UTILIZING REVERSE-STEMMING AND OTHER TECHNIQUES TO DEVELOP A DATA ANNOTATION TOOL FOR CODE-MIXED TEXT DATA FOR EFFICIENT CUSTOM ENTITY TAGGING

Sakalasooriya S.A.H.A.

(IT19051208)

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

Department of Computer Systems Engineering

Sri Lanka Institute of Information Technology
Sri Lanka

September 2022

UTILIZING REVERSE-STEMMING AND OTHER TECHNIQUES TO DEVELOP A DATA ANNOTATION TOOL FOR CODE-MIXED TEXT DATA FOR EFFICIENT CUSTOM ENTITY TAGGING

Sakalasooriya S.A.H.A.

(IT19051208)

The dissertation was submitted in partial fulfilment of the requirements for the B.Sc. Special Honours degree in Information Technology

Department of Computer Systems Engineering

Sri Lanka Institute of Information Technology
Sri Lanka

September 2022

DECLARATION, COPYRIGHT STATEMENT AND THE STATEMENT OF THE SUPERVISORS

I declare that this is my own work, and this dissertation 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 my 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. Also, I hereby grant to Sri Lanka Institute of Information Technology the non-exclusive right to reproduce and redistribute my dissertation in whole or part in future works (such as articles or books).

Name	Student ID	Signature
Sakalasooriya S.A.H.A.	IT19051208	

The above candidate is carrying out research for the undergraduate dissertation under my supervision.

Name of the supervisor: Dr. Lakmini Abeywardhana
Signature of the supervisor: Date: 09/09/2022

Name of the co-supervisor: Ms. Dinuka Wijendra
Signature of the co-supervisor: Date: 09/09/2022

ABSTRACT

Named entity recognition is a key component in conversational AIs. It helps conversational AI to identify the key elements in a given text, as an example names of people, places, brands, monetary values etc. This process is done by artificial neural networks or statistical methods. Annotated text data is a vital requirement to train a name entity recognition model. therefore. Training a name entity recognition model requires considerable amount of annotated text dataset. Text annotations is a tedious task because it requires domain knowledge and human interaction for each word. Therefore, few tools are available for text data annotation for make the process easier, but there is no optimized tool for Sinhala - English code-switching text data annotation. Sinhala is a native language in Sri Lanka which is being used by approximately 19 million people. Most of the Sinhalese use Sinhala - English code-switching typing style when they are typing. This paper presents a methos to increase the efficiency of text annotation task by providing name entity suggestions during the Sinhala-English code-switching text data annotation.

Keywords: custom named entity tagging, text annotating, named entity recognition, code-switching text data

ACKNOWLEDGEMENT

I would like to extend my sincere gratitude to all the parties that helped me to complete this research including Dr Lakmini Abeywardhana (research team supervisor), and Ms Dinuka Wijendra (research team co-supervisor) for their assistance, direction, and suggestions. This project would not have been achieved without their valuable help. In addition, I am grateful to the Sri Lanka Institute of Information Technology for the support they have provided by conducting support sessions and document templates. They provided detailed student handbook and rule books which I used to create the conversational AI dataset that was used in this research. A well-prepared research timeline, document templates, and instructions were made possible with help from the SLIIT Research Project Team (CDAP) and academics who led essential Research Project Module (RP) sessions. Again, I would like to be thankful to all my lecturers and teachers to give me priceless knowledge.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

The list of all the abbreviations used in this report is in the following table.

Abbreviation	Description
AI	Artificial intelligence
AIML	Artificial Intelligence Mark-up Language
CLI	command line interface
CNN	Convolutional neural network
GUI	graphical user interface
LSTM	Long short-term memory
ML	Machine learning
NER	named entity recognition
NLP	Natural language processing
NLU	Natural language understanding
RNN	Recurrent Neural Network
XML	Extensible markup language
CDD	Conversation Driven Development
CaaS	Chatbot as a service

1. INTRODUCTION

1.1 Background & Literature survey

Humans are daily engaging with various kinds of data types to fulfil their requirements. As an example, text data when reading news and magazines, audio data when listening to music or radio, images when watching pictures, and videos when watching films. These observations are interpreted by the human brain to extract the required information to guide our actions or feelings. Text data represents the human languages which are used to speak and write human thoughts. Since humans are social creatures text data is one of the most popular forms of media among others. As a result of advancements in machine learning technologies, there are many machine learning techniques and models specially designed for text data. Before feeding text data into machine learning models there are some pre-processing steps to follow such as lower casing. removal of punctuations. removal of emojis. removal of stop words, stemming, lemmatization and stop word removal. These techniques are used to enhance the results. Text annotation also can be considered as a pre-processing task. A metadata tag is provided during the annotation process to provide features to text data. This information provides tags that indicate criteria like words, phrases, and feelings. This helps to understand the meanings or emotions behind human speech during training an NLP model. The objective is to improve the NLP model's understandability of spoken or written language. Large volumes of annotated data are used by NLP algorithms and AI models. Therefore, text annotation must be done accurately and completely due to its widespread use. This allows computers to read, comprehend, evaluate, and produce text in ways that are useful for technological interactions with people more effectively. According to the 2020 State of AI and Machine Learning report [2], 70% of businesses are using text as a type of their AI products. This allows businesses to cost savings and revenue generation across all industries.

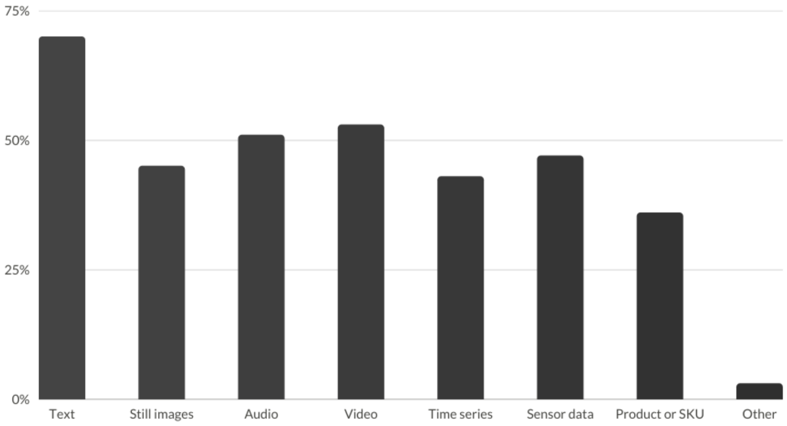


Figure 1.1. 1. 2020 State of AI and Machine Learning report – Data type usage in businesses

The significance of model training with high-quality text data is significantly important for machines to become more adept at understanding human language. In all circumstances, accurate, thorough text annotation is the first step in the creation of accurate training data. AI models will perform poorly with grammatical flaws or issues with clarity, understanding the context, and misinterpretations because of poorly done text annotations. If AI models train on correctly annotated text data, AI models will learn to perform effectively enough for the natural language. [3] These models can handle the more routine and repetitive tasks that are handled by people such as digital customer assistance, intelligent help desks, call centre management, language assistants, machine translators, customer feedback analysis, and more. This enables a business to concentrate on more strategic goals by freeing up time, money, and resources. Additionally, this will increase customer satisfaction as well, because of the quick responses and all-day availability. According to the survey most people tend to use conversational AIs to archive their goals.

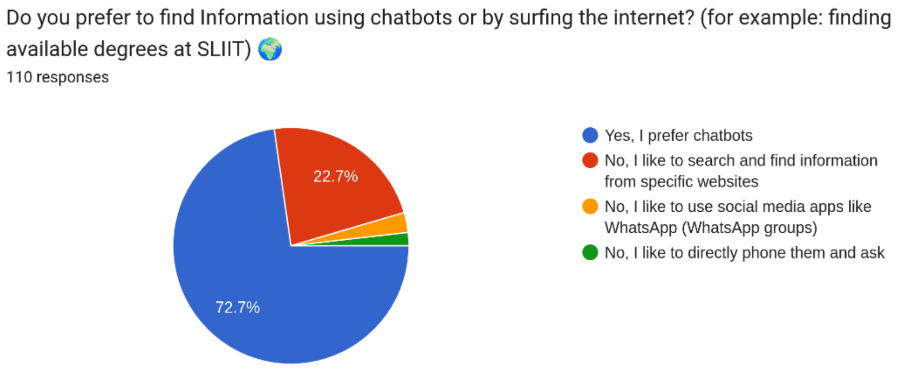


Figure 1.1. 2. Responses for do you prefer to find information using chatbots or by surfing the internet? (For example: finding available degrees at SLIIT)

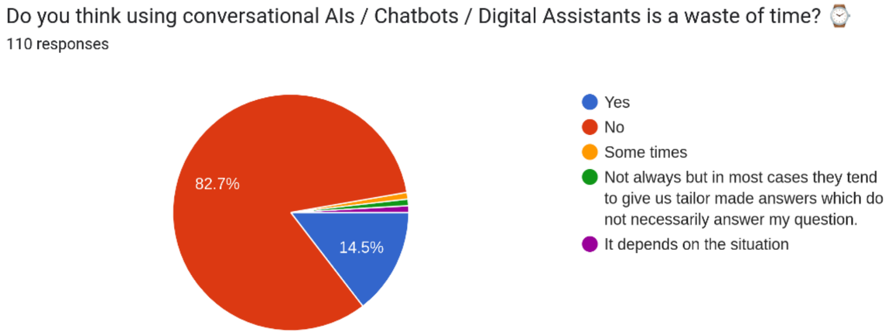


Figure 1.1. 3. Responses for do you think using conversational AIs / Chatbots / Digital Assistants is a waste of time?

For organizations in every domain, the ability to simplify processes by using high-quality text data is having a significant impact on customer satisfaction and financial performance. Text annotations can be of many different forms, sentiment annotation, intent annotation, semantic annotation, and relationship annotation. These annotation types are applied to many different human languages. Sentiment Annotation classifies text as positive, negative, or neutral after evaluating the attitudes and feelings by understanding the context. Text underlying need or desire of the meaning, analysed by intent annotation. Which divides it into various categories. As an example, request, command, or acknowledgement. Semantic text annotation is given additional tags that refer to numerous concepts and entities. As an example, persons, locations, or subjects. The relationship annotation aims to depict various connections between various elements of the document. Resolving dependencies and coreferences are examples of relationship annotation. These text annotation techniques are applied to conversational AIs as well, also known as chatbots.

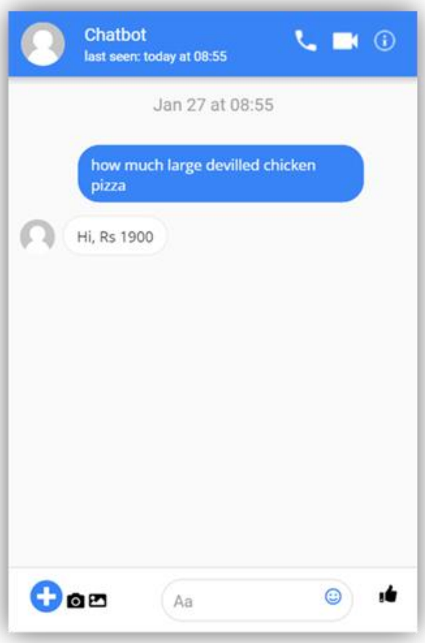
Conversational AI is software that simulates human conversations. Through the advancement of NLP and ML technologies, chatbots can understand human needs through textual inputs.[3] Therefore, chatbots can use to increase customer service and satisfaction at anything by placing a chatbot when customer interaction happens with the business. When a user asks a question from a chatbot it needs to understand the input text. This process is called natural language understanding (NLU). The named entity recognition (NER) approach can be used for NLU tasks. [4][5] There are neural network-based models (Bidirectional LSTM, and Bidirectional LSTM-CNNs) and statistical models such as conditional random field to identify name entities. In this research, only neural network-based models are considered. Neural network-based models require a considerable amount of training data to converge by adjusting weights. Usually, when building a Conversational AI, it should focus on a domain, otherwise, chatbots will not perform well according to business requirements. Therefore, the data set should be created according to the selected domain, by covering the domain-specific vocabulary.

As previously mentioned, [6] adding linguistic information to a corpus is called text annotation or named entity tagging. Named entity tagging is responsible for indicating the classes of words in a corpus. As an example, named entities are predefined categories of real-world things, such as a person, location, time, expressions, organization, and product. Also, it can be a domain-specific class of words. If the corpus contains an enormous number of words, it takes a considerable amount of time to do the tagging process. Because every word in the corpus should be read and identified by the person who is doing the annotation task. Therefore, creating data sets for named entity recognition models is a very time-consuming task and it requires expert knowledge.

Table 1.1. 1. Examples for domain specific named entities

Domain	Name entity	Example words
Biological domain [7]	virus	Coronavirus, Coxsackievirus, Dugbe
	bacteria	Hafnia spp, Escherichia coli, Citrobacter koseri
Commerce domain	Currency type	Yen, LKR, USD (United States Dollar), ₹
	payment method	visa card, master card, PayPal, crypto

Named entity recognition helps to determine the most suitable intent for a given text. This is done by the intent classification component of the chatbot. In this case, the text is the user's question, the intent is the chatbot's reply. If the chatbot does not have an ability to identify name entities chatbot will work in form filling way as shown in Figure 1.1. 5

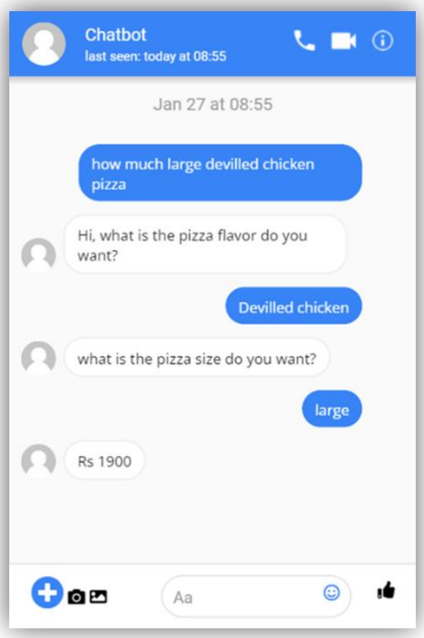


Jan 27 at 08:55

how much large devilled chicken pizza

Hi, Rs 1900

Figure 1.1. 4. Chatbot with named entity identification



how much large devilled chicken pizza

Hi, what is the pizza flavor do you want?

Devilled chicken

what is the pizza size do you want?

large

Rs 1900

Figure 1.1. 5. Chatbot without named entity identification

As shown in Figure 1.1. 5 when the user asks, “how much large, devilled chicken pizza,” the chatbot failed to identify the size and flavour of the pizza. If the chatbot could be able to identify named entities, the chatbot can provide the price of the pizza in the first reply by understanding flavour type and the size of pizza as shown in Figure 1.1. 5. According to the survey results, most of the people like to interact chatbots with automatically identifying name entities instead of form filling style.

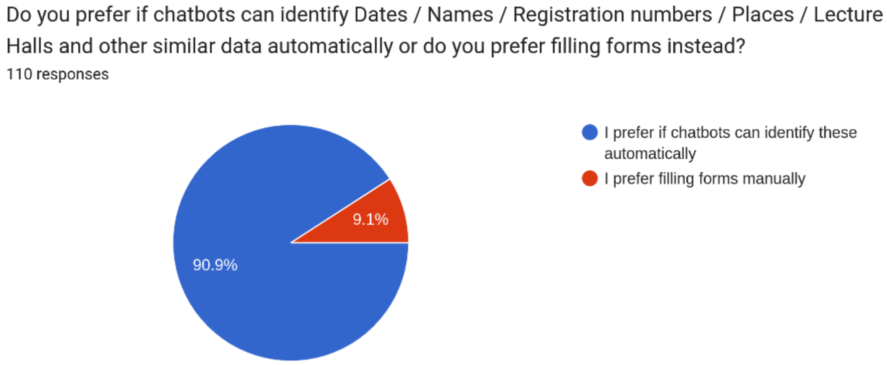


Figure 1.1. 6. Responses for do you prefer if chatbots can identify Dates / Names / Registration numbers / Places / Lecture Halls and other similar data automatically or do you prefer filling forms instead?

1.2 Research Gap

Researchers in named entity recommendation area are used diverse kinds of approaches to implement automated and semi-automated text annotation tools. [8] BRAT is a web-based annotating tool that has a named entity recommendation system using statistical method. It uses a semantic class disambiguation system with numerous outputs and probability estimations to reducing ambiguity of auto annotation. Additionally, it has rich collection of search features which enables users to conduct document, annotations, relations, event structures, and plain text searches. [9] GATE is a multi-role web-based annotation environment, and it facilitates unique text annotation processes and offers methodological and tool assistance to the many hierarchies of team roles. It fully supports every stage of an annotation project's lifecycle. It mainly focused on collaborative annotation because of the time taken to text data annotation and user experience. However, collaborative annotation is not focused on this paper. Additionally, it has feature to auto annotate previously annotated words when user provides a partially annotated document. [10] YEDDA offers a complete range of features for text annotation as well as administrator assessment and analysis. It provides a graphical user interface and command line interface as well as powerful shortcut keys that can be configured with custom labels to annotate entities, it improves the ineffectiveness of conventional text annotation with these features. By providing pre-annotated text, YEDDA also can provide recommendations by using by Maximum Matching algorithm which is a text segmentation algorithm. [11] ANEA tool uses Wiktionary to make a connection between words in a given text. Wiktionary is an online dictionary it contains details and explanations about the word. Therefore, ANEA recognises entity groupings automatically by identifying Wiktionary descriptions and performing a search of relevant terms. It performs a double optimization task by calculating Cross-similarity and maximizations of words within an entity group and their average similarity to determine candidate labels. The drawback of this approach is if Wiktionary does not contain the required word this tool will not be able to guess the word entity. The lack of Sinhala words in Wiktionary affects negatively to Sinhala text data annotation using this approach.

Since all the previously mentioned text annotation tools are not providing name entity suggestions by considering variations of the base words, SIENA provides name entity suggestions by considering variations of the base words. SIENA tool is specially designed for Sinhala English code-switching text annotation tasks. It can identify variations of Sinhala and English words. By using the SIENA tool, text annotating persons can do their task effectively on Sinhala English code-switching corpus when they require a domain adaptation of named entity recognition.

Table 1.2. 1. Comparison with existing research related to text annotation tools

Tool name	**Collaboration features**	**Name Entity recommendations**	**Sinhala word variation identification**
ANEA	no	yes *(via external knowledge source)*	no
BRAT	no	yes *(via semantic class disambiguation)*	no
GATE	yes	no	no
YEDDA	no	yes *(via maximum matching algorithm)*	no
SIENA *(This research component)*	no	yes	yes

2. RESEARCH PROBLEM

With the advancement of technology, machine learning models are becoming more complex and denser, because of that the amount of data they needed to train is increasing, therefore a large amount of text data is required to prepare for the model training. As an example, spacy name entity recognition AI model is trained by 741 MBs of text data.[12] NCBI-Disease, BC5CDR and DFKI are popular datasets that used to train name entity recognition models which has large number of word count. Training chatbot components also required a large corpus because of that annotating a such large data set is a very tedious task. As well as when the dataset or the corpus contains enormous number of words, it takes a considerable amount of time to do the tagging process because every word in the corpus should be read and identified by the person who is doing the annotation task. Therefore, annotating data sets for named entity recognition models is a very time-consuming task. Name entity tagging requires domain knowledge as well because the person who does the annotation needs to read the whole text and understand the meaning, therefore it requires expert knowledge regarding the domain. Finding people with domain knowledge is also a challenging task. Also, there is a chance they are not able to do the text annotation due to their lack of technical knowledge.

In this research, only Sri Lankan people are focused, therefore most Sri Lankans are using the Sinhala-English code-switching language style when they interact with a chatbot. This statement is proven by the survey. [13] Sinhalese language, also known as Sinhala (සිංහල) is one of the two official languages of Sri Lanka, with about 16 million speakers out of the total population of 21 million also Sinhala is not a worldwide spread language like English. Due to those reasons, there is a lack of NLP tools and Sinhala specified name entity tagging tools designed for the Sinhala language. Therefore, Sinhala English mixed data annotation is even time consuming due to the absence of Sinhala English code-mixed annotation tools and technologies.

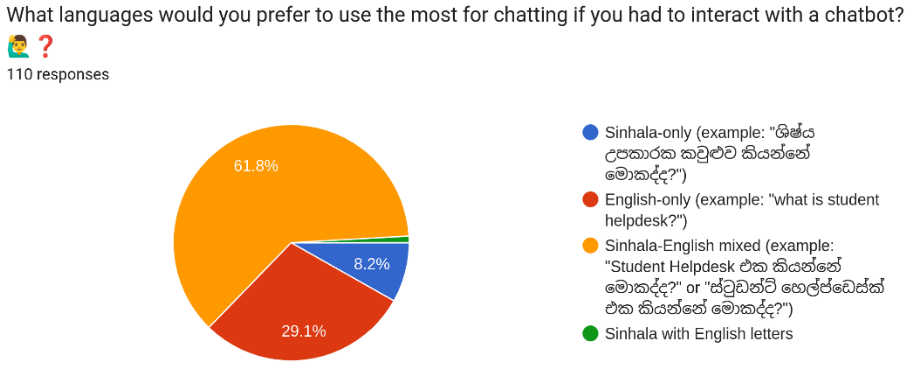


Figure 2. 1. Responses for What languages would you prefer to use the most for chatting if you had to interact with a chatbot?

