (chapter 3) of this report explains the main objective and specific objectives mentioned above in detail.

# Methodology

## Requirements Gathering and Analysis

The requirements gathering phase mainly focused on studying existing research on XAI concepts for NLP and the necessity of an XAI component for machine learning-based products, especially chatbot frameworks such as Rasa. After identifying the current requirement for an XAI component to interpret machine learning models used in modern conversational AI frameworks, the survey mentioned in section 1.3 was conducted to study the following regarding this research component.

1. The percentage of users who understand how modern artificial intelligence-based chatbots work.
2. The percentage of users who would like to know how machine learning models make predictions when responding to user queries

### Functional requirements

The functional requirements of the XAI approach, DIME, are as follows.

1. The proposed XAI approach DIME should logically calculate the global feature importance for any given fully trained DIET classifier.
2. DIME should normalize global feature importance scores between an acceptable range.
3. DIME should generate local model explanations using global feature importance scores as a part of its algorithm.
4. DIME should consider global feature importance to perform feature selection when generating local model explanations.
5. DIME should be applicable to any machine learning text classification model that outputs confidence scores for predictions.
6. DIME should ask the users number of features to generate explanations to reduce the number of calculations drastically.
7. DIME should visualize locally generated model explanations in a human-interpretable manner.
8. DIME should be seamlessly combinable with Rasa conversational AI assistants to generate explanations for non-technical users.

### Non-functional requirements

The non-functional requirements of DIME are as follows.

1. DIME should perform calculations efficiently.
2. DIME should provide reliable local model explanations.
3. DIME should provide simple and easy-to-interpret model explanation visualizations.
4. DIME algorithm should be modular.

## Feasibility Study

A feasibility study was carried out to investigate the technical, financial, legal, operational, and scheduling feasibility of the research component.

### Technical feasibility

The research component requires having in-depth knowledge in Rasa framework, Rasa machine learning models, data science, and software development. There should be enough computing resources to train Rasa machine learning models. Training machine learning models in Rasa does not require advanced computing resources such as GPUs, TPUs, or High RAM/CPU instances, although it is better if such resources are available. Hard drive space should be a foremost concern as the Rasa framework may need a fair amount of disk space for caching model files. There should be a properly configured cloud infrastructure for the overall deployment of the final product.

### Financial feasibility

The cloud infrastructure for training machine learning models or deployment should not be costly. There can be minor charges for domain name purchasing and cloud infrastructure resource usage, and they are acceptable since the final product contains a market value that can cover the costs involved.

### Legal feasibility

Datasets used in the research components should be publicly available datasets. Any scraped data from websites that are not public should have the approval of the original owners and should have written consent. The research component should not extract content as it is from previous work done without the written permission of the original authors. Any python packages or other software used must be open source or legally purchased, and the original authors must be credited if requested.

### Operational feasibility

This research component must not conflict with other research components of the same research project or other research projects. The research component should address the gap in the explainability of the DIET classifier and should preserve novelty in contrast to the previous work done. The research component should be viable to fulfill the scope stated by the research project requirements.

### Scheduling feasibility

The tasks of the research component should adhere to the Gantt chart mentioned in section 5 of this proposal report and the pre-defined research project milestones.

## Preparation of Datasets

All research components require two datasets in total. Both datasets are domain-specific datasets with Sinhala-English code-switched text data. However, there are notable differences between the datasets and subsections 3.3.1 and 3.3.2 explain these differences clearly. Both datasets utilize data augmentation techniques mainly to overcome the low-resource nature of Sinhala text data gathered, capture as many as code-switched phrases and collect as many distinct writing patterns as possible. All research group members will generate four versions of the samples in both datasets according to different code-switching styles. Duplicate data removal will be employed to ensure the quality of the collected data.

This research component only utilizes the domain-specific dataset mentioned in subsection 3.3.2. to train a Rasa conversational AI model since developing DIME only requires a fully trained Rasa model.

### General dataset for machine learning model training

The first domain-specific dataset for machine learning model training and NLP text pre-processing tool development contains scraped Sinhala-English code-switched 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). In addition to that, publicly available documents, such as PDFs and Docx files taken from the above SLIIT-based websites since they are both public and official. Dataset will utilize data augmentation techniques to overcome the low-resource issue.

### Domain specific dataset for conversational AI training

The second domain-specific dataset for training the Rasa-based conversational AI contains handcrafted Sinhala-English code-switched textual data. This dataset should be designed carefully as a training dataset for the intent classification task of conversational AI according to the guidelines provided by Rasa. As in the previous case, the second dataset also will utilize data augmentation techniques to overcome the low-resource issue. There will be around 78 intents (classes), and each class initially contains a minimum of 10 examples as a standard. Note that the number of training examples may increase in the future to increase the overall performance of conversational AI.

## Individual Component Architecture

The implementation of DIME, the XAI approach developed in this research component, will contain several sub-components, as illustrated in Figure 3.1.

Diagram

Description automatically generated

Figure 3.1: High-level architectural diagram of DIME

The core python package of DIME will contain two sub-packages to calculate the global feature importance and generate local model explanations. The "local explainer" module will use the results generated by the "global explainer" module. The global explainer will utilize the data and model handling modules to perform the required IO operations. The model handler will execute model loading, and the data handler module will handle training data-related operations. DIME will consist of a CLI interface and a local server built with Flask to reveal visualizations for the explanations generated. CLI tool and Notebooks will only get raw values instead of visualizations, leaving room for further improvements to the DIME module.

Graphical user interface

Description automatically generated

Figure 3.2: High-level architectural diagram of the purposed solution with all research components integrated

Figure 3.2 illustrates how DIME fits into the overall system architecture after integrating all research components. The "DIME XAI component + Flask API" is where DIME fits in. DIME component will use a docker container for the deployment, and it will take advantage of the reverse proxy to avoid directly exposing the server contents and the port for security concerns.

## Machine Learning Model Training and Testing

The proposed XAI component requires a fully trained Rasa model with a DIET classifier included as an NLP pipeline component. The Rasa NLU component of the Rasa package allows to load a trained model and parse data to get predictions. The model predictions include intent predicted by the DIET classifier and the entities extracted (Figure 3.3). DIME will use the confidence scores given with the model predictions of the DIET classifier to generate global and local model explanations. The Rasa conversational AI will be trained using the dataset mentioned in subsection 3.3.2.

A screenshot of a computer

Description automatically generated with medium confidence

Figure 3.3: A sample of predictions given by a fully trained Rasa model, including intent, intent ranking, and entities predicted using DIET classifier explored with Postman.

## Generating Model Explanations

### Calculating global feature importance

The global feature importance calculation will take the intent predictions for all training data instances given by a fully trained Rasa model with a DIET classifier attached. The DIME algorithm will take the training dataset instead of a separate testing dataset to capture feature importance for all tokens that the model has seen during training. When calculating the global importance of a specific word, the DIME algorithm will get the predictions from the original model for the training data instances by removing the token of interest from the training data. It will perform this process repeatedly and find the probability of the correct model predictions. The algorithm will then consider the change in probability of accurate model predictions as the global feature importance of the token of interest.

Consider the sample training data instances and the corresponding predictions given in Table 3.1. First, the algorithm finds the probability of correct predictions without alternating training data. In the following case, the percentage of accurate predictions given by the model is about 0.6667. If the token of interest is "IT", it will be removed from training, and the probability of correct predictions given by the model is calculated. The percentage of accurate predictions is around 0.3333 after the alternations are done (Table 3.2). Then the change in probability of accurate model prediction is calculated and assigned as the global feature importance of the token "IT". Thus, the feature importance of IT roughly equals to 0.3334. The above process is repeated for all terms, and the results will be cached to minimize the number of repeated calculations.

Table 3.1: Sample training data and predictions

|  |  |  |
| --- | --- | --- |
| **Original training data** | **Predicted Class** | **Is correct prediction** |
| SLIIT එකේ තියන IT Degrees මොනවද? | Ask\_Hotline | Incorrect |
| IT ඩිග්‍රියක් කරන්න ඕන requirements මොනවද? | Ask\_Requirements | Correct |
| ස්ලිට් එකේ Data Science Degrees තියනවද | Ask\_Degrees | Correct |

Table 3.2: Sample altered training data and respective predictions

|  |  |  |
| --- | --- | --- |
| **Altered training data after removing the token “IT”** | **Predicted Class** | **Is correct prediction** |
| SLIIT එකේ තියන Degrees මොනවද? | Ask\_Hotline | Incorrect |
| ඩිග්‍රියක් කරන්න ඕන requirements මොනවද? | Ask\_Requirements | Incorrect |
| ස්ලිට් එකේ Data Science Degrees තියනවද | Ask\_Degrees | Correct |

Text

Description automatically generated with medium confidence

Figure 3.4: Providing the original dataset and recording the model accuracy

Text

Description automatically generated

Figure 3.5: Providing the altered dataset by removing “IT” and recording the model accuracy

The implementation of the research component will attempt another approach called permutation feature importance as an auxiliary approach to calculate the global feature importance. Instead of removing features from training data, permutation feature importance randomly shuffles the value of a token to calculate the mean decrease in the probability of accurate model predictions.