3. RESEARCH OBJECTIVES

3.1 Main Objectives

The main objective of implementing the SIENA tool is to increase the efficiency of the text annotation process for the Sinhala-English code-switching corpus by providing accurate name entity recommendations and auto annotation.

3.2 Specific Objectives

To archive the main objective, the specific objectives that need to be fulfilled are identified as follows

1. Implement name entity suggestion algorithm

SIENA shows the name entity suggestion list according to the descending order of similarity score when the user selects a phrase.

2. Implement auto annotation algorithm

An algorithm to annotate all variations of phrases at once by selecting a base word of a phrase.

3. Make SIENA compatible with frameworks

After annotation is done. The tool should be able to export data with the compatibility of famous libraries which are used to build NER models. For this research component Rasa NLU data format was selected.

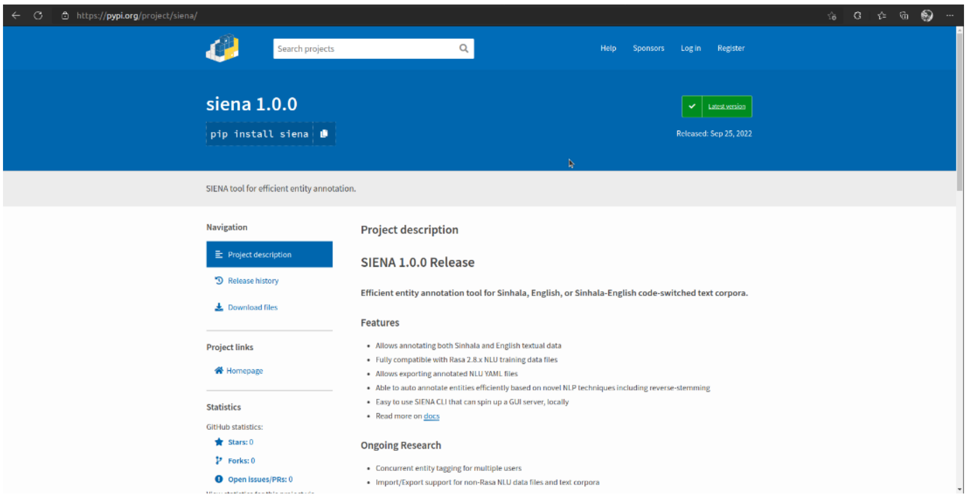
4. Develop visualizations technique to provide user friendly suggestions

Name entity suggestions will appear as a sorted list according to the similarity score below to the user selected phrases. Therefore, the user can easily select the most appropriate name entity.

5. Implement knowledge base as a portable component

Users can import or export knowledge base from or into SIENA. Therefore, users can share the knowledgebase which includes data for name entity suggestion and auto annotation algorithms among other instances of SIENA.

4. METHODOLOGY



siena 1.0.0

```
pip install siena
```

Released: Sep 25, 2022

SIENA tool for efficient entity annotation.

Project description

SIENA 1.0.0 Release

Efficient entity annotation tool for Sinhala, English, or Sinhala-English code-switched text corpora.

Features

- Allows annotating both Sinhala and English textual data
- Fully compatible with Rasa 2.8.x NLU training data files
- Allows exporting annotated NLU YAML files
- Able to auto annotate entities efficiently based on novel NLP techniques including reverse-stemming
- Easy to use SIENA CLI that can spin up a GUI server, locally
- Read more on docs

Ongoing Research

- Concurrent entity tagging for multiple users
- Import/Export support for non-Rasa NLU data files and text corpora

GitHub statistics:

Stars: 0

Forks: 0

Open issues/PRs: 0

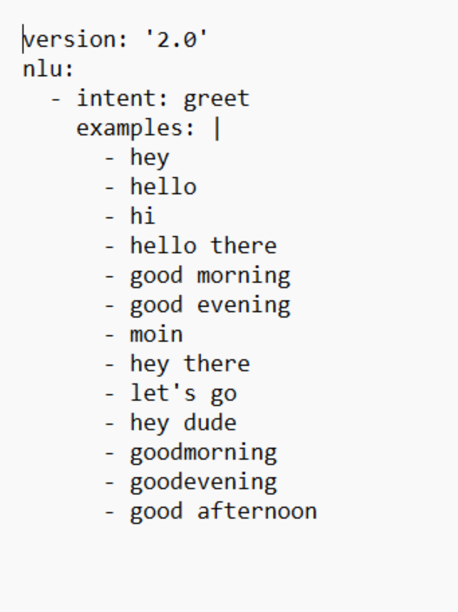
Figure 4. 1. SIENA python package

The requirements-gathering phase of SIENA tool done by studying existing research on name entity annotation and related products. All the functional and non-functional requirements are covered by implemented SIENA tool as described in above and later sections of this paper and mentioned in Table 4. 1. SIENA[1] is a text annotation tool specially designed for Sinhala – English text data. This tool helps the user to annotate text more effectively by providing suggestions and auto annotation. SIENA can install as a python package, and it supports rasa NLU data format for the annotation.

Table 4. 1. Functional and non-functional requirements

Functional requirements	Non functional requirements
Recommend and auto annotate name entities	Maintainability
Import / Upload corpus into SINEA	Security
Export annotated text from SIENA	Reliability
Import / Export knowledge base into SINEA	Portability

[1] https://pypi.org/project/siena



```
version: '2.0'
nlu:
  - intent: greet
    examples: |
      - hey
      - hello
      - hi
      - hello there
      - good morning
      - good evening
      - moin
      - hey there
      - let's go
      - hey dude
      - goodmorning
      - goodevening
      - good afternoon
```

Figure 4. 2. Sample NLU file format

SIENA tool can start with the command "siena server" inside the RASA bot folder. SIENA will search all the NLU files inside the "data" folder and shows those files inside the SIENA tool's files section area.

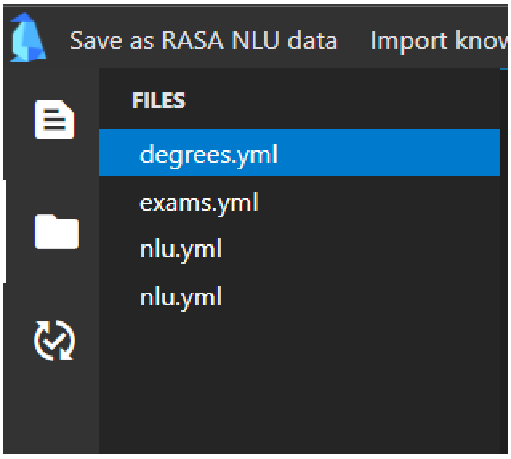
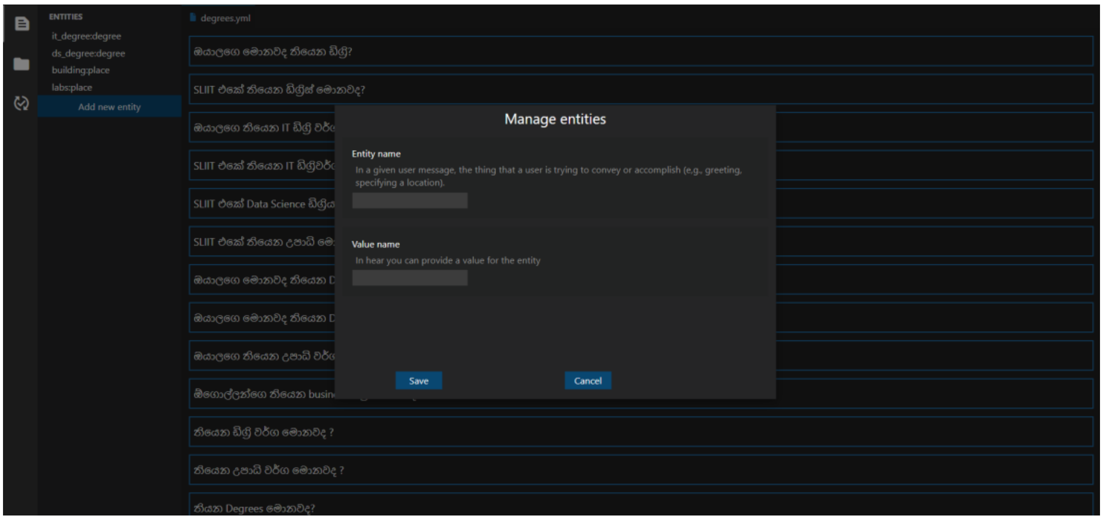


Figure 4. 3. NLU file list

As well as user can define the name entities using the tool and entered name entities will appear in the “entities” section.



ඔයාලගෙ මොනවද තියෙන ඩිග්‍රි?

SLIIT එකේ තියෙන ඩිග්‍රීස් මොනවද?

ඔයාලගෙ තියෙන IT ඩිග්‍රි වර්

SLIIT එකේ තියෙන IT ඩිග්‍රිවර්

SLIIT එකේ Data Science ඩිග්‍රිය

SLIIT එකේ තියෙන උපාධි මෙ

ඔයාලගෙ මොනවද තියෙන D

ඔයාලගෙ මොනවද තියෙන D

ඔයාලගෙ තියෙන උපාධි වර්

ඔගොල්ලන්ගෙ තියෙන busin

තියෙන ඩිග්‍රි වර්ග මොනවද ?

තියෙන උපාධි වර්ග මොනවද ?

තියන Degrees මොනවද?

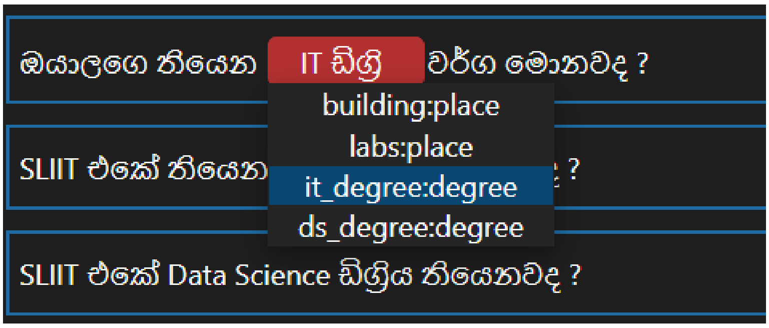
Manage entities

Entity name
In a given user message, the thing that a user is trying to convey or accomplish (e.g., greeting, specifying a location).

Value name
In hear you can provide a value for the entity

Figure 4. 4. Insert name entities to SIENA

Once the user annotates a named entity, SIENA keeps the annotated name entity and the base form of the phrase inside the knowledge base. Deriving base words is done by the stemming algorithm described in section 4.2. When the user selects a phrase for the annotation, SIENA converts it to its base form and calculates similarity scores against all the base word entries in the knowledge base. To calculate similarity between base forms SIENA uses a text similarity algorithm as described in section 4.1. Then SIENA will sort name entities according to descending order of calculated score and show the sorted list of name entities to the user.



ඔයාලගෙ තියෙන IT ඩිග්‍රි වර්ග මොනවද ?

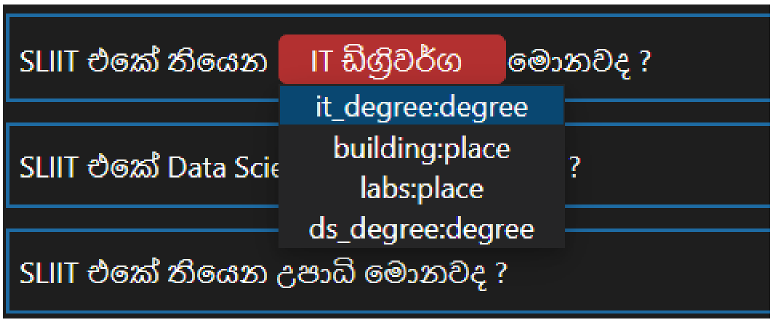
SLIIT එකේ තියෙන ? ?

SLIIT එකේ Data Science ඩිග්‍රිය තියෙනවද ?

Figure 4. 5. Initial name entity suggestion list order

As an example, Figure 4. 5 shows the initial state of the entity list before annotating “IT ඩිග්‍රි” word. After user annotate it as “it_degree:degree” and select “IT ඩිග්‍රිවර්ග”,

SIENA rearrange the entity list according to the similarity results of the knowledge base entries as shown in Figure 4. 6.



SLIIT එකේ තියෙන IT ඩිග්‍රිවර්ග මොනවද ?

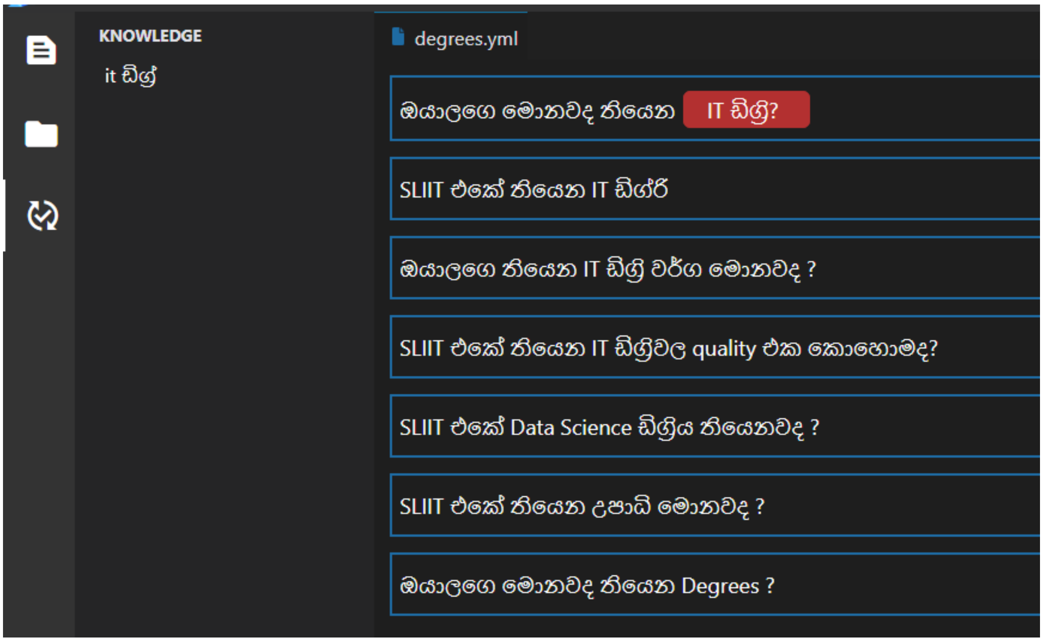
it_degree:degree
building:place
labs:place
ds_degree:degree

SLIIT එකේ Data Scie ?

SLIIT එකේ තියෙන උපාධි මොනවද ?

Figure 4. 6. Sorted name entity suggestion list

SIENA support auto annotation through the reverse stemming approach which is described in section 4.3. this allows user to assign same name entity to all the variations of the base word by clicking the base word in the "Knowledge" section of SIENA tool. SIENA shows a mapped name entity to each base word as a tooltip to improve user experience.



ඔයාලගෙ මොනවද තියෙන IT ඩිග්‍රි?

SLIIT එකේ තියෙන IT ඩිග්රී

ඔයාලගෙ තියෙන IT ඩිග්‍රි වර්ග මොනවද ?

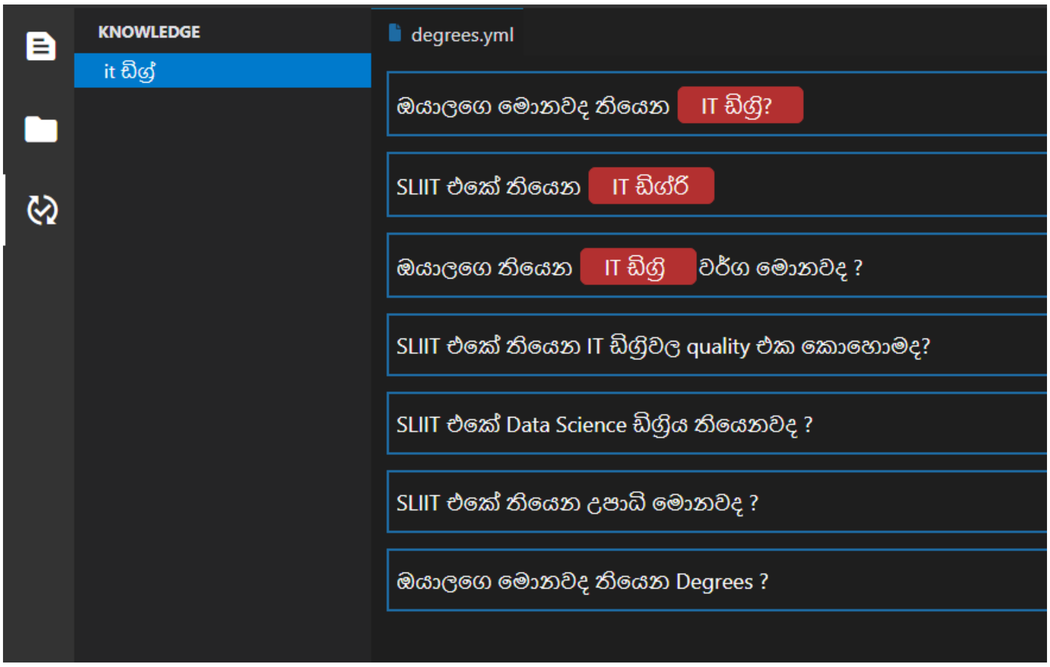
SLIIT එකේ තියෙන IT ඩිග්‍රිවල quality එක කොහොමද?

SLIIT එකේ Data Science ඩිග්‍රිය තියෙනවද ?

SLIIT එකේ තියෙන උපාධි මොනවද ?

ඔයාලගෙ මොනවද තියෙන Degrees ?

Figure 4. 7. Before auto annotation



ඔයාලගෙ මොනවද තියෙන IT ඩිග්‍රී?

SLIIT එකේ තියෙන IT ඩිග්රි

ඔයාලගෙ තියෙන IT ඩිග්‍රී වර්ග මොනවද ?

SLIIT එකේ තියෙන IT ඩිග්‍රීවල quality එක කොහොමද?

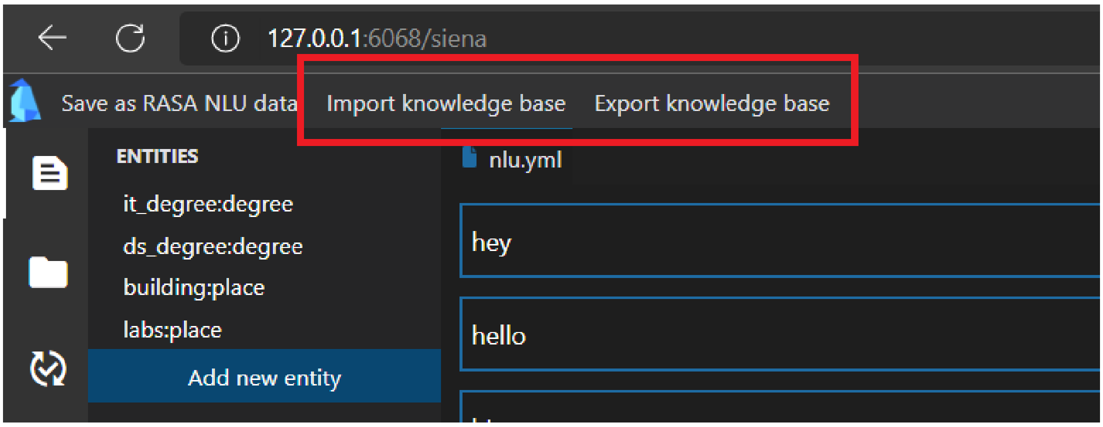
SLIIT එකේ Data Science ඩිග්‍රිය තියෙනවද ?

SLIIT එකේ තියෙන උපාධි මොනවද ?

ඔයාලගෙ මොනවද තියෙන Degrees ?

Figure 4. 8. After auto annotation

As well as SIENA can share gathered knowledge among other SIENA instances through the portable knowledge base component.



it_degree:degree

ds_degree:degree

building:place

labs:place

hey

hello

Figure 4. 9. Import or export knowledge base

4.1 Text Similarity Algorithm

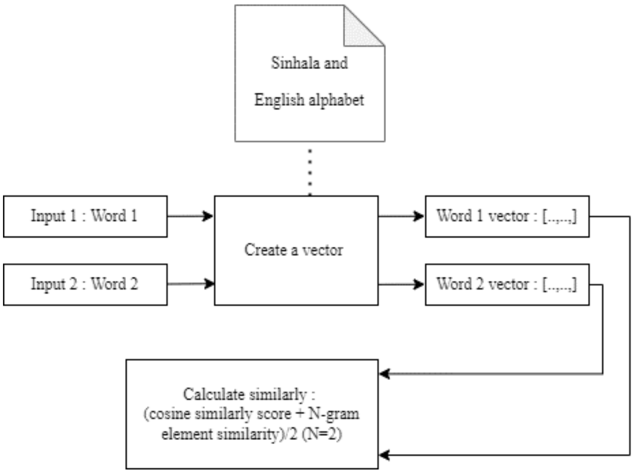

Sinhala and English alphabet
Input 1 : Word 1
Input 2 : Word 2
Create a vector
Word 1 vector : [..,..,]
Word 2 vector : [..,..,]
Calculate similarly : (cosine similarly score + N-gram element similarity)/2 (N=2)

Figure 4.1. 1. Text similarity algorithm

This algorithm is used to measure the similarity between two given phrases. Since this research only focused on Sinhala and English, this similarly algorithm will calculate similarly only for Sinhala and English data. Out of scope languages are ignored without throwing an error. As the first step two given phrases are converted into two vectors. To vectorise the phrase, this algorithm counts all the English and Sinhala letters in given phrases against each letter of the Sinhala and English alphabets. The Sinhala alphabet contains 1001 letters, and the English alphabet contains 26 letters. Therefore, this will produce a 1027-letter long vector for each phrase. To calculate the similarity between two vectors, this algorithm takes cosine similarly. Then it calculates bigram letter wise similarity and provides an average value of bigram similarity and the cosine similarity as the final similarity score. To calculate bigram similarity, it breaks down phrases into a sequence of neighbouring pairs of letters. As an example, word “ABCD” will break into “AB”, “BC”, and “CD”. Then it calculates the proportion of the matching elements.

4.2 Stemming Algorithm

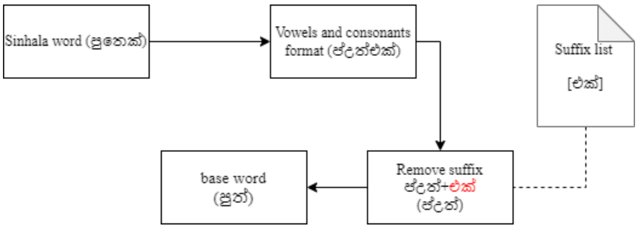
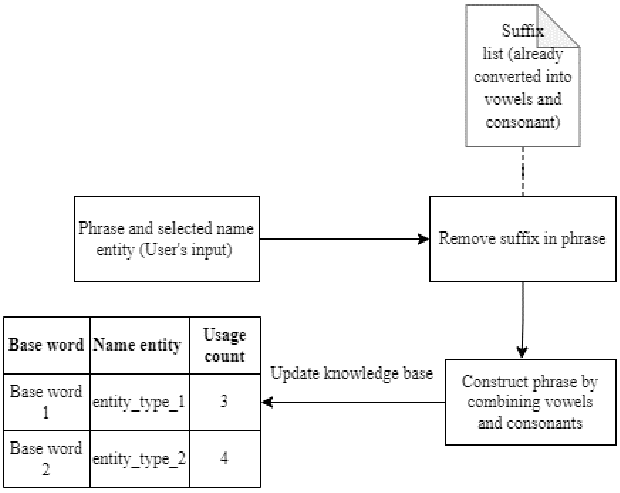


Figure 4.2. 1. Stemming algorithm

This algorithm is used to derive the base form of a given word without considering its morphological information.[14] In the Sinhala language, it is possible to construct words by joining prefixes, root words, and suffixes in a meaningful manner according to the නාම වරනැගිල්ල: Sinhala noun declension rules. A combination of prefixes and base words can have multiple suffixes. But prefixes significantly change the meaning of the word, unlike suffixes. For example, the Sinhala word පොත (book) can have multiple representations in terms of noun declension, such as පොත් (The books), පොතක් (a book), පොතට (To the book), පොතකට (To a book), පොතෙන් (From the book), and පොතකින් (From a book) which have the same base form පොත් combined with various suffixes. These combinations have almost the same meaning from the perspective of entity annotation. Therefore, the next section of this paper will propose a method for annotating entities by assigning entity types to the base form of a specified word. To remove suffixes, Sinhala words are converted into vowels and consonant formats. It makes it easy to separate suffixes from a given word because of the Sinhala word conjugation rules. The converted text was then cross checked with a Sinhala suffixes list which is also in vowels and consonants format. Then suffix is removed after the identification. Then remaining word is converted into normal form by combining vowels and consonants. As an example, word "ඇතා" (Tusker) need to be converted into "ඇත්" therefore, this algorithm will convert into its vowels and consonant format "ඇත්ආ" then tool looks longest suffix that can be removed from given phrase by checking suffix list which also converted into vowels and consonant format. In this case, the selected suffix will be "ආ" and its vowels and consonant format also "ආ". Then word "ඇත්" is remains. Then "ඇත්" need to be converted to its normal form. In this case vowels and consonant format of the word “ඇත්” is same as its normal form.

4.3 Reverse Stemming Algorithm


Suffix list (already converted into vowels and consonant)
Phrase and selected name entity (User's input)
Remove suffix in phrase
Construct phrase by combining vowels and consonants
Update knowledge base

Base word	Name entity	Usage count
Base word 1	entity_type_1	3
Base word 2	entity_type_2	4

Figure 4.3. 1. Reverse stemming algorithm – inserting to knowledge base

This algorithm is used to perform auto-annotation tasks based on previously annotated data. Unlike stemming algorithms, this is used to create a connection between base words and their original phrases. Therefore, when a user initially assigns a name entity to a phrase, this approach converts that phrase into base form and saves it inside the knowledge base with the name entity that was provided by the user. SIENA also saves the usage count for each entry in the knowledge base. When user selects a phrase SIENA converts it to its base form and searches for it in the knowledge base. If it finds a base word inside the knowledge base, the algorithm will assign it to all the variations of that base word. To find the base word variations, SIENA counts the number of words inside the user selected phrase and creates groups of words by preserving the sequence of sentences also each group consists of same count of words. Then SIENA iterate over all groups and converts to base form and checks if it matches with the user selected base form. If it matches SIENA will assign selected base word’s name entity to that group and reconstructs the sentence from groups of words.

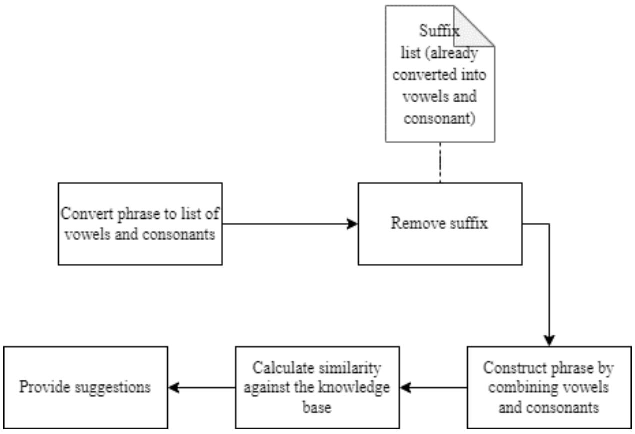

Suffix list (already converted into vowels and consonant)
Convert phrase to list of vowels and consonants
Remove suffix
Construct phrase by combining vowels and consonants
Calculate similarity against the knowledge base
Provide suggestions

Figure 4.3. 2. Reverse stemming algorithm - Providing suggestions

4.4 Preparation of Datasets

4.4.1 General dataset for algorithm evaluation

[15] To create the general dataset for algorithm evaluation and NLP text pre-processing tasks, used a publicly available Sinhala English newspaper dataset. It consists of Sinhala and English word mixed data. To do the user testing portion of 4.4.2 domain specific dataset is used.

4.4.2 Domain specific dataset for conversational AI training

The second domain specific dataset for text annotation with the SIENA tool is made using manually created Sinhala-English code-switched text. This dataset should be appropriately prepared as it will be used as a training dataset for conversational AI's intent classification challenge There will be roughly seventy-eight intents (78 classes), seven hundred examples (1700 data points) and each class have minimum of ten examples in each class as a standard. It is important to note that enhancing the overall effectiveness of conversational AI is dependent on the data set. Therefore, chatbot developers can change the number of question examples per intent at any time to improve conversational AI's overall performance and the domain adaptation.

4.5 Individual Component Architecture

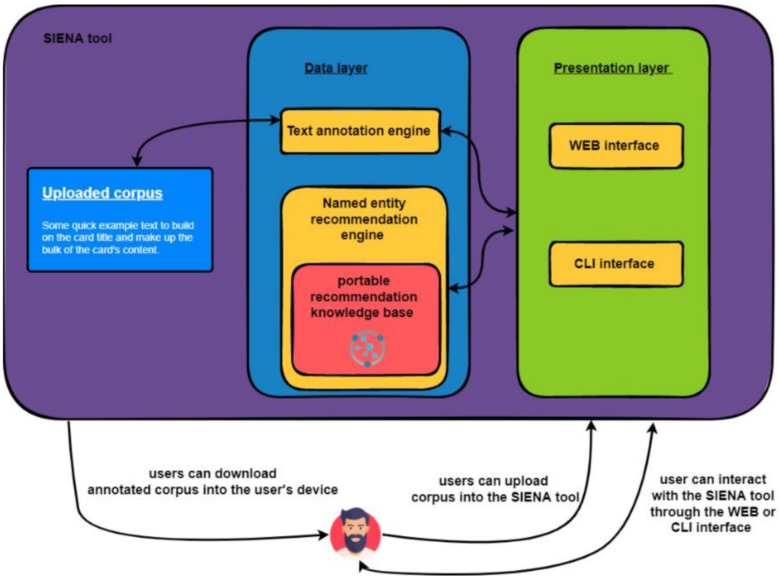

SIENA tool
Data layer
Presentation layer
Text annotation engine
WEB interface
Uploaded corpus
Some quick example text to build on the card title and make up the bulk of the card's content.
Named entity recommendation engine
CLI interface
portable recommendation knowledge base
users can download annotated corpus into the user's device
users can upload corpus into the SIENA tool
user can interact with the SIENA tool through the WEB or CLI interface

Figure 4.5. 1. Individual Component Architecture

SIENA tool consists of two subcomponents data layer and presentation layer. Data layer is responsible for auto annotation, name entity suggestions and other visualisations, presentation layer is responsible for provide interface to user interaction. SINEA tool will be used to annotate corpus which is used as training data by custom name entity recognition model in RASA NLU pipeline as shown in Figure 4.5. 2 and Figure 4.5. 3.

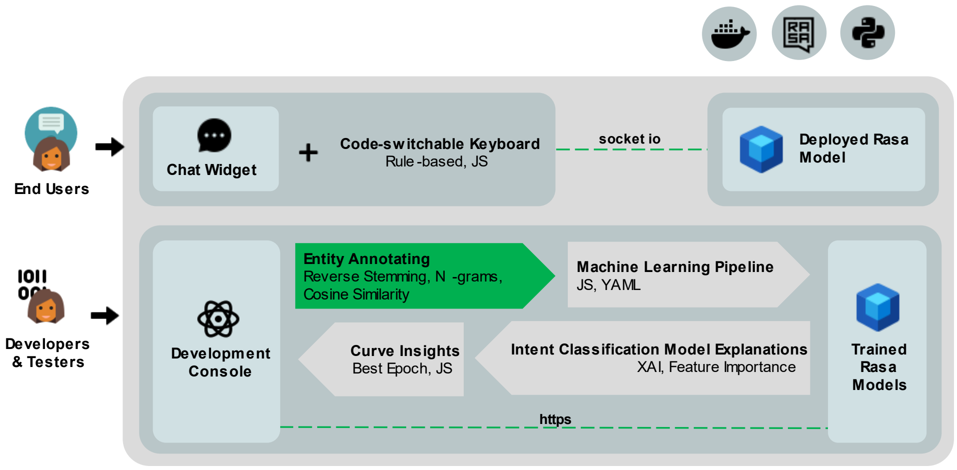


Figure 4.5. 2. Overall system flow

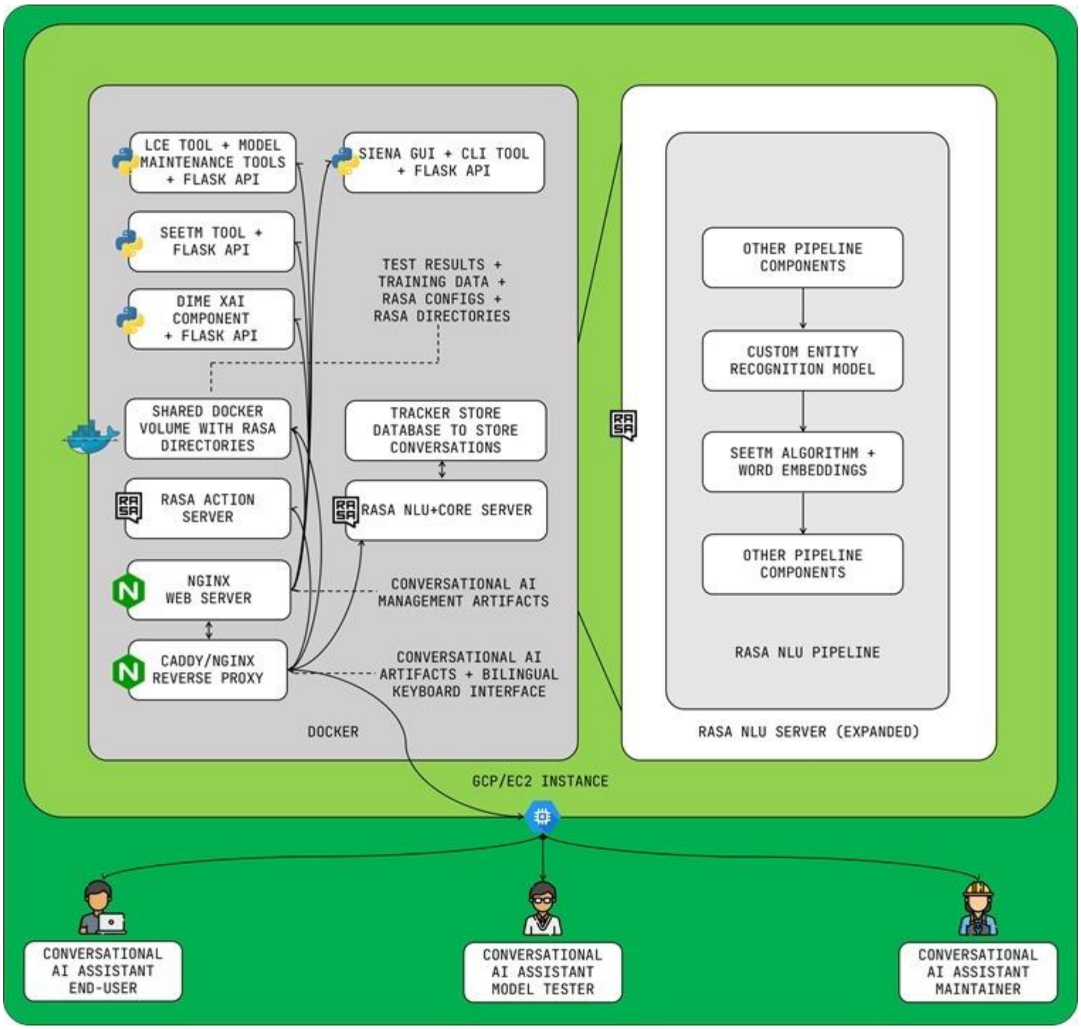

LCE TOOL + MODEL MAINTENANCE TOOLS + FLASK API
SIENA GUI + CLI TOOL + FLASK API
SEETM TOOL + FLASK API
TEST RESULTS + TRAINING DATA + RASA CONFIGS + RASA DIRECTORIES
DIME XAI COMPONENT + FLASK API
SHARED DOCKER VOLUME WITH RASA DIRECTORIES
TRACKER STORE DATABASE TO STORE CONVERSATIONS
RASA ACTION SERVER
RASA NLU+CORE SERVER
NGINX WEB SERVER
CONVERSATIONAL AI MANAGEMENT ARTIFACTS
CADDY/NGINX REVERSE PROXY
CONVERSATIONAL AI ARTIFACTS + BILINGUAL KEYBOARD INTERFACE
DOCKER
OTHER PIPELINE COMPONENTS
CUSTOM ENTITY RECOGNITION MODEL
SEETM ALGORITHM + WORD EMBEDDINGS
OTHER PIPELINE COMPONENTS
RASA NLU PIPELINE
RASA NLU SERVER (EXPANDED)
GCP/EC2 INSTANCE
CONVERSATIONAL AI ASSISTANT END-USER
CONVERSATIONAL AI ASSISTANT MODEL TESTER
CONVERSATIONAL AI ASSISTANT MAINTAINER

Figure 4.5. 3. Overall system architecture

4.6 Tools and Technologies

SIENA tool was developed using Python 3.8 with the help of Pandas, and NumPy packages. SIENA will be deployed as a flask app inside a docker container. With the help of Microsoft Azure as the cloud hosting service. As well as SIENA was released as a python package available in “pip”. SIENA version controlling is done by GitLab and GitHub. The integrated product version is controlled by GitLab, and the pip package’s version control is done by GitHub.

Table 4.6. 1. Tools and technologies

Task	Tools to be used
Algorithm Implementation and SIENA tool development	Python 3.8, NumPy, Pandas, PyCharm, Visual Studio Code, Google CoLab
Server development as a standalone frontend for the SIENA UI	Flask, JavaScript, CSS
Cloud Infrastructure management	Microsoft Azure
Conversational AI development by Integrating all research components (Backend)	Rasa 2.8.8, Gensim, spaCy, Docker
Conversational AI frontend development	React JS, Bootstrap, socket.io, JavaScript
Overall system deployment	Caddy 2, NGINX, Git, Docker, docker-compose

4.7 Commercialization Aspect of the Product

SIENA is a subcomponent of Kolloqe. Kolloqe is a chatbot development tool which aims to code less chatbot development and maintenance service and it will provide the service as a CaaS. Kolloqe consists of four main components, training data annotation is done by SIENA tool. Chatbot model training configuration and model evaluation, portable Sinhala typing keyboard and token mapping, and provide chatbot model explanation. Since SIENA is text annotation tool kollowe uses SIENA to annotate Rasa NLU text files efficiently. Business can purchase CaaS to reduce their chatbot development cost and the chatbot maintenance cost. Kolloqe provides several subscriptions by applying different limitations. Therefore, subscription fee is the main income of kolloqe. As well as kolloqe provides one time purchase on premises solution.

5. TESTING & IMPLEMENTATION RESULTS & DISCUSSION

5.1 Results

The main objective of implementing SIENA is reduce the time taken to text annotation. To achieve the main objective SIENA consists of name entity suggestion feature (shown in Figure 4. 6), auto annotation feature (shown in Figure 4. 8) and portable knowledge base component (shown in Figure 4. 9) To test the effectiveness of SIENA tool, provided seventeen lines of text document which contained sentences related to SLIIT degrees. This document provided text dataset to four users and recorded time taken to annotate with SIENA tool and the manual annotation method. SIENA tool achieved 54.7% average effectiveness when compared to manual annotation as shown in the Table 5.1. 1.

Table 5.1. 1. SIENA user testing results

User	**Time taken to annotate using SIENA (seconds)**	**Time taken to manual annotating (seconds)**	**Time efficiency percentage (%)**
User 1	135	184	36.3
User 2	114	178	56.1
User 3	106	181	70.7
User 4	125	195	56.0
Average	120	184.5	54.7

5.1.1 Text similarity algorithm results

The following matrix shows how the combined similarity algorithm works with different words.

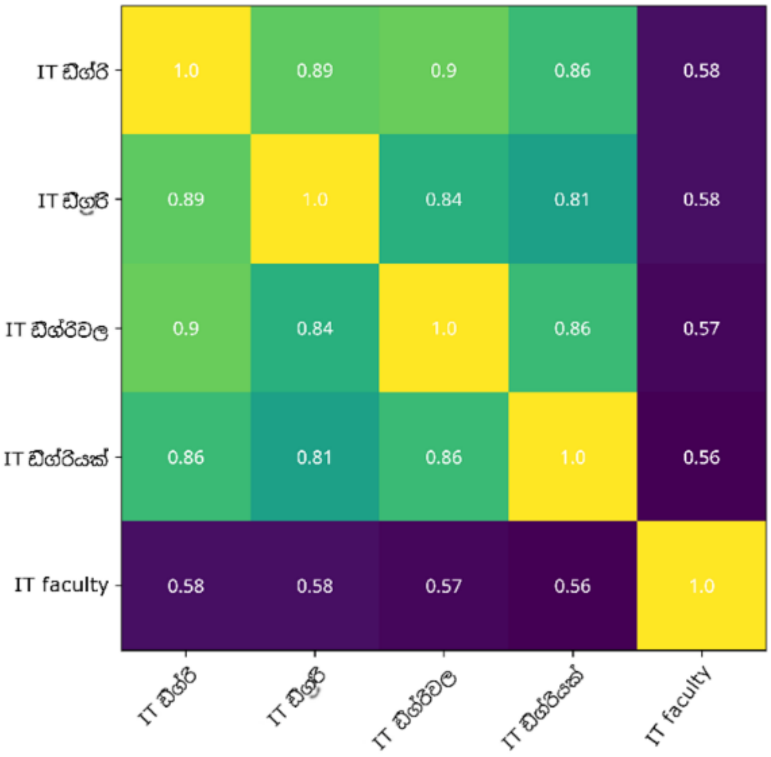


Figure 5.1.1. 1. Text similarity algorithm results

5.1.2 Stemming algorithm results

Since the stemming algorithm and reverse stemming approach shares same accuracy, results are mentioned in the section 5.1.3. The following table shows the stemming algorithm works with different words.