### Generating local model explanations

The local model explanations consider a single data instance to generate model prediction explanations. The implementation incorporates Shapley value calculation to find the contribution of each word in each data instance towards the prediction given by the model. Shapley values are calculated by generating new data instances by constructing subsets of features of original instance tokens and finding the corresponding change in the model confidence score. Calculation of change in model confidence for each subset of tokens of an instance can be computationally expensive. Thus, local model explanations will utilize global feature importance found as explained previously as a technique to perform feature selection. However, a user must state the number of tokens and the feature selection technique to be used, considering them as hyperparameters. An alternative method will be provided to pass the minimum global feature importance as a percentage to only select features with an importance score equal to or higher than the specified value. For this, the global feature importance scores for all tokens will be sent through a Softmax function. A Softmax function is a mathematical function that can fit all global feature importance scores to a probability distribution that sums up to 1. The local model explanation scores also will be sent through a Softmax function to get the contribution of each token towards the prediction as a percentage/probability.

Diagram

Description automatically generated

Figure 3.6: Generating local explanations using global feature importance

## User Interfaces and Text Visualizations

DIME will mainly have two interfaces, including a web interface and a CLI. The CLI will allow generating explanations and viewing the raw scores in the terminal. The CLI is useful when DIME is executed within a server with no access to a GUI or if it is impossible to run the webserver. However, the CLI interface will be less interpretable. The web interface will comprise interpretable explanations with word highlighting and feature contribution scores towards a specific prediction.

Graphical user interface

Description automatically generated

Figure 3.7: Proposed web interface wireframe of the DIME local server for model explanation visualizations.

## Tools and Technologies

Python 3.8 will be used for the implementation of the DIME XAI algorithm. Python packages such as NumPy, Pandas, and Sklearn will be utilized for handling the dataset. Rasa 2.8.12 will be used to handle Rasa machine learning models and train the conversational AI for both testing and final implementation. PyCharm, Visual Studio Code, and Google CoLab will be used as Integrated Development Environments. Anaconda Python distribution will be used for convenient python environment management.

A local server will be developed using Flask to build the required API endpoints and the interactive web application for the visualizations. DIME package and required Rasa source code scripts will be included in a docker container for efficient deployment. The DIME docker container will be integrated with the docker containers of packages from the other research components using the docker-compose tool, and the overall system will be deployed on either a GCP or an EC2 instance according to the final system requirements. A reverse proxy will be set up using Caddy Server 2.0 and NGINX webservers to secure the production server by minimizing the exposed ports to the public. The summary of the tools to be used is as follows.

Table 3.3: Summary of Tools and Technologies to be used according to the tasks

|  |  |
| --- | --- |
| **Task** | **Tools to be used** |
| Algorithm Implementation and DIME package development | Python 3.8, NumPy, Pandas, sklearn, Rasa 2.8.12, PyCharm, Visual Studio Code, Google CoLab |
| Server development as a standalone frontend for the DIME visualizations | Flask, JavaScript, Bootstrap, React JS (optional), Chart JS, Streamlit |
| NoSQL Database development | MongoDB Atlas, MongoDB Compass |
| Cloud Infrastructure management | One out of Google Cloud Platform or Amazon Web Services |
| Conversational AI development by Integrating all research components (Backend) | Rasa 2.8.12, MongoDB Atlas, Gensim, spaCy, Docker |
| Conversational AI frontend development | React JS, Bootstrap, socket.io, JavaScript |
| Overall system deployment | Caddy 2, NGINX, Git, Docker, docker-compose |

# Description of Personal and Facilities

Timeline

Description automatically generatedFigure 4.1 breaks down all tasks of this individual research component under 5 main tasks. Deployment is done at the end by integrating all four research components.

Figure 4.1: Work breakdown structure of the individual research component

# Gantt Chart



Figure 5. 1: Gantt chart for overall project and the individual component

# Commercialisation Plan

Each research component of the overall research project provides applications for various Natural Language Processing and Machine Learning model evaluation-related tasks. Although the outputs of each research component can help design many products and services, they were all integrated to build a conversational AI that fully supports Sinhala-English code-switching. The main reason for developing a conversational AI are as follows.