Table 5.1.2. 1. Stemming algorithm results

Word	SIENA base word
IT ඩිග්‍රියේ	it ඩිග්‍රිය්
IT ඩිග්‍රියක්	it ඩිග්‍රිය්
IT ඩිග්‍රියක	it ඩිග්‍රිය්
IT ඩිග්රි	it ඩිග්‍ර්

IT ඩිග්‍රි	it ඩිග්‍ර්
IT ඩිග්‍රිවල	it ඩිග්‍ර්
IT ඩිග්‍රිය	it ඩිග්‍ර්

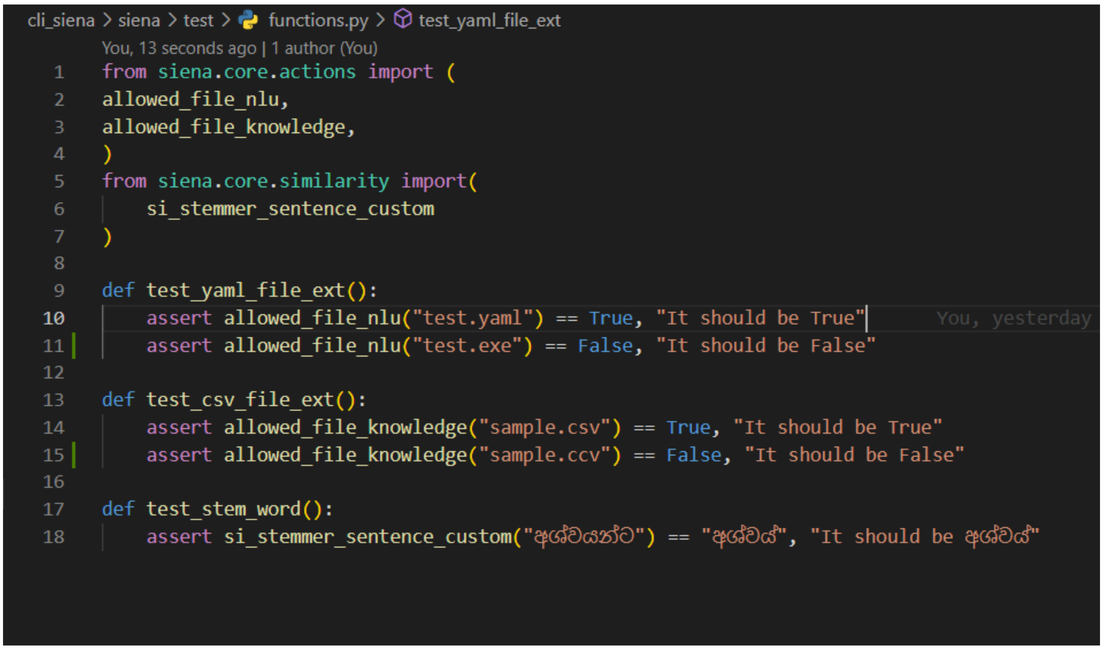
5.1.3 Reverse stemming results

Evaluation of the SIENA reverse stemming (base word mapping) technique performed using the conversational AI training dataset yielded an average accuracy of 57.1%. this is calculated by getting the percentage of number of correctly identified base words over the total number of variations of the base words for a given phase.

5.1.4 SIENA test results

SIENA tool was tested by performing unit testing and integration testing. Unit tests are done by the python inbuilt keyword “assert”. “test_yaml_file_ext” and “test_csv_file_ext” are file extension validators for YAML NUL and CSV files. “test_stem_word” is the function of the stemming algorithm.

$$Accuracy = \frac{Number\ of\ correctly\ identified\ base\ words}{Total\ number\ of\ base\ words\ for\ given\ phase} \times 100\% \qquad (1)$$


cli_siena > siena > test > functions.py > test_yaml_file_ext
You, 13 seconds ago | 1 author (You)
You, yesterday

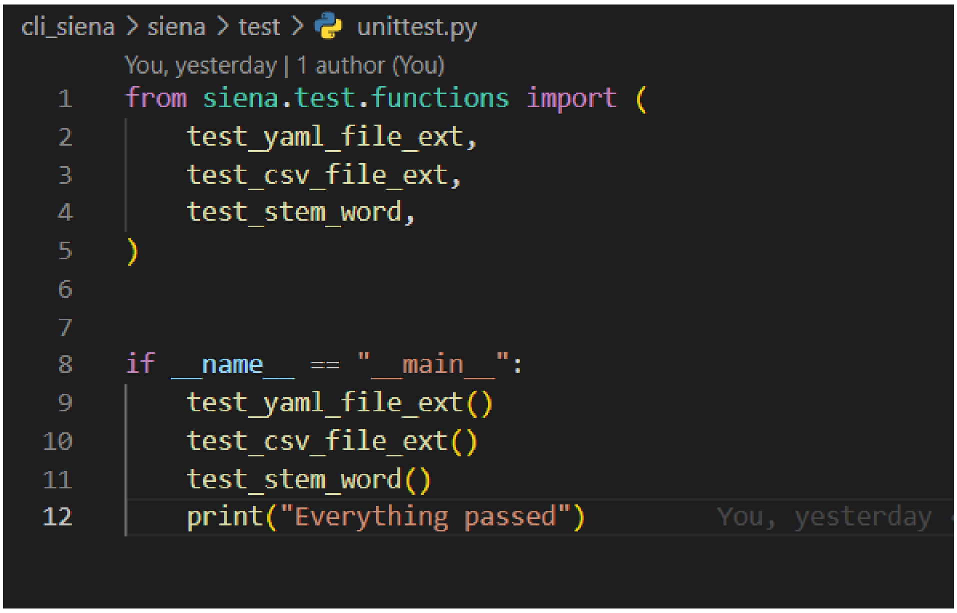
```python
from siena.core.actions import (
allowed_file_nlu,
allowed_file_knowledge,
)
from siena.core.similarity import(
    si_stemmer_sentence_custom
)

def test_yaml_file_ext():
    assert allowed_file_nlu("test.yaml") == True, "It should be True"
    assert allowed_file_nlu("test.exe") == False, "It should be False"

def test_csv_file_ext():
    assert allowed_file_knowledge("sample.csv") == True, "It should be True"
    assert allowed_file_knowledge("sample.ccv") == False, "It should be False"

def test_stem_word():
    assert si_stemmer_sentence_custom("අශ්වයන්ට") == "අශ්වය්", "It should be අශ්වය්"
```

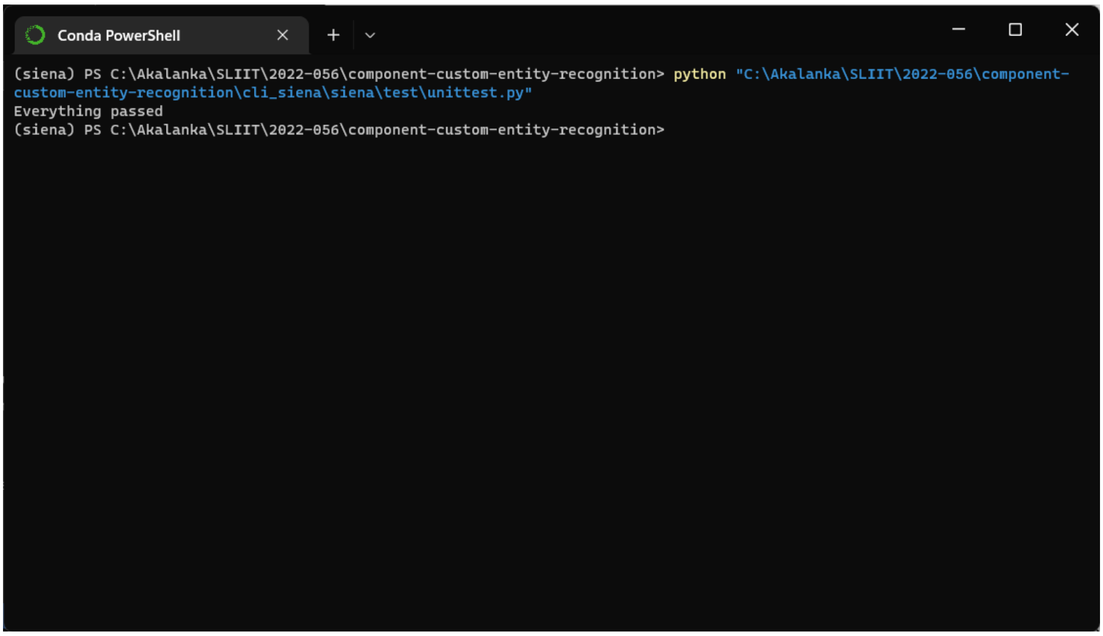
Figure 5.1.4. 1. Unit tests


cli_siena > siena > test > unittest.py
You, yesterday | 1 author (You)
You, yesterday

```python
from siena.test.functions import (
    test_yaml_file_ext,
    test_csv_file_ext,
    test_stem_word,
)


if __name__ == "__main__":
    test_yaml_file_ext()
    test_csv_file_ext()
    test_stem_word()
    print("Everything passed")
```

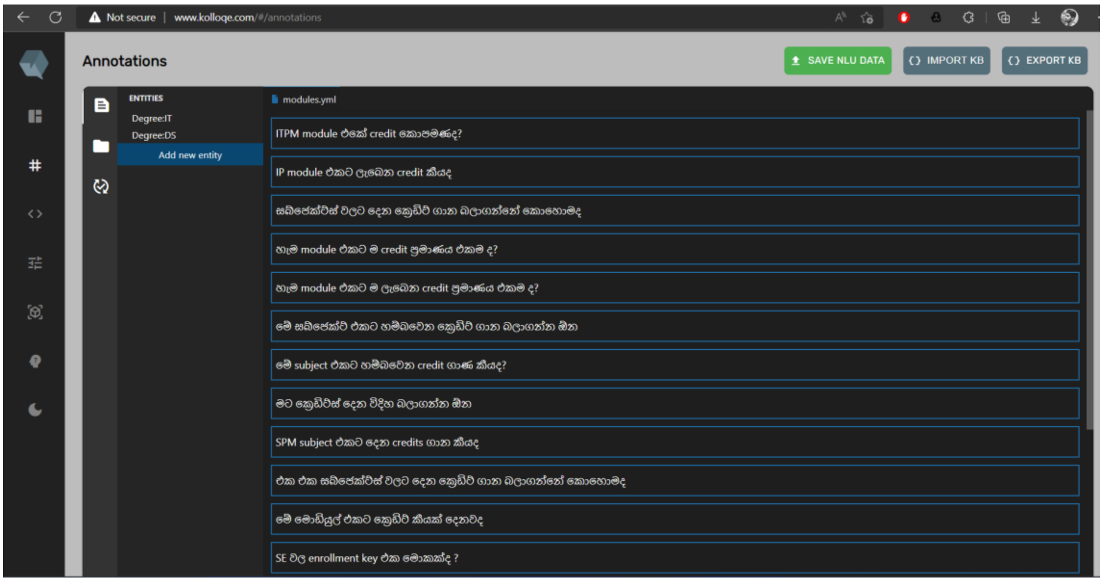
Figure 5.1.4. 2. Unit tests



```
(siena) PS C:\Akalanka\SLIIT\2022-056\component-custom-entity-recognition> python "C:\Akalanka\SLIIT\2022-056\component-custom-entity-recognition\cli_siena\siena\test\unittest.py"
Everything passed
(siena) PS C:\Akalanka\SLIIT\2022-056\component-custom-entity-recognition>
```

Figure 5.1.4. 3. Unit tests

Integration testing is done with the help of all the team members by testing each other's individual components. After the testing was completed. SIENA and other components are deployed into a Microsoft Azure virtual machine instance.



ITPM module එකේ credit කොපමණද?

IP module එකට ලැබෙන credit කීයද

සබ්ජෙක්ට්ස් වලට දෙන ක්‍රෙඩිට් ගාන බලාගන්නේ කොහොමද

හැම module එකට ම credit ප්‍රමාණය එකම ද?

හැම module එකට ම ලැබෙන credit ප්‍රමාණය එකම ද?

මේ සබ්ජෙක්ට් එකට හම්බවෙන ක්‍රෙඩිට් ගාන බලාගන්න ඕන

මේ subject එකට හම්බවෙන credit ගාණ කීයද?

මට ක්‍රෙඩිට්ස් දෙන විදිහ බලාගන්න ඕන

SPM subject එකට දෙන credits ගාන කීයද

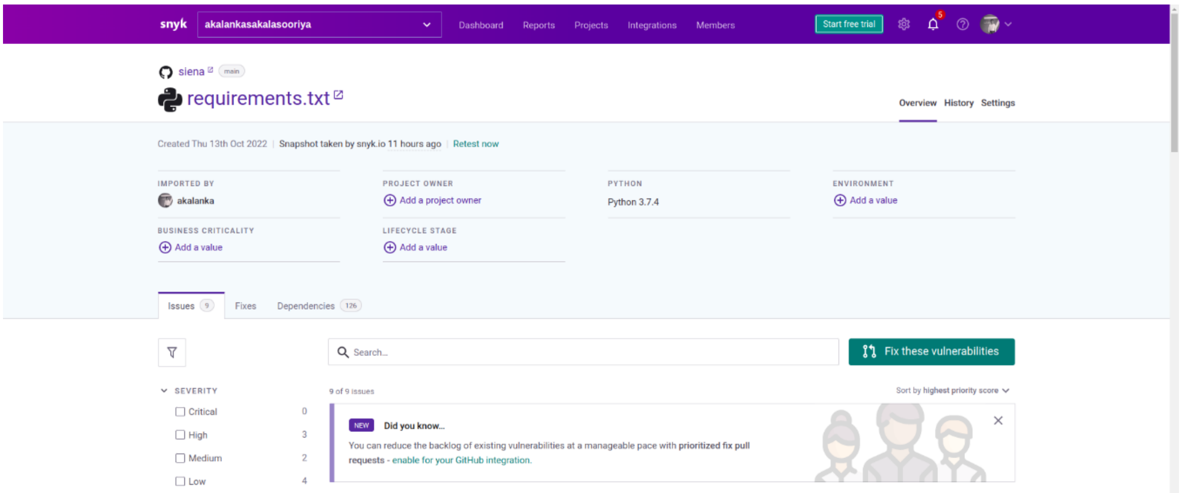
එක එක සබ්ජෙක්ට්ස් වලට දෙන ක්‍රෙඩිට් ගාන බලාගන්නේ කොහොමද

මේ මොඩියුල් එකට ක්‍රෙඩිට් කීයක් දෙනවද

SE වල enrollment key එක මොකක්ද ?

Figure 5.1.4. 4. SIENA integrated into kolloqe

To analyse security issues SIENA is continuously scanned by a popular vulnerability scanning tool “snyk”. It provides suggestions to improve security. As well as It notifies vulnerabilities in third party packages used in SIENA tool.



siena main

requirements.txt

Created Thu 13th Oct 2022 | Snapshot taken by snyk.io 11 hours ago | Retest now

IMPORTED BY akalanka

PROJECT OWNER Add a project owner

PYTHON Python 3.7.4

ENVIRONMENT Add a value

BUSINESS CRITICALITY Add a value

LIFECYCLE STAGE Add a value

SEVERITY

☐ Critical 0

☐ High 3

☐ Medium 2

☐ Low 4

9 of 9 issues

Sort by highest priority score

NEW Did you know...

You can reduce the backlog of existing vulnerabilities at a manageable pace with prioritized fix pull requests - enable for your GitHub integration.

Figure 5.1.4. 5. Results of snyk vulnerability analysis tool

SIENA tool’s main functionalities are test by using the following test cases.

Table 5.1.4. 1. Test case 001

<table>
<tr><td colspan="6"><b>Project ID:</b> 2022-056-IT19051208</td></tr>
<tr><td colspan="6"><b>Project Name:</b> SIENA</td></tr>
<tr><td colspan="6"><b>Project Function:</b> Name entity suggestions</td></tr>
<tr><td colspan="3"><b>Test case ID:</b> 001</td><td colspan="3"><b>Test case designed by:</b>
<b>ID No:</b> IT19051208
<b>Name:</b> Sakalasooriya S.A.H.A.</td></tr>
<tr><td colspan="6"><b>Test Priority (High/Medium/Low):</b> Medium</td></tr>
<tr><td colspan="6"><b>Test Description:</b> The name entity suggestion list should be rearranged once user annotate a similar phrase.</td></tr>
<tr><td colspan="6"><b>Prerequisite:</b> User should enter list of name entities to SIENA tool using the entity management window.</td></tr>
<tr><td colspan="6"><b>Test Steps:</b>
Step 1: Load a file to start the annotation
Step 2: Highlight first word
Step 3: Click on a name entity from the list (near to the highlighted area)
Step 4: Highlight similar word (used in step 2)
Step 5: Observer the order of name entity suggestion list.</td></tr>
<tr><td>Test ID</td><td>Test Inputs</td><td>Expected Outputs</td><td>Actual Output</td><td>Result (Pass/Fail)</td><td>Comments</td></tr>
</table>

001	**Input:** Selected “IT ඩිග්‍රී” as the first word and assigned “degree_type” as the name entity. Selected “IT ඩිග්‍රියේ” as the second word	Entity list rearrange by coming up “degree_type” as the first in the entity list	Entity list rearranged by coming up “degree_type” as the first in the entity list	pass	Entity suggestion function is working properly.

Table 5.1.4. 2. Test case 002

<table>
<tr><td colspan="6"><b>Project ID:</b> 2022-056-IT19051208</td></tr>
<tr><td colspan="6"><b>Project Name:</b> SIENA</td></tr>
<tr><td colspan="6"><b>Project Function:</b> Auto annotation</td></tr>
<tr><td colspan="3"><b>Test case ID:</b> 002</td><td colspan="3"><b>Test case designed by:</b>
<b>ID No:</b> IT19051208
<b>Name:</b> Sakalasooriya S.A.H.A.</td></tr>
<tr><td colspan="6"><b>Test Priority (High/Medium/Low):</b> Medium</td></tr>
<tr><td colspan="6"><b>Test Description:</b> Auto annotate all the variation of base word when user selects a base word from auto annotation menu.</td></tr>
<tr><td colspan="6"><b>Prerequisite:</b> User should annotate at least one name entity to appear base word in auto annotation windows.</td></tr>
<tr><td colspan="6"><b>Test Steps:</b>
Step 1: Load a file to start the annotation
Step 2: Click on base word list icon
Step 3: Click on base word
Step 4: Observer the auto annotation</td></tr>
<tr><td>Test ID</td><td>Test Inputs</td><td>Expected Outputs</td><td>Actual Output</td><td>Result (Pass/Fail)</td><td>Comments</td></tr>
<tr><td>002</td><td><b>Input:</b> Selected “it ඩිශ්‍රී” as the base word.</td><td>Auto annotate all the identified variations of the selected base word.</td><td>Auto annotated all the identified variations of the selected base word.</td><td>pass</td><td>Auto annotation function is working properly.
Auto annotation success notifications works properly</td></tr>
</table>

Table 5.1.4. 3. Test case 003

Project ID: 2022-056-IT19051208	
Project Name: SIENA	
Project Function: Exporting knowledge base	
Test case ID: 003	**Test case designed by:** **ID No:** IT19051208 **Name:** Sakalasooriya S.A.H.A.
Test Priority (High/Medium/Low): Medium	
Test Description: Export knowledge base as a CSV file.	
Prerequisite: User should annotate at least one name entity to contain details inside the exported knowledge base.	
Test Steps: Step 1: Click on export knowledge button Step 2: Save the downloaded file.	

Test ID	Test Inputs	Expected Outputs	Actual Output	Result (Pass/Fail)	Comments
003	**Input:** No inputs	Web browser ask to save the knowledge base file.	Web browser asked to save the knowledge base file.	pass	Knowledge base export function works properly. Knowledge base export notification works properly.

Table 5.1.4. 4. Test case 004

Project ID: 2022-056-IT19051208	
Project Name: SIENA	
Project Function: Auto annotation	
Test case ID: 004	**Test case designed by:** **ID No:** IT19051208 **Name:** Sakalasooriya S.A.H.A.
Test Priority (High/Medium/Low): Medium	

<table>
<tr><td colspan="6"><b>Test Description:</b> Auto annotate all the variation of base word when user selects a base word from auto annotation menu.</td></tr>
<tr><td colspan="6"><b>Prerequisite:</b> User should have up and running SIENA instance.</td></tr>
<tr><td colspan="6"><b>Test Steps:</b>
Step 1: Click on import knowledge base button
Step 2: Find previously exported knowledge base file
Step 3: Upload selected file</td></tr>
<tr><td>Test ID</td><td>Test Inputs</td><td>Expected Outputs</td><td>Actual Output</td><td>Result (Pass/Fail)</td><td>Comments</td></tr>
<tr><td>004</td><td><b>Input:</b> Navigate to previously exported knowledge base file and select it in file selection dialog box.</td><td>Update current knowledge base</td><td>Updated current knowledge base</td><td>pass</td><td>Knowledge base import function works properly.
Knowledge base import notification works properly.</td></tr>
</table>

5.2 Research Findings

Initially, SIENA used cosine similarity to calculate word similarity, since cosine similarity results are right-skewed and provide close scores to 100%. As well as cosine similarity algorithm does not care about the sequence of the letters when calculating similarity. As an example, cosine similarity provides 100% similarity for word “ABC” and word “CBA”. Therefore, text similarity measurement approach employed by the SIENA tool provides the average of cosine and bi-gram element similarity scores. It makes results nearly symmetrical distribution as shown in the Figure 5.2. 1. These similarity results are taken by cross-calculating randomly selected 1000 words from Sinhala newspaper corpus [15]. The bi-gram similarity algorithm is more sensitive to character variations when compared to the cosine similarity algorithm. As well as the bi-gram similarity algorithm adds a sequential feature to the combined algorithm because the cosine similarly algorithm does not care about positional information of letter in each phrase.

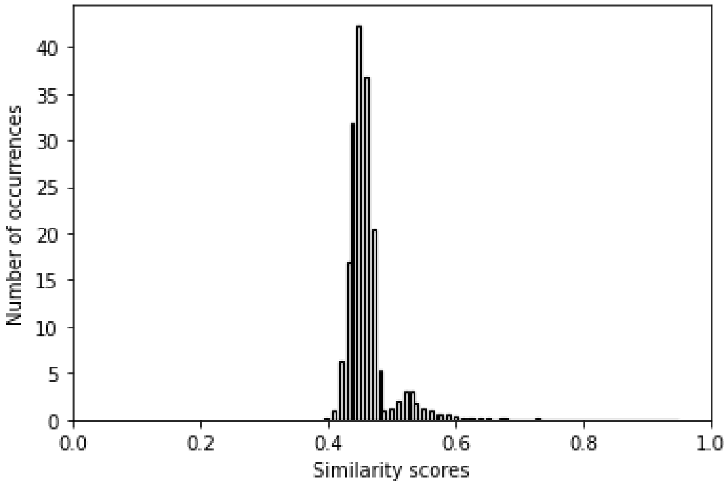


Figure 5.2. 1. Combined text similarly algorithm

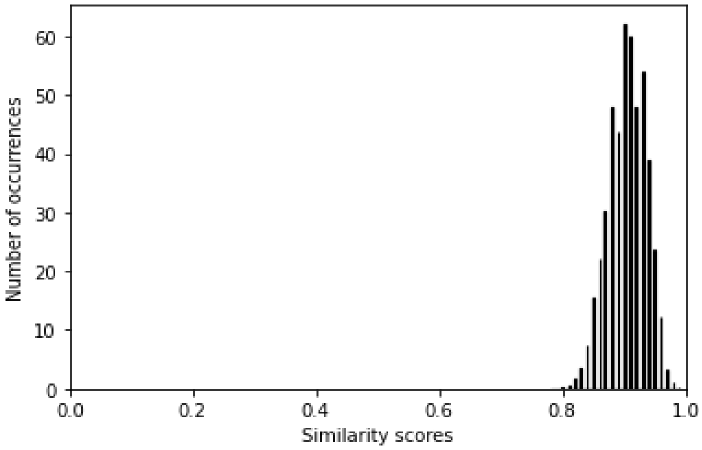


Figure 5.2. 2. Cosine text similarly algorithm

5.3 Discussion

Since SIENA could be able to find the variations of base wards, SIENA can auto annotate name entities when there is an already name entity assigned to the base word. But as mentioned in the results section the accuracy of the auto annotation is low due to the problems that occurred in the stemming algorithm. As shown in table 5.3.1 both "IT ඩිග්‍රියක" and "IT ඩිග්‍රි" words should be mapped into a single base form. But the current Sinhala stemming algorithm is not much advanced and depends on a rule-based suffix removal approach, in some cases stemming algorithm cannot cover some Sinhala word conjugation rules. Because of this auto annotation algorithm will not be able to find all the variations of base words. To overcome this situation stemming algorithm needs to be improved further. Currently SIENA stemming algorithm only supports English and Sinhala languages. This same approach can be used to stem Tamil word as well because Tamil letters are also construed with vowels and consonants like in Sinhala.

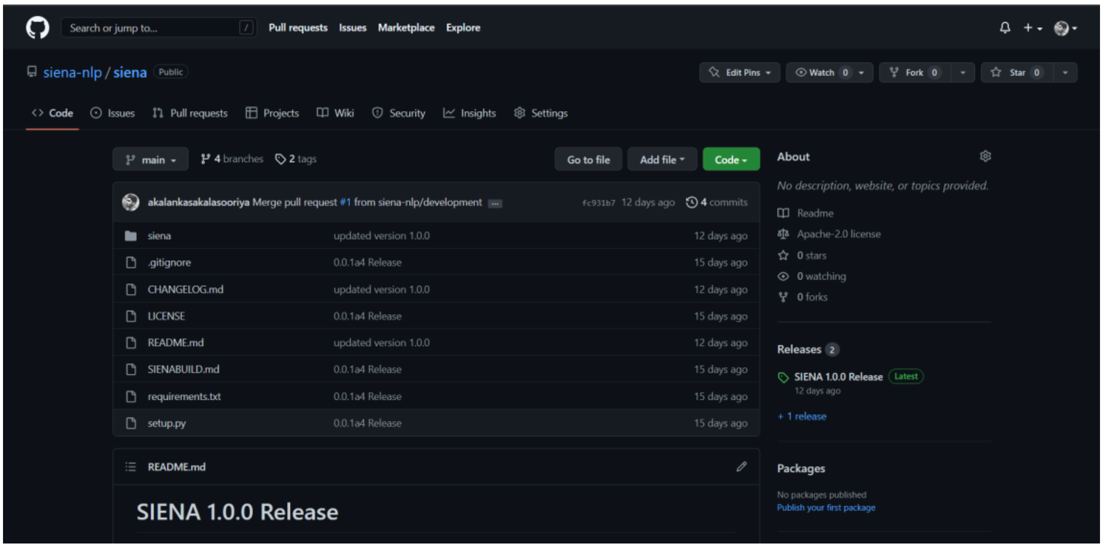
Table 5.3. 1. SIENA stemming issues

Word	SIENA base word
IT ඩිග්‍රියක	it ඩිග්‍රිය්
IT ඩිග්‍රි	it ඩිග්‍ර්

5.4 Contribution

SIENA is an open-source project available on GitHub[2] and anyone can read, distribute, and modify the source. SIENA is released as a python package which is available on PyPI. Therefore, SIENA can install on any platform that Python supports. Since SIENA is an open-source tool, the transparency of the tool is achieved. As well as developers can further improve the SIENA tool as a community driven project. This will motivate developers to implement or improve Sinhala language focused NLP tools further.

[2] https://github.com/orgs/siena-nlp/repositories



akalankasakalasooriya Merge pull request #1 from siena-nlp/development … fc931b7 12 days ago 4 commits

siena	updated version 1.0.0	12 days ago
.gitignore	0.0.1a4 Release	15 days ago
CHANGELOG.md	updated version 1.0.0	12 days ago
LICENSE	0.0.1a4 Release	15 days ago
README.md	updated version 1.0.0	12 days ago
SIENABUILD.md	0.0.1a4 Release	15 days ago
requirements.txt	0.0.1a4 Release	15 days ago
setup.py	0.0.1a4 Release	15 days ago

README.md

SIENA 1.0.0 Release

About

No description, website, or topics provided.

Readme
Apache-2.0 license
0 stars
0 watching
0 forks

Releases 2

SIENA 1.0.0 Release Latest
12 days ago

+ 1 release

Packages

No packages published
Publish your first package

Figure 5.4. 1. SIENA GitHub page

6. CONCLUSION

Text data annotation is time consuming task, and it requires human interaction and domain knowledge. There are several text annotation tools are available but none of them are not specially enhanced for Sinhala and English code-mixed data annotation. Unlike other text data annotation tools SIENA can identify variations of Sinhala base words and auto annotate all the variation at once. By using SIENA tool, users can effectively annotate Sinhala and English text data. Currently, SIENA only accepts Rasa NLU data format because SIENA is a part of a chatbot creation and evaluation tool kolloqe which is a Rasa oriented tool. Since SIENA is a python web-based application it can use any operating system which supports python. SIENA tool was tested on Microsoft Windows, Mac OS, and Linux operating systems. Also, SIENA has name entity suggestion and auto annotation feature which reduce the time taken to annotate. The knowledge base is another vital component which helps to name entity suggestions and auto annotation. Since the knowledge base component is a portable component, it can export its content to another instance of SIENA by exporting the knowledge base and SIENA can import external knowledge as well. Therefore, SIENA can share the gathered knowledge among other instances as needed. This feature is helpful to increase the efficiency when annotating same domain data because it can provide better suggestions and accurate auto annotation at the beginning of the annotation. Therefore, SIENA tool can be used to annotate Rasa NLU text files and it makes it easier to develop Rasa chatbots with the support of Sinhala English code-mixed conversations. Therefore, businesses tend to use chatbots with Sinhala English support as their business interface, this will enhance the most of Sri Lankan customer experience and business growth.

REFERENCES

[1]. ieee-dataport.org, ‘How to Cite References: IEEE Documentation Style,’ [Online]. Available: https://ieee-dataport.org/sites/default/files/analysis/27/IEEE%20Citation%20Guidelines.pdf [Accessed: 04-Oct-2022]

[2]. The State of AI and Machine Learning, [online], Available: https://resources.appen.com/wp-content/uploads/2020/06/Whitepaper-State-of-Ai-2020-Final.pdf [Accessed: 04-Oct-2022]

[3]. Siddhant Meshram, Namit Naik, Megha VR, Tanmay More, Shubhangi Kharche, Conversational AI: Chatbots, Available: [Online]. https://ieeexplore.ieee.org/document/9498508

[4]. Jason P.C. Chiu, Eric Nichols, ‘Named Entity Recognition with Bidirectional LSTM-CNNs’, [Online]. Available: tacl_a_00104.pdf (silverchair.com)

[5]. Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, Chris Dyer, ‘Neural Architectures for Named Entity Recognition’ [Online]. Available: https://arxiv.org/pdf/1603.01360.pdf

[6]. Geoffrey Leech, Lancaster University, Developing Linguistic Corpora: A Guide to Good Practice, Available: [Online]. https://users.ox.ac.uk/~martinw/dlc/chapter2.htm

[7]. viralzone.expasy.org, Human viruses table, Available: [Online]. https://viralzone.expasy.org/678

[8]. Pontus Stenetorp, Sampo Pyysalo, Goran Topic, Tomoko Ohta, Sophia Ananiadou, and Jun’ichi Tsujii, ‘BRAT: a Web-based Tool for NLP-Assisted Text Annotation’, [Online]. Available: https://aclanthology.org/E12-2021.pdf

[9]. Kalina Bontcheva, Hamish Cunningham, Ian Roberts, Angus Roberts, Valentin Tablan, Niraj Aswani, Genevieve Gorrell, ‘GATE Teamware: a web-based, collaborative text annotation framework,’ [Online]. Available: https://www.jstor.org/stable/42636386

[10]. Jie Yang, Yue Zhang, Linwei Li, Xingxuan Li, ‘YEDDA: A Lightweight Collaborative Text Span Annotation Tool,’ [Online]. Available: https://aclanthology.org/P18-4006.pdf

[11]. Anastasia Zhukova, Felix Hamborg, Bela Gipp, ‘ANEA: Automated (Named) Entity Annotation for German Domain-Specific Texts,’ [Online]. Available: https://arxiv.org/pdf/2112.06724.pdf

[12]. Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li, A Survey on Deep Learning for Named Entity Recognition, [Online]. Available: https://ieeexplore.ieee.org/document/9039685

[13]. G. Thilini Weerasuriya , Supunmali Ahangama, Maheshi Nandathilaka, A Rule-based Lemmatizing Approach for Sinhala Language, [Online]. Available: https://www.researchgate.net/publication/333769052_A_Rule-based_Lemmatizing_Approach_for_Sinhala_Language

[14]. J.B Dissanayake, Basaka mahima, ISBN: 9789556963656

[15]. Nisansa de Silva, SiClaEn dataset, [Online]. Available: https://osf.io/tdb84/

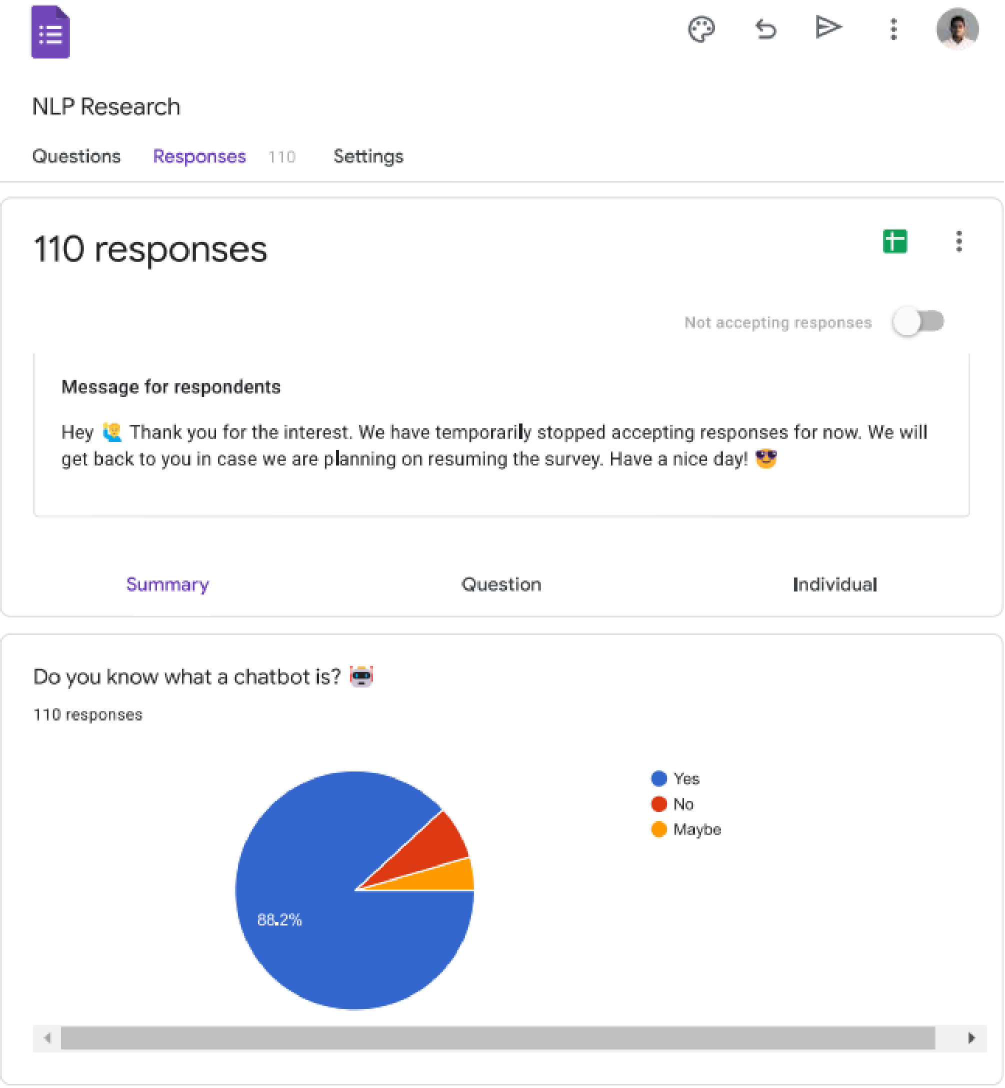
[16]. A. M. Turing, ‘computing machinery and intelligence,’ [Online]. Available:https://academic.oup.com/mind/article/LIX/236/433/986238

GLOSSARY

Term	Definition
Artificial Intelligence (AI)	Artificial intelligence is intelligence demonstrated or acquired by machines
Machine Learning (ML)	Machine learning is an area of research focused on comprehending and developing "learning" techniques, or techniques that use data to enhance performance on a certain set of tasks.
Natural Language Processing (NLP)	A branch of linguistics, computer science, and artificial intelligence called "natural language processing" is concerned with how computers and human languages interact.
Convolutional Neural Network (CNN)	A CNN is a particular type of network design for deep learning algorithms that is utilized for tasks like image recognition and pixel data processing.
Long short-term memory (LSTM)	An artificial neural network called long short-term memory is utilized in deep learning and artificial intelligence and it has feedback connections.
Named-entity recognition (NER)	NER is a subtask of information extraction which looks for named entities mentioned in unstructured text.
Natural-language understanding (NLU)	NLU is the understanding of the meaning and structure of human language by computers.

APPENDICES

Appendix A: Survey Form



NLP Research

Questions Responses 110 Settings

110 responses

Not accepting responses

Message for respondents

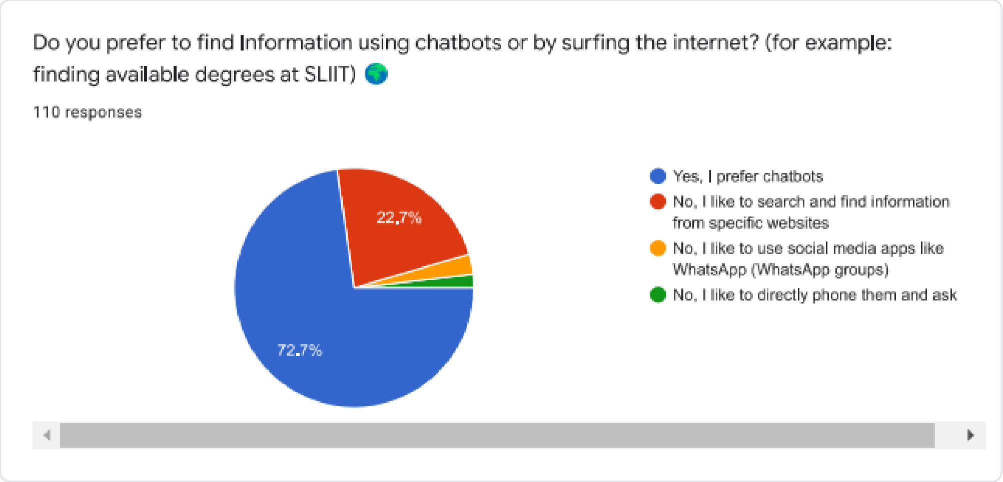
Hey 🙋 Thank you for the interest. We have temporarily stopped accepting responses for now. We will get back to you in case we are planning on resuming the survey. Have a nice day! 😎

Summary Question Individual

Do you know what a chatbot is? 🤖

110 responses

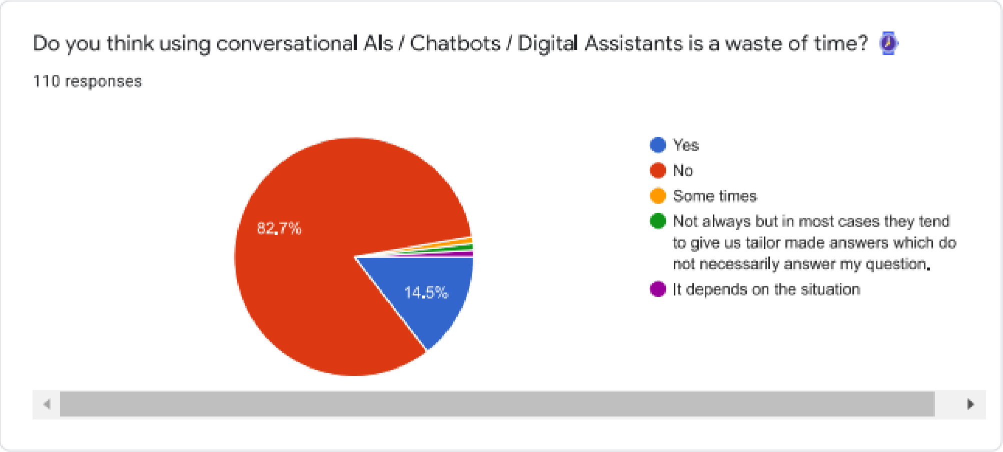
Figure A.1: Complete survey form questions and responses – part 1