1. Conversational AIs are a trending topic, and there is an increasing demand for Sinhala-based conversational AIs. Many businesses are looking to increase their customer reach by using conversational AIs for a vast range of business tasks, including providing technical assistance, automating manual tasks such as booking tickets, and providing the information requested by the customers. Although that is the case, it is hard to find Sinhala-based conversational AIs, especially in Sri Lanka. Thus, developing Sinhala-based conversational AIs was identified as a potential business opportunity.
2. Although there are many conversational AI development frameworks, most of them lack evaluation tools to debug machine learning models. In frameworks like Rasa, model evaluations are highly technical, and the average developers without machine learning knowledge fail to identify problems and debug the machine learning models. Solving this issue by allowing non-technical users to maintain and evaluate machine learning models and conversational AIs is another business opportunity where evaluation tools can be built and released as add-on features.
3. Businesses are moving towards cloud-based solutions, especially SaaS products from traditional standalone applications, web applications, and on-premises solutions, where the maintenance cost is high. A cloud-based highly configurable conversational AI would be an appropriate solution for many businesses where the effort for maintenance is considerably low.

The end product of the overall research concentrates on designing a solution for the above potential business ideas and opportunities. The conversational AI developed as the research end-product is mainly a SaaS or simply a CaaS (conversational AI-as-a-service) product that a business can purchase with a set of add-on features, as illustrated in Figure 6.1. In addition to the CaaS packages, on-premises and demo cloud-based conversational AI packages are available for affordability and convenience. The end product of the overall research has the following archivable user benefits.

1. Businesses can purchase a CaaS package and eliminate conversational AI maintenance efforts.
2. Developers can use code-less maintenance tools to maintain conversational AIs they purchase without having in-depth knowledge about the backend deployment and the conversational AI development framework.
3. Businesses can purchase evaluation tools as add-ons and generate model explanations and evaluate machine learning models themselves to avoid extra maintenance costs. Here, the generated evaluations are easy to understand by non-technical users, which is not a feature of any existing chatbot framework.
4. Users of conversational AI can easily use the Sinhala-English code-switchable keyboard interface, and businesses can attract more users from having this feature as it eliminates the need to use third-party Sinhala typing services.
5. Businesses have the freedom to purchase either the CaaS packages or on-premises packages as per their need, while anyone can test the demo conversational AI for a period of 1 month before deciding to purchase any of the other packages that have a cost assigned.
6. Businesses can reach a vast customer range through Sinhala-English code-switching-based conversational AIs, especially within the Sri Lankan market, and it is possible to eliminate the need for having a dedicated customer care staff. It will dramatically lower the expenses of the business.

The proposed commercialisation plan of the end product of the overall research component contains a set of convenient packages and the offered features of each package differ based on the cost attached to it. The distribution of the features and the package cost were carefully planned and designed by analyzing the existing purchasable packages of similar SaaS products and conversational AIs. Table 6.1 clearly illustrates the feature variation and the cost difference of the packages offered by the proposed commercialisation plan.

Table 6.1: Feature variations and cost difference of packages proposed by the commercialisation plan

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Feature** | **Packages** | | | | |
| **Demo** | **On-Prem** | **CaaS** | | |
| **Starter** | **Pro** | **Genius** |
| Intents | 10 | Unlimited | 20 | 180 | 400 |
| API Integrations | 2 | Unlimited | 2 | 110 | 200 |
| Bot Analytics |  |  | - |  |  |
| CDD |  |  |  |  |  |
| Sinhala Entity Annotating |  | - | - |  |  |
| ML Evaluation Tools | - | - | - | - |  |
| Maintenance Fee | - | 2 Free + $9.99 per additional call | - | - | - |
| Trial Duration | 1 Month | - | - | - | - |
| Package Price | Free | $199.99 | $9.99 | $34.99 | $49.99 |

A picture containing text, outdoor

Description automatically generated

Figure 6.1: proposed commercialisation plan

# Budget and Budget Justification

The budget justification for all four research components is mentioned in Table 7.1.