Do you prefer to find Information using chatbots or by surfing the internet? (for example: finding available degrees at SLIIT) 🌍

110 responses

22.7%
72.7%
● Yes, I prefer chatbots
● No, I like to search and find information from specific websites
● No, I like to use social media apps like WhatsApp (WhatsApp groups)
● No, I like to directly phone them and ask

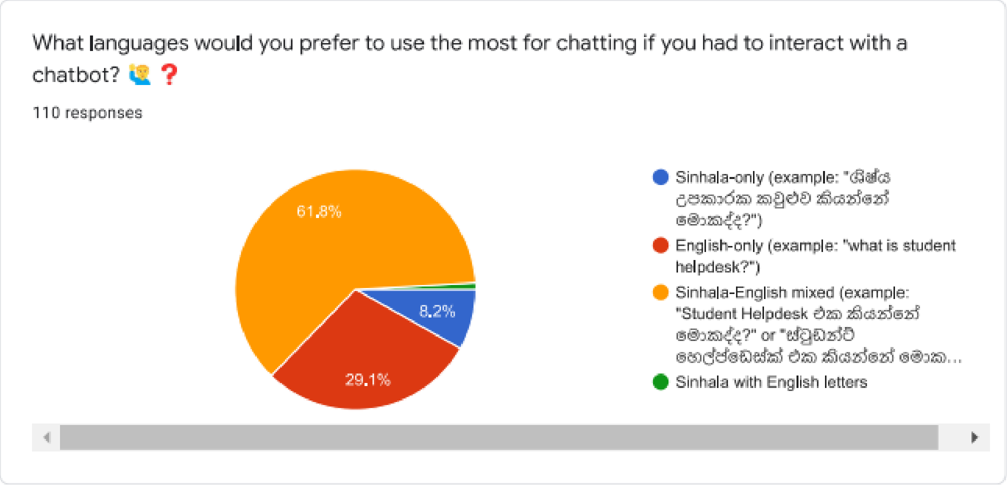


Do you think using conversational AIs / Chatbots / Digital Assistants is a waste of time? 🧭

110 responses

82.7%
14.5%
● Yes
● No
● Some times
● Not always but in most cases they tend to give us tailor made answers which do not necessarily answer my question.
● It depends on the situation

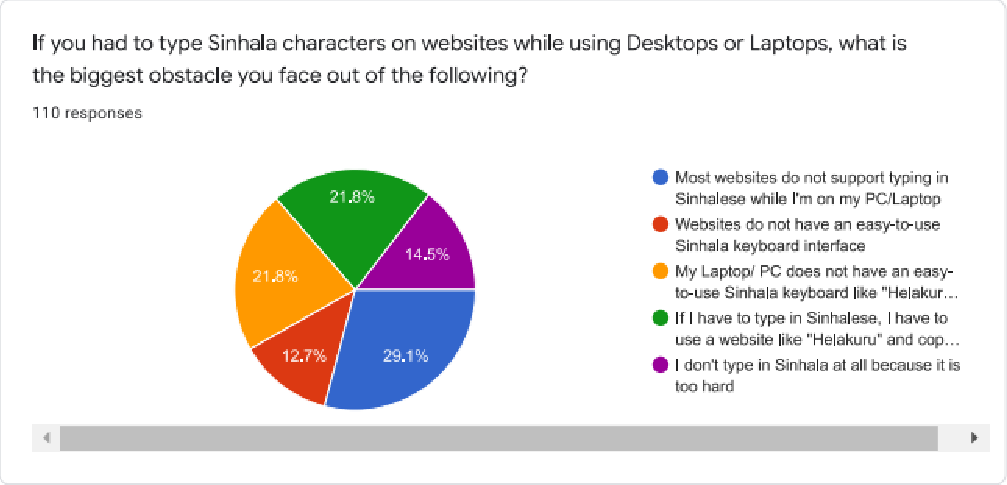
Figure A.2: Complete survey form questions and responses – part 2



What languages would you prefer to use the most for chatting if you had to interact with a chatbot? 🙋 ❓

110 responses

- ● Sinhala-only (example: "ශිෂ්‍ය උපකාරක කවුළුව කියන්නේ මොකද්ද?")
- ● English-only (example: "what is student helpdesk?")
- ● Sinhala-English mixed (example: "Student Helpdesk එක කියන්නේ මොකද්ද?" or "ස්ටුඩන්ට් හෙල්ප්ඩෙස්ක් එක කියන්නේ මොක…
- ● Sinhala with English letters

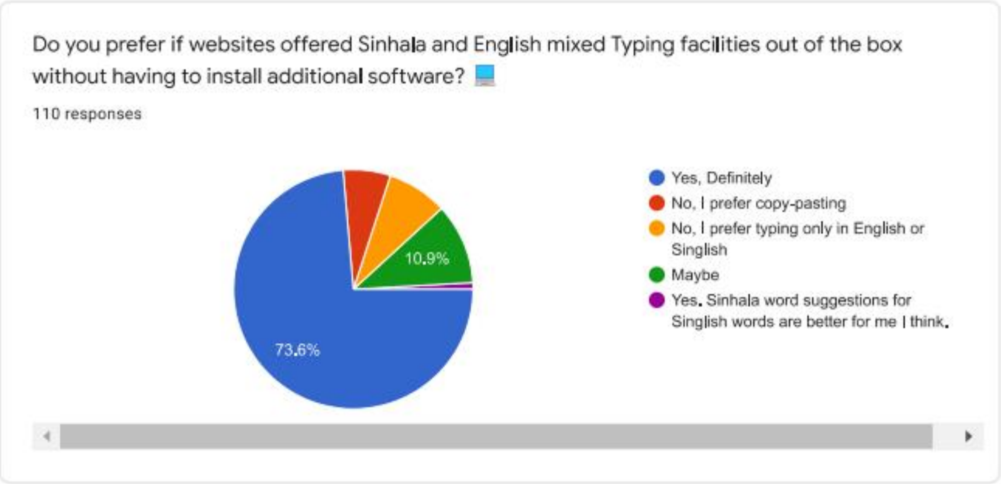


If you had to type Sinhala characters on websites while using Desktops or Laptops, what is the biggest obstacle you face out of the following?

110 responses

- ● Most websites do not support typing in Sinhalese while I'm on my PC/Laptop
- ● Websites do not have an easy-to-use Sinhala keyboard interface
- ● My Laptop/ PC does not have an easy-to-use Sinhala keyboard like "Helakur…
- ● If I have to type in Sinhalese, I have to use a website like "Helakuru" and cop…
- ● I don't type in Sinhala at all because it is too hard

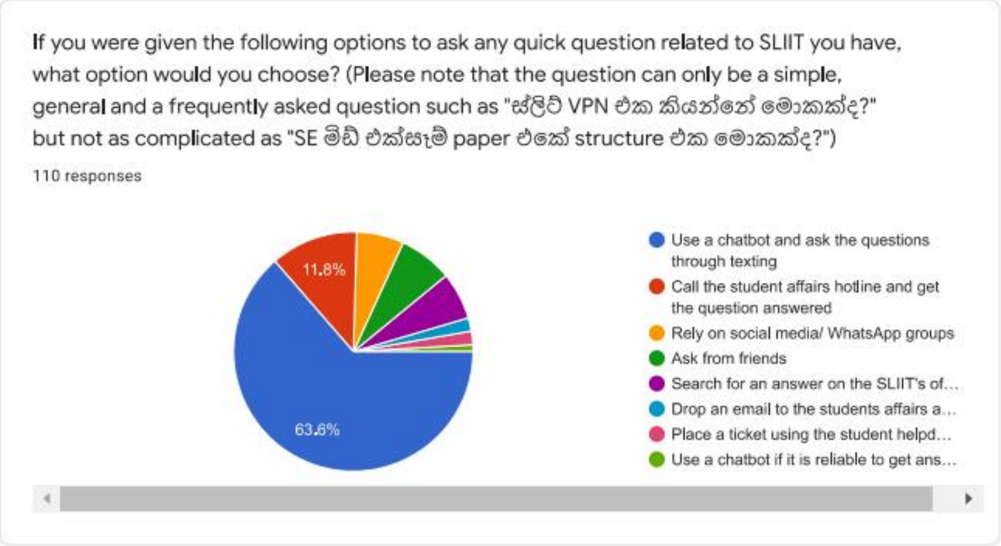
Figure A.3: Complete survey form questions and responses – part 3



Do you prefer if websites offered Sinhala and English mixed Typing facilities out of the box without having to install additional software? 💻

110 responses

● Yes, Definitely
● No, I prefer copy-pasting
● No, I prefer typing only in English or Singlish
● Maybe
● Yes. Sinhala word suggestions for Singlish words are better for me I think.
10.9%
73.6%

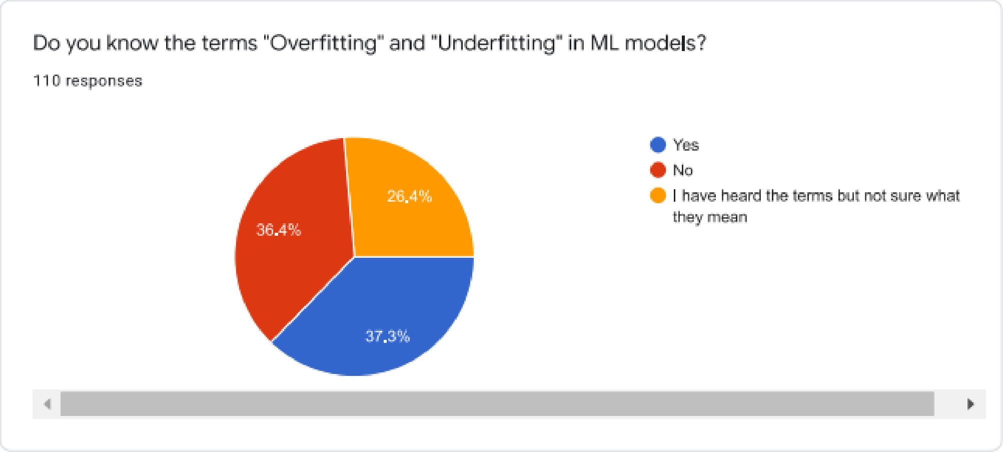


If you were given the following options to ask any quick question related to SLIIT you have, what option would you choose? (Please note that the question can only be a simple, general and a frequently asked question such as "ස්ලිට් VPN එක කියන්නේ මොකක්ද?" but not as complicated as "SE මිඩ් එක්සෑම් paper එකේ structure එක මොකක්ද?")

110 responses

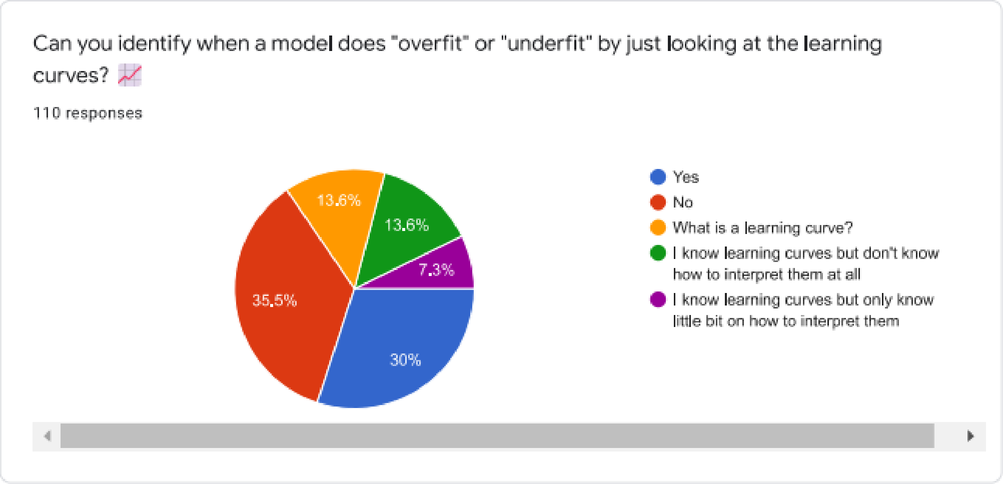
● Use a chatbot and ask the questions through texting
● Call the student affairs hotline and get the question answered
● Rely on social media/ WhatsApp groups
● Ask from friends
● Search for an answer on the SLIIT's of…
● Drop an email to the students affairs a…
● Place a ticket using the student helpd…
● Use a chatbot if it is reliable to get ans…
11.8%
63.6%

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



Do you know the terms "Overfitting" and "Underfitting" in ML models?

110 responses

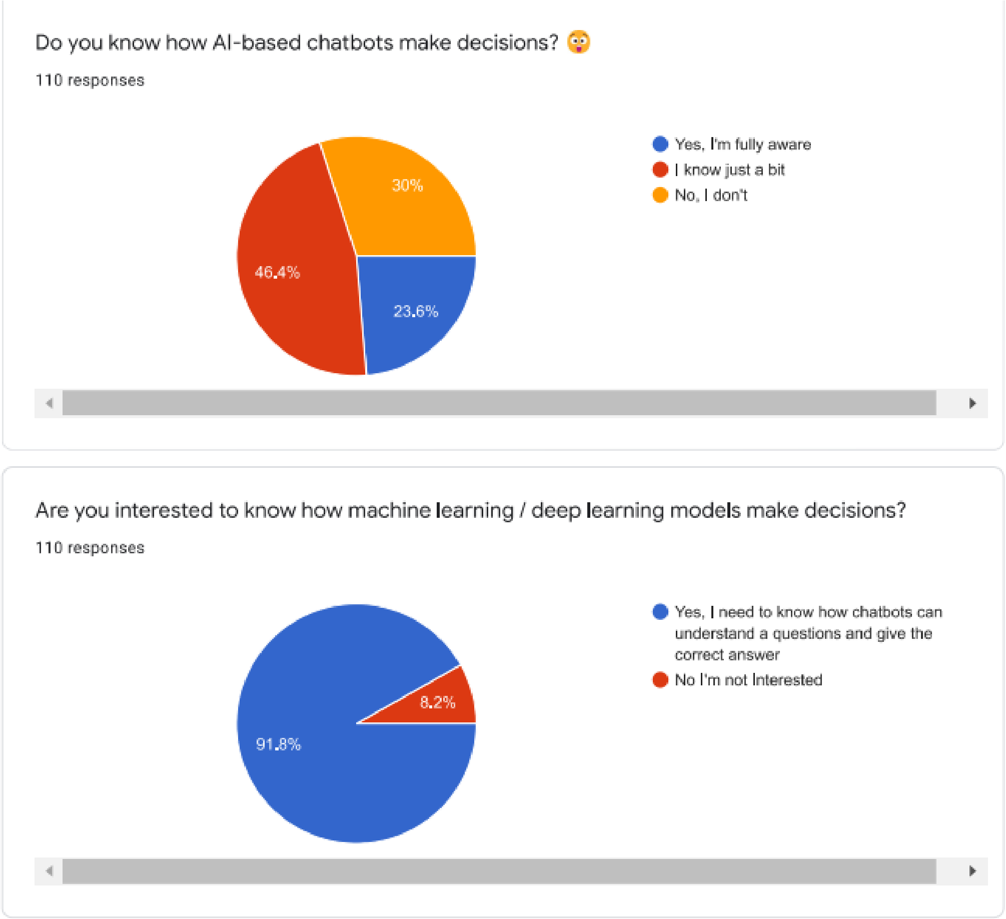


Can you identify when a model does "overfit" or "underfit" by just looking at the learning curves? 📈

110 responses

13.6%
13.6%
7.3%
35.5%
30%
● Yes
● No
● What is a learning curve?
● I know learning curves but don't know how to interpret them at all
● I know learning curves but only know little bit on how to interpret them

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



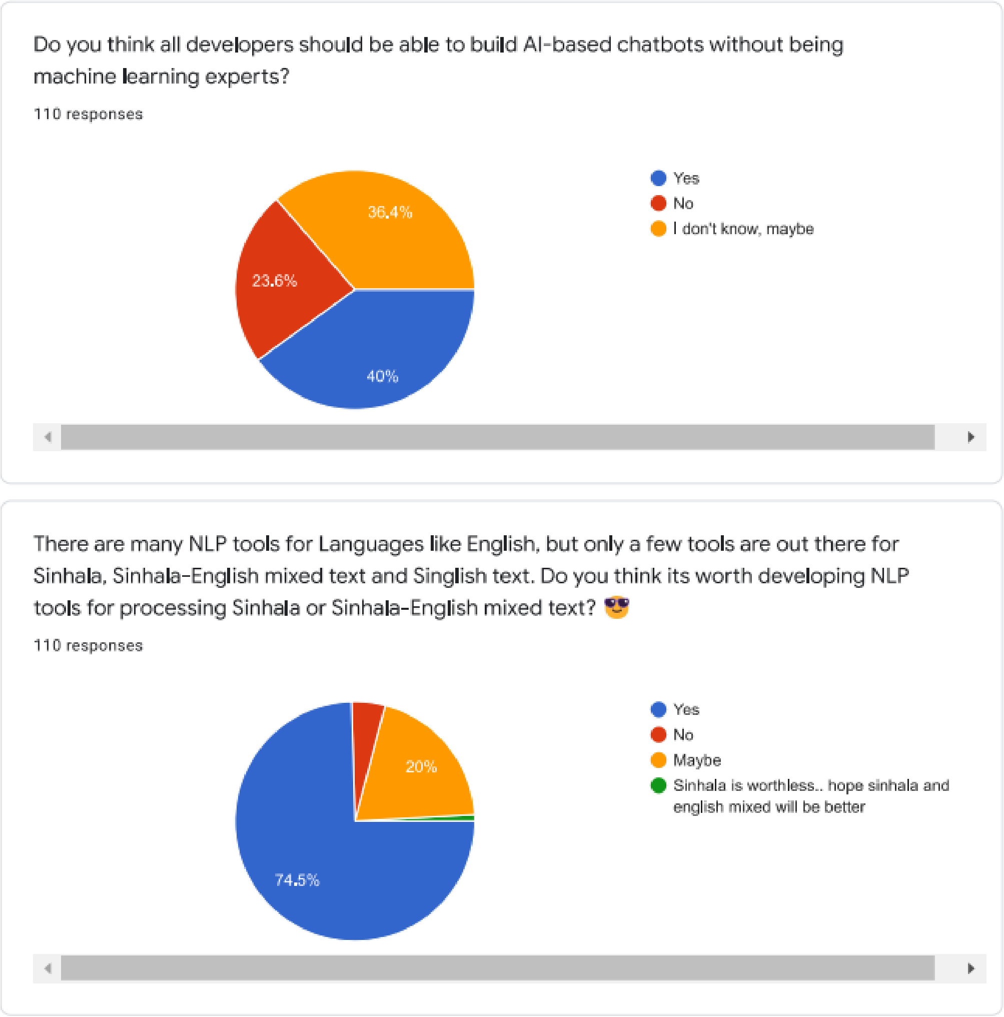
Do you know how AI-based chatbots make decisions? 😳

110 responses

Are you interested to know how machine learning / deep learning models make decisions?

110 responses

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



Do you think all developers should be able to build AI-based chatbots without being machine learning experts?