Table 7.1: Budget justification for the overall research project

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Component Name** | **Individual Item Price (LKR)** | **Number of Items** | **Duration** | **Total Item Price (LKR)** |
| Domain Name | 2148.43/year | 1 | 1 year | 2148.43 |
| GCP Instance | 10683.84/month | 1 | 6 months | 64103.06 |
| Reference Book: Basaka Mahima by J.B. Dissanayake | 1250.00 | 1 | - | 1250.00 |
| Research Paper Publication | 25000.00 | 1 | - | 25000.00 |
| Grand Total | - | - | - | 92,501.49 |

# References

1. 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>
2. T. Bunk, D. Varshneya, V. Vlasov, A. Nichol, “DIET: Lightweight Language Understanding for Dialogue Systems,” 2020. [Online]. Available: <https://arxiv.org/abs/2004.09936>
3. G. Caldarini, S. Jaf, K. McGarry, “A Literature Survey of Recent Advances in Chatbots,” 2022. [Online]. Available: <https://arxiv.org/abs/2201.06657>
4. F. K. Došilović, M. Brčić and N. Hlupić, "Explainable artificial intelligence: A survey," 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2018, pp. 0210-0215, doi: 10.23919/MIPRO.2018.8400040.
5. M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, P. Sen, “A Survey of the State of Explainable AI for Natural Language Processing,” 2020. [Online]. Available: <https://arxiv.org/abs/2010.00711>
6. Y. Jin and B. Sendhoff, "Pareto-Based Multiobjective Machine Learning: An Overview and Case Studies," IEEE Trans. Syst. Man Cybern. Part C Appl. Rev., vol. 38, no. 3, pp. 397-415, May 2008.
7. G. Montavon, W. Samek, and K.-R. Muller, "Methods for interpreting and understanding deep neural networks," Digit. Signal Process., vol. 73, pp. 1-15, Feb. 2018
8. S. R. Islam, W. Eberle, S. K. Ghafoor, en M. Ahmed, “Explainable Artificial Intelligence approaches: A survey,” 2021. [Online]. Available: <https://arxiv.org/abs/2101.09429>
9. P. Gohel, P. Singh, en M. Mohanty, “Explainable AI: current status and future directions”, 2021. [Online]. Available: <https://arxiv.org/abs/2107.07045>
10. S. Lundberg, S. Lee, “A Unified Approach to Interpreting Model Predictions,” 2017. [Online]. Available: <https://arxiv.org/abs/1705.07874>
11. 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>
12. Xainlp2020.github.io, “*XAI for Natural Language Processing*”, [online] Available at: <https://xainlp2020.github.io/xainlp> [Accessed 24 January 2022].
13. D. Rajagopal, V. Balachandran, E. Hovy, en Y. Tsvetkov, “SelfExplain: A self-explaining architecture for neural text classifiers”, 2021. [Online]. Available: <https://arxiv.org/abs/2103.12279>
14. A. Shrikumar, P. Greenside, en A. Kundaje, “Learning important features through propagating activation differences”, 2017. [Online]. Available: <https://arxiv.org/abs/1704.02685>
15. Lloyd S Shapley. “A value for n-person games”. In: Contributions to the Theory of Games 2.28 (1953), pp. 307–317.
16. C. Molnar, 2022. “*Interpretable Machine Learning*”, 2022. [online] Christophm.github.io. Available at: <https://christophm.github.io/interpretable-ml-book> [Accessed 24 January 2022].
17. 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>

# Appendices

## Appendix A: Survey Form

Figure A.1: Complete survey form questions and responses – part 1

Graphical user interface, application

Description automatically generated

Chart, pie chart

Description automatically generated

Figure A.2: Complete survey form questions and responses – part 2

Chart, pie chart

Description automatically generated

Figure A.3: Complete survey form questions and responses – part 3

Chart, application, pie chart

Description automatically generated

Figure A.4: Complete survey form questions and responses – part 4

Chart, pie chart

Description automatically generated

Figure A. 5: Complete survey form questions and responses – part 5

Chart, pie chart

Description automatically generated

Figure A.6: Complete survey form questions and responses – part 6

Chart, pie chart

Description automatically generated

Figure A.7: Complete survey form questions and responses – part 7

Graphical user interface, chart, application, pie chart

Description automatically generated

Figure A.8: Complete survey form questions and responses – part 8

Chart, application, pie chart

Description automatically generated

Figure A.9: Complete survey form questions and responses – part 9

Chart, pie chart

Description automatically generated

Figure A.10: Complete survey form questions and responses – part 10

Graphical user interface, application

Description automatically generated

Figure A.11: Complete survey form questions and responses – part 11

## Appendix B: Supervision Confirmation Emails

Graphical user interface, text, application, email

Description automatically generated

Figure B.1: Research project supervision confirmation email

Graphical user interface, text, application, email

Description automatically generated

Figure B. 2: Research project co-supervision confirmation email