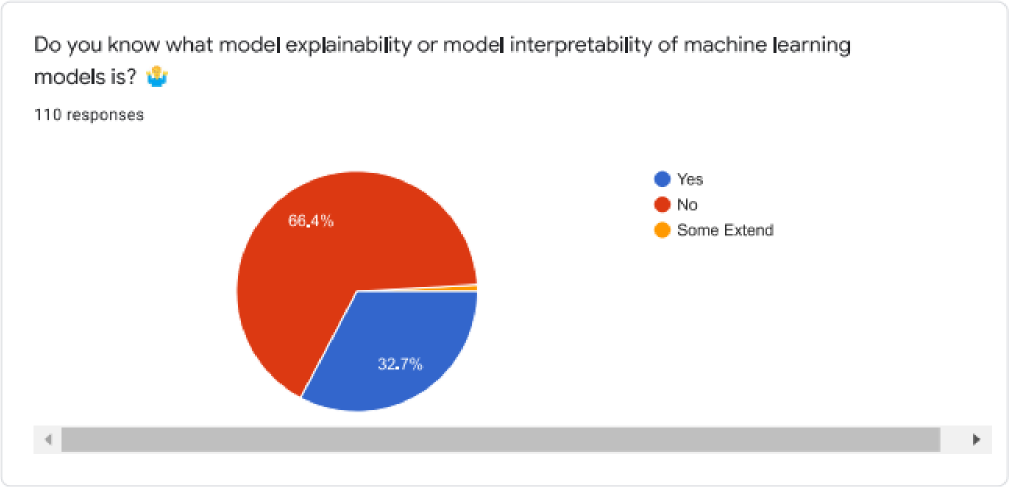
110 responses

There are many NLP tools for Languages like English, but only a few tools are out there for Sinhala, Sinhala-English mixed text and Singlish text. Do you think its worth developing NLP tools for processing Sinhala or Sinhala-English mixed text? 😎

110 responses

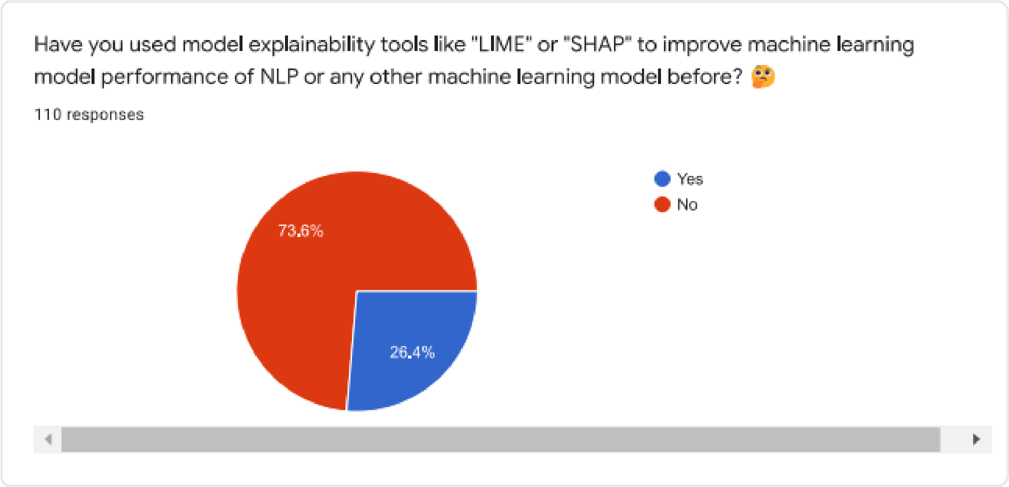
20%
74.5%
● Yes
● No
● Maybe
● Sinhala is worthless.. hope sinhala and english mixed will be better

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



Do you know what model explainability or model interpretability of machine learning models is? 🤷

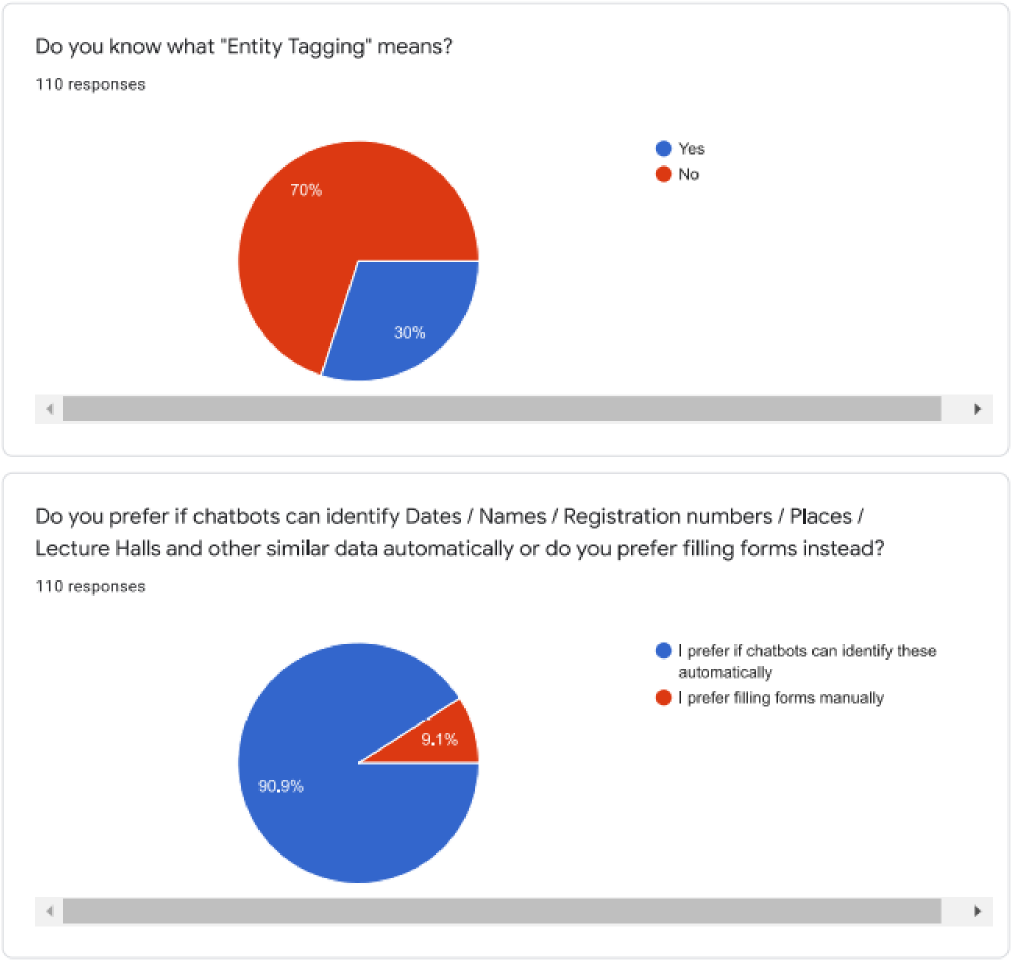
110 responses



Have you used model explainability tools like "LIME" or "SHAP" to improve machine learning model performance of NLP or any other machine learning model before? 🤔

110 responses

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



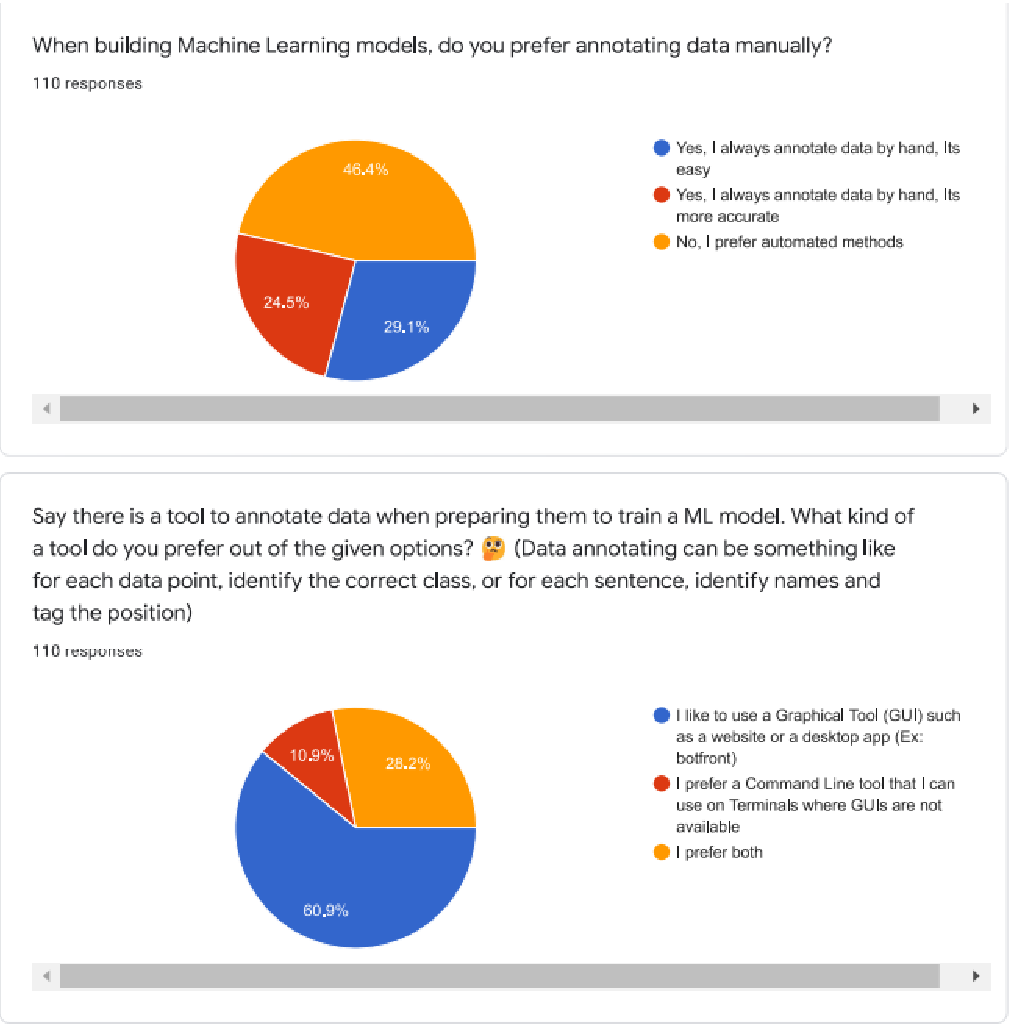
Do you know what "Entity Tagging" means?

110 responses

Do you prefer if chatbots can identify Dates / Names / Registration numbers / Places / Lecture Halls and other similar data automatically or do you prefer filling forms instead?

110 responses

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



When building Machine Learning models, do you prefer annotating data manually?

110 responses

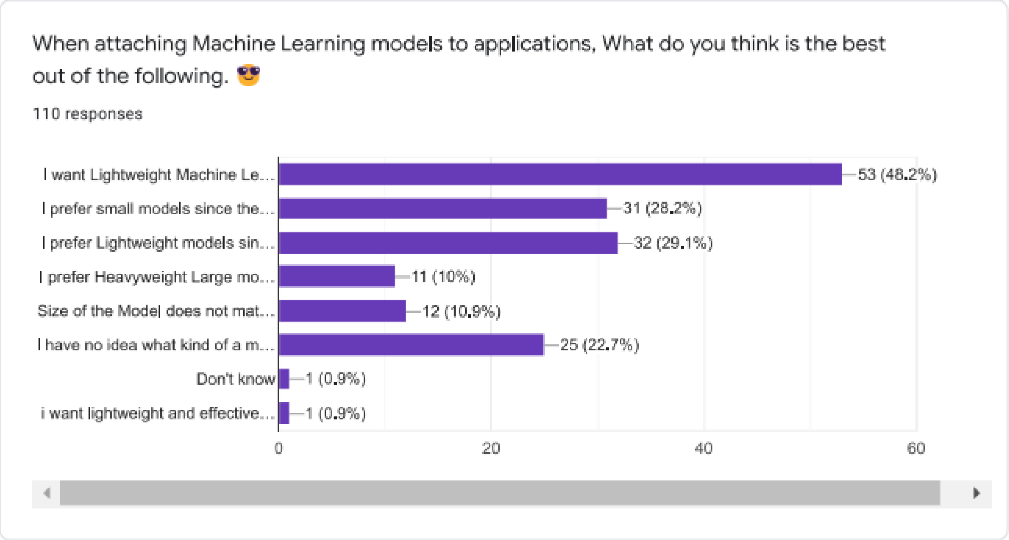
● Yes, I always annotate data by hand, Its easy
● Yes, I always annotate data by hand, Its more accurate
● No, I prefer automated methods
46.4%
24.5%
29.1%

Say there is a tool to annotate data when preparing them to train a ML model. What kind of a tool do you prefer out of the given options? 🤔 (Data annotating can be something like for each data point, identify the correct class, or for each sentence, identify names and tag the position)

110 responses

● I like to use a Graphical Tool (GUI) such as a website or a desktop app (Ex: botfront)
● I prefer a Command Line tool that I can use on Terminals where GUIs are not available
● I prefer both
10.9%
28.2%
60.9%

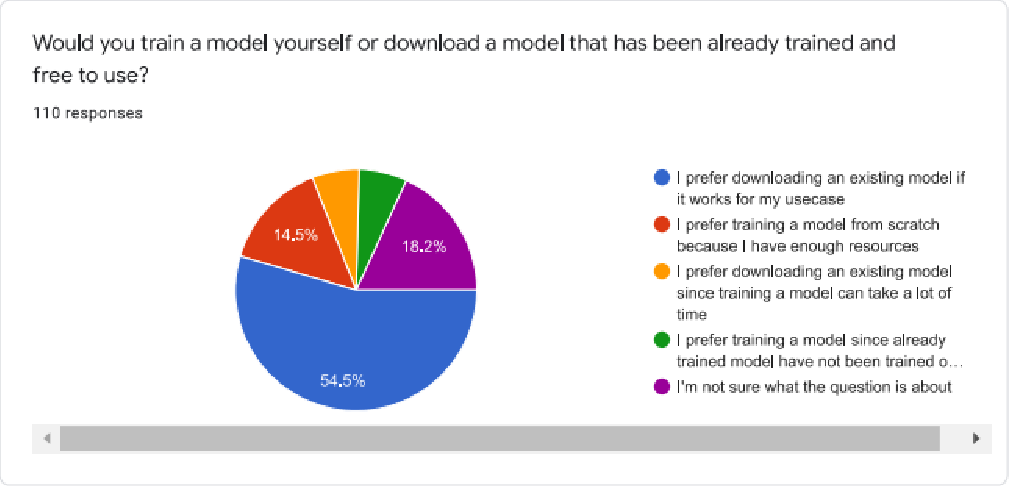
Figure A.10: Complete survey form questions and responses – part 10



When attaching Machine Learning models to applications, What do you think is the best out of the following. 😎

110 responses

I want Lightweight Machine Le…
53 (48.2%)
I prefer small models since the…
31 (28.2%)
I prefer Lightweight models sin…
32 (29.1%)
I prefer Heavyweight Large mo…
11 (10%)
Size of the Model does not mat…
12 (10.9%)
I have no idea what kind of a m…
25 (22.7%)
Don't know
1 (0.9%)
i want lightweight and effective…
1 (0.9%)
0
20
40
60



Would you train a model yourself or download a model that has been already trained and free to use?

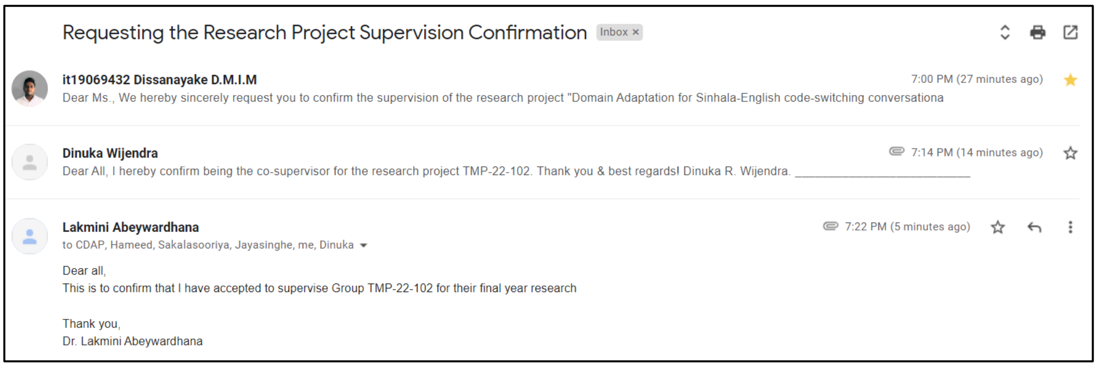
110 responses

14.5%
18.2%
54.5%
● I prefer downloading an existing model if it works for my usecase
● I prefer training a model from scratch because I have enough resources
● I prefer downloading an existing model since training a model can take a lot of time
● I prefer training a model since already trained model have not been trained o…
● I'm not sure what the question is about

Thank you! 🙇‍♀️

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

Appendix B: Supervision Confirmation Emails



Requesting the Research Project Supervision Confirmation

it19069432 Dissanayake D.M.I.M 7:00 PM (27 minutes ago)

Dear Ms., We hereby sincerely request you to confirm the supervision of the research project "Domain Adaptation for Sinhala-English code-switching conversationa

Dinuka Wijendra 7:14 PM (14 minutes ago)

Dear All, I hereby confirm being the co-supervisor for the research project TMP-22-102. Thank you & best regards! Dinuka R. Wijendra. ___________________________

Lakmini Abeywardhana 7:22 PM (5 minutes ago)

to CDAP, Hameed, Sakalasooriya, Jayasinghe, me, Dinuka

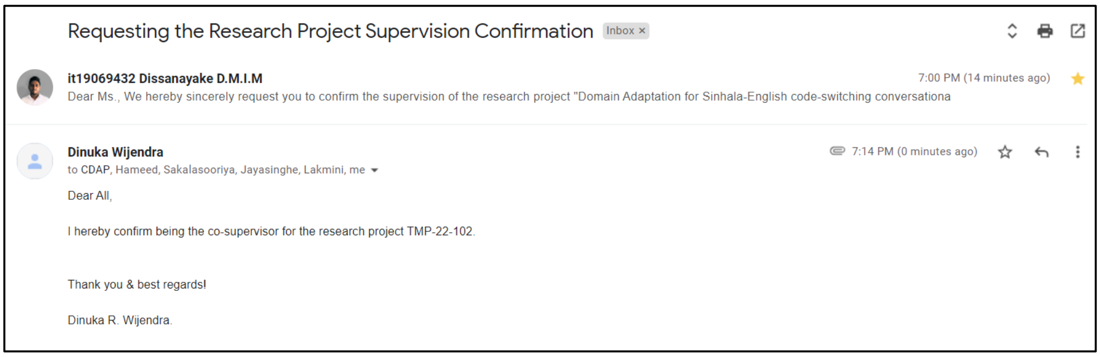
Dear all,

This is to confirm that I have accepted to supervise Group TMP-22-102 for their final year research

Thank you,

Dr. Lakmini Abeywardhana

Figure B.1: Research project supervision confirmation email



Requesting the Research Project Supervision Confirmation

it19069432 Dissanayake D.M.I.M 7:00 PM (14 minutes ago)

Dear Ms., We hereby sincerely request you to confirm the supervision of the research project "Domain Adaptation for Sinhala-English code-switching conversationa

Dinuka Wijendra 7:14 PM (0 minutes ago)

to CDAP, Hameed, Sakalasooriya, Jayasinghe, Lakmini, me

Dear All,

I hereby confirm being the co-supervisor for the research project TMP-22-102.

Thank you & best regards!

Dinuka R. Wijendra.

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

Appendix C: SIENA user interfaces

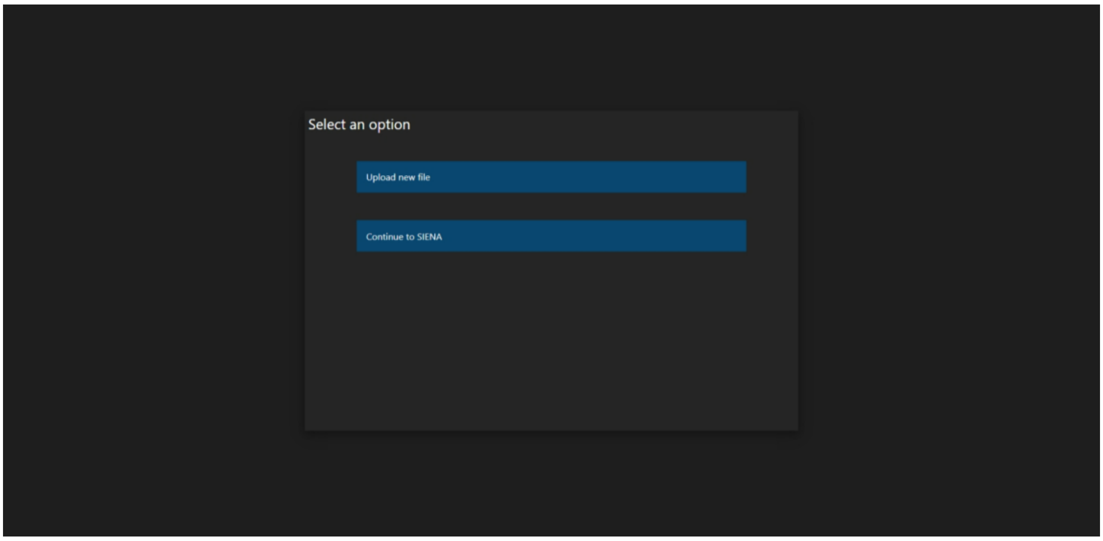


Figure C.1: SIENA initial pathway selection user interface

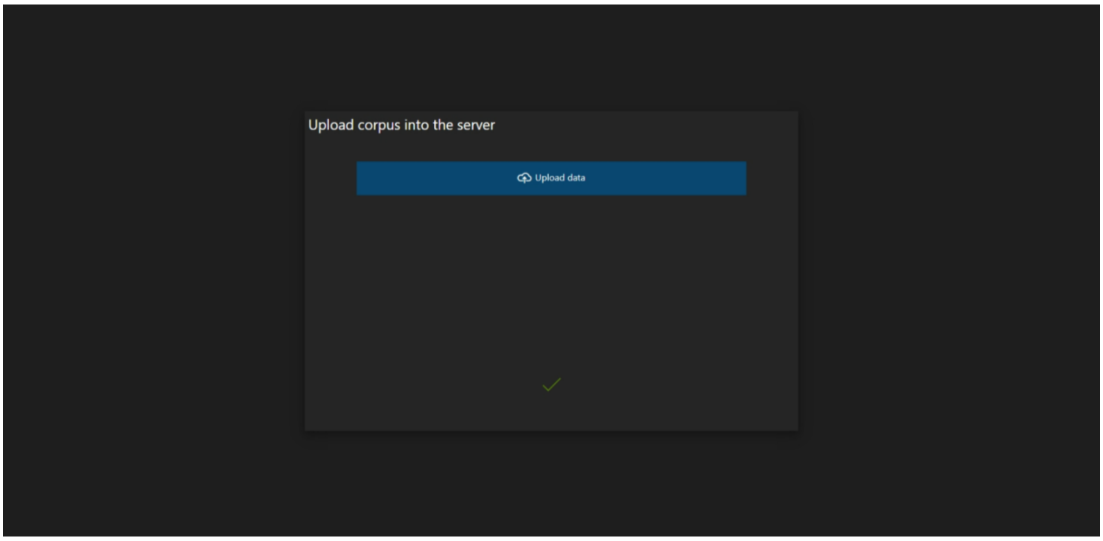


Figure C.2: SIENA file selection user interface

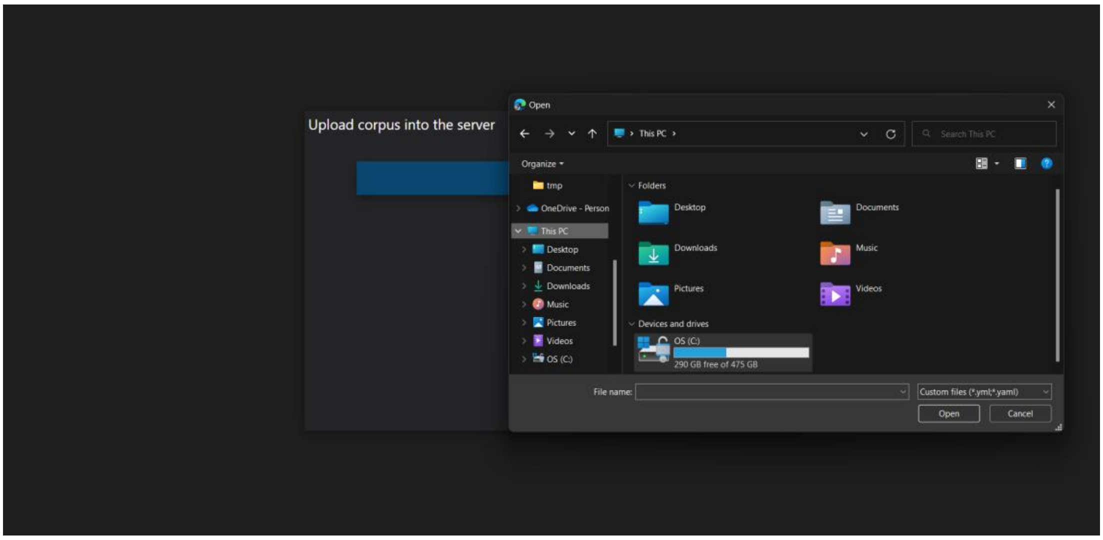
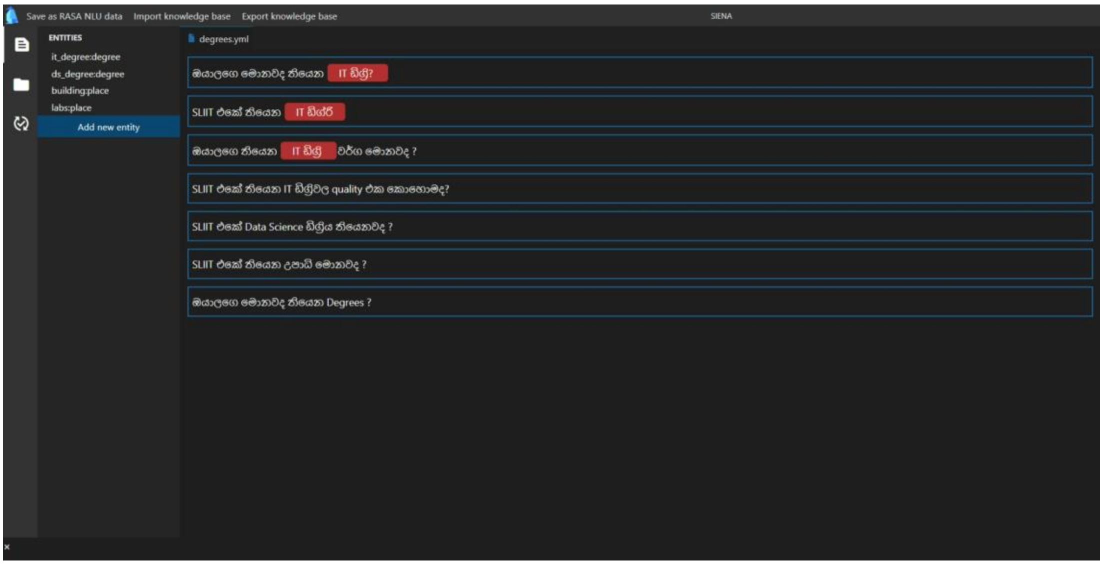


Figure C.3: SIENA file picker



ඔයාලගෙ මොනවද තියෙන IT ඩිග්‍රී?

SLIIT එකේ තියෙන IT ඩිග්‍රී

ඔයාලගෙ තියෙන IT ඩිග්‍රී වර්ග මොනවද ?

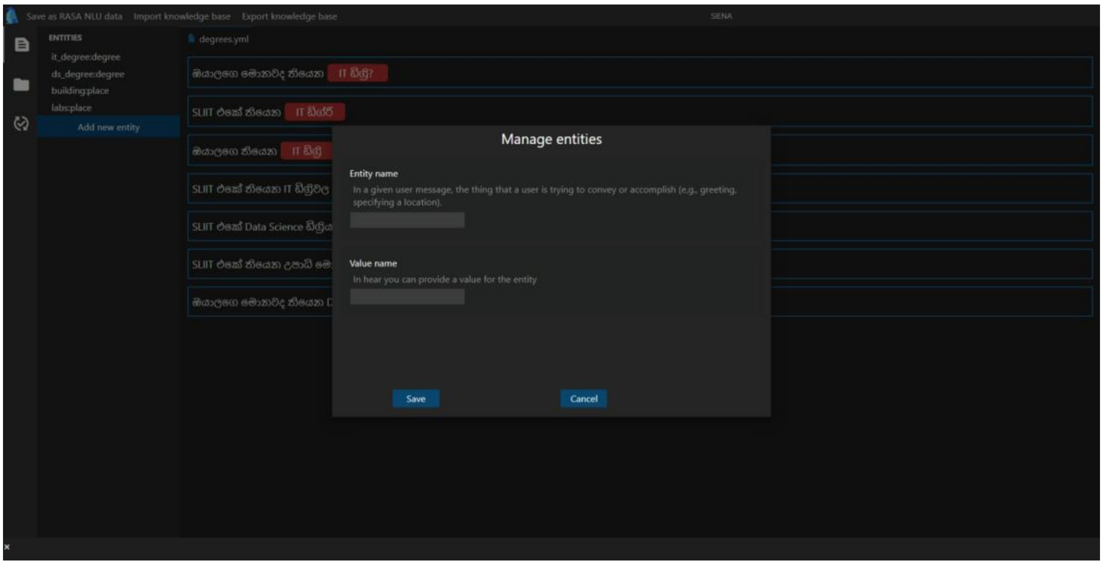
SLIIT එකේ තියෙන IT ඩිග්‍රීවල quality එක කොහොමද?

SLIIT එකේ Data Science ඩිග්‍රීය තියෙනවද ?

SLIIT එකේ තියෙන උපාධි මොනවද ?

ඔයාලගෙ මොනවද තියෙන Degrees ?

Figure C.4: SIENA entity list side menu



ඔයාලගෙ මොනවද තියෙන IT ඩිග්‍රී?

SLIIT එකේ තියෙන IT ඩිග්රී

ඔයාලගෙ තියෙන IT ඩිග්‍රී

SLIIT එකේ තියෙන IT ඩිග්‍රීවල

SLIIT එකේ Data Science ඩිග්‍රිය

SLIIT එකේ තියෙන උපාධි මෙ

ඔයාලගෙ මොනවද තියෙන D

Manage entities

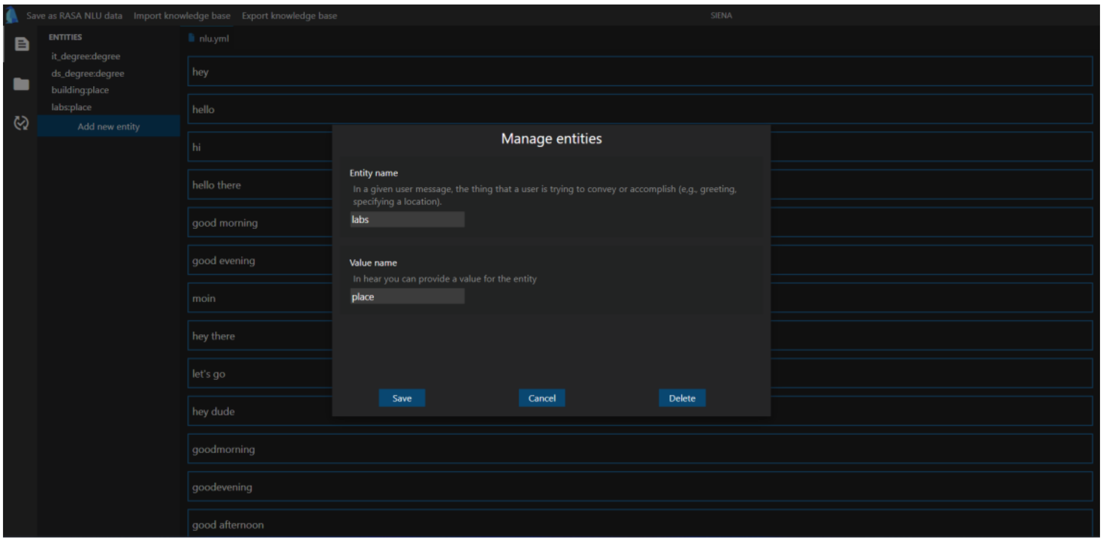
Entity name

In a given user message, the thing that a user is trying to convey or accomplish (e.g., greeting, specifying a location).

Value name

In hear you can provide a value for the entity

Figure C.5: SIENA add new entity user interface



hey

hello

hi

hello there

good morning

good evening

moin

hey there

let's go

hey dude

goodmorning

goodevening

good afternoon

Manage entities

Entity name

In a given user message, the thing that a user is trying to convey or accomplish (e.g., greeting, specifying a location).

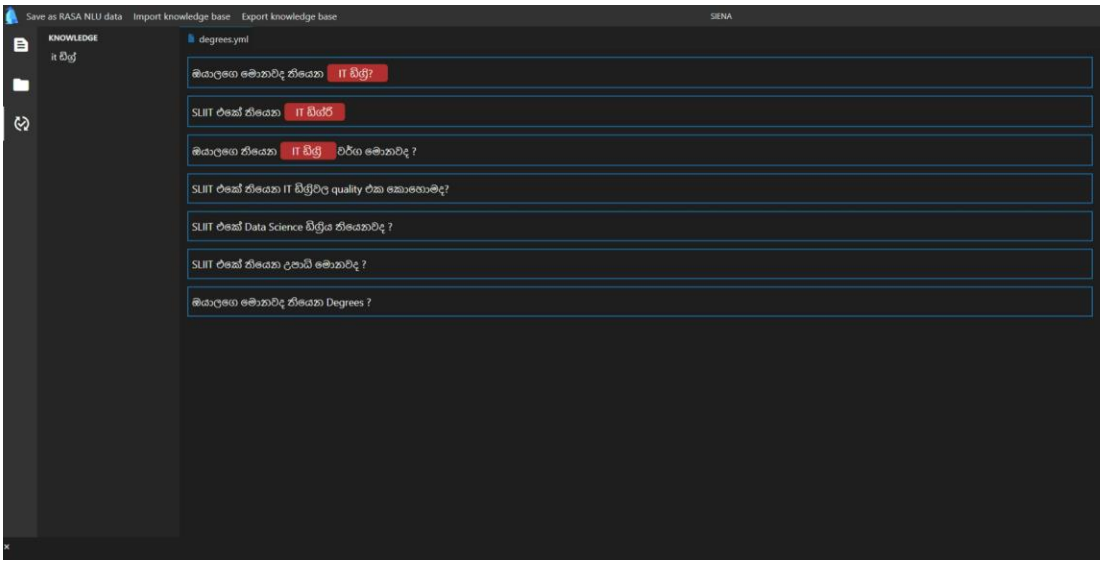
labs

Value name

In hear you can provide a value for the entity

place

Figure C.6: SIENA edit entity user interface



ඔයාලගෙ මොනවද තියෙන IT ඩිග්‍රී?

SLIIT එකේ තියෙන IT ඩිග්රී

ඔයාලගෙ තියෙන IT ඩිග්‍රී වර්ග මොනවද ?

SLIIT එකේ තියෙන IT ඩිග්‍රීවල quality එක කොහොමද?

SLIIT එකේ Data Science ඩිග්‍රිය තියෙනවද ?

SLIIT එකේ තියෙන උපාධි මොනවද ?

ඔයාලගෙ මොනවද තියෙන Degrees ?

Figure C.7: SIENA base word list side menu

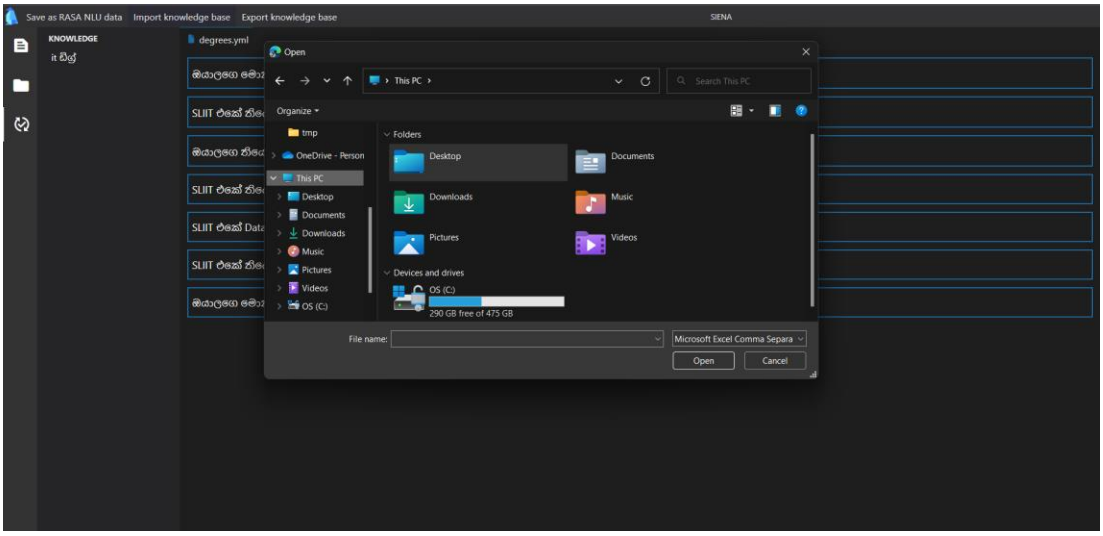
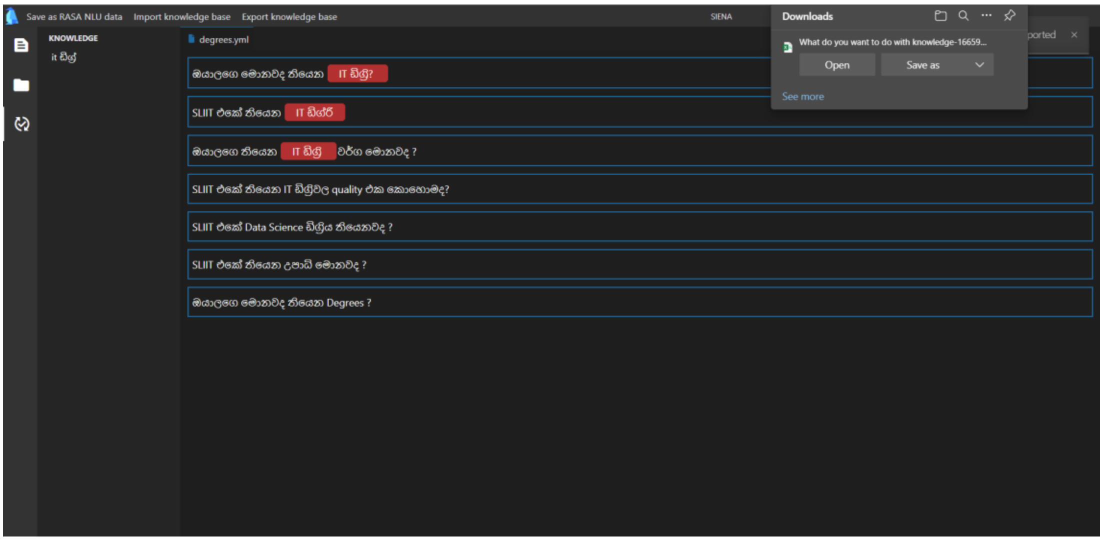


Figure C.8: SIENA knowledge base file picker



What do you want to do with knowledge-16659...

ඔයාලගෙ මොනවද තියෙන IT ඩිග්‍රී?

SLIIT එකේ තියෙන IT ඩිග්රි

ඔයාලගෙ තියෙන IT ඩිග්‍රී වර්ග මොනවද ?

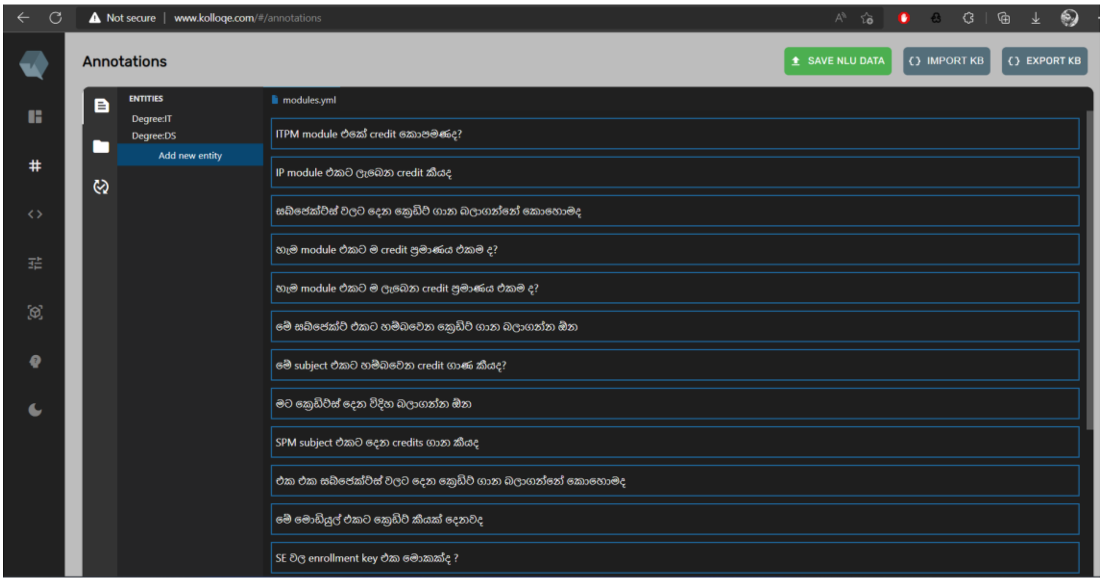
SLIIT එකේ තියෙන IT ඩිග්‍රීවල quality එක කොහොමද?

SLIIT එකේ Data Science ඩිග්‍රිය තියෙනවද ?

SLIIT එකේ තියෙන උපාධි මොනවද ?

ඔයාලගෙ මොනවද තියෙන Degrees ?

Figure C.9: SIENA knowledge base export



ITPM module එකේ credit කොපමණද?

IP module එකට ලැබෙන credit කීයද

සබ්ජෙක්ට්ස් වලට දෙන ක්‍රෙඩිට් ගාන බලාගන්නේ කොහොමද

හැම module එකට ම credit ප්‍රමාණය එකම ද?

හැම module එකට ම ලැබෙන credit ප්‍රමාණය එකම ද?

මේ සබ්ජෙක්ට් එකට හම්බවෙන ක්‍රෙඩිට් ගාන බලාගන්න ඕන

මේ subject එකට හම්බවෙන credit ගාණ කීයද?

මට ක්‍රෙඩිට්ස් දෙන විදිහ බලාගන්න ඕන

SPM subject එකට දෙන credits ගාන කීයද

එක එක සබ්ජෙක්ට්ස් වලට දෙන ක්‍රෙඩිට් ගාන බලාගන්නේ කොහොමද

මේ මොඩියුල් එකට ක්‍රෙඩිට් කීයක් දෙනවද

SE වල enrollment key එක මොකක්ද ?

Figure C.10: Integrated SIENA into Kolloqe