# An Inappropriate Word Detector for The Sinhala to English and English to Sinhala Translator (SEES)

**D. S. U. Arachchi[1], R. P. N. M. Herath[2], M. B. P. T. H. Gunaratne[3],K. T. Hansana[4], E. Weerasinghe[5], D. I. De Silva[6]**

*dilanshanuka999@gmail.com[1],naveenmalshan6@gmail.com[2], bathiyapathum@icloud.com[3], thilinahansana1100@gmail.com[4], eishan.w@sliit.lk[5], dilshan.i@sliit.lk[6],*

*Dept. ofComputer Scienceand Software Engineering,Sri Lanka Institute of Information Technology*
*Malabe, Sri Lanka.*

***Abstract:*** -This paper describes the creation and application of web-based tool for translating between English and Sinhala as well as Sinhala and English with features that improve user involvement, feedback, and content filtering. The main objective of the app is to give users accurate translations with inappropriate word detection. This translation tool's main features include a comprehensive user experience with inappropriate word detection. The application provides the relevant Sinhala meaning when a user enters an English term. Users have the ability to provide feedback on errors in translation, recognizing the possibility of meaning inconsistencies. This user-driven feedback system encourages ongoing enhancements to translation quality. Another important factor addresses the user population, which focuses on children. The system carefully saves any inappropriate word input by a user, especially children, within an unchangeable history. Parents can monitor their children's behaviour through this historical record by reading the entries and giving advice. This app helps to translate English as well as prevent children from being misled.

***Keywords****: Sinhala to English translator, English to Sinhala translator, bad word detector*

## 1. Introduction

The language Sinhala, a member of the Indo-Aryan linguistic family, it has some unique characteristics not found in other Indo-Aryan languages due to the influence of other languages in the area. In Sinhala, there are 14 vowels (the sounds people produce with their voice) and 26 consonants (the noises people make with their lips) [1]. Sinhala is the oldest language in the world, spoken by about sixteen million people; approximately thirteen million of them are native speakers [2], it is written from left side to right side in literature. In this style of writing, the subject or object of a phrase must agree with the time, gender, and the number of subjects in the verb (the action word). Unlike English, every Sinhala word is spoken exactly as it is written [3].

English is a language that originated in England and is part of the West German language family [3]. It is the primary language in the United Kingdom, the United States, and countries that were once a part of the British Empire, such as Australia, Canada, and New Zealand [1]. English has grown in popularity all across the world as a result of its widespread use as an important language. People from countries where English is not the primary language make a concerted effort to learn it. Knowing how to read and speak English will help individuals accomplish better in their professions and academics. It also provides them with an additional advantage.

The recognition of a translation gap encountered by many groups of consumers, especially children, is a vital componentthat propels the effort of the project team [4]. When people attempt to translate an insulting or inappropriate language, it frequently leads to unintended or even harmful results. For instance, young children may accidentally attempt to translate inappropriate words, which can cause misunderstandings, confusion, and even the spread of harmful content.

In conclusion, the primary objective of this research is the development of a text translation tool that enables text conversion between English and Sinhala. And the main goal is to create an effective filtering system that efficiently separates content from inappropriate language, keeping children from grabbing harmful information. SEES web application has a unique feature that enables parents to keep track of their children's search actions, particularly by spotting potentially harmful content. This project demonstrates the dedication of the team to user engagement and responsible technology use. By engaging to this content, the authors aim to provide readers with a better understanding of how this research project functions in its complex combination of technology, language, culture, and the common desire for worldwide connectivity.

This research offers an informative examination effort of the research team to improve Sinhala and English communication and vice versa through the SEES Web Application. They illustrate the unique characteristics of SEES web application and understand the languages it interacts with by demonstrating the difficulties caused by the translation gap, particularly in connection to children's usage. This creative approach, which is willing to successfully address these difficulties, is an instance of the researcher's commitment to responsible technology deployment and user involvement. In the SEES project, technology converges with language, culture, and the need for true global connection in a complex journey. The research paper cordially extends an invitation for readers to explore the complexity that the article reveals as they read it.

## 2. Objectives

In today's world, machine translation has become quite popular because it helps to save time and effort when turning words, sentences, and paragraphs from one language into another. People have made many translation systems for different languages, but they mostly focus on languages like Indo-European and Indo-Aryan [5].

However, there's something important to know about the languages Sinhala and English. They're quite different in how they're built, so they're called non-similar languages. Because of this, there aren't a lot of systems that can change things easily between Sinhala and English. This happens because the way these two languages are put together is different [3].

The task of translating non-similar languages is much more challenging than translating similar languages. For instance, since English, French, and German are related languages with similar alphabets, translating from one to the other is simpler than translating from Sinhalese to English [6].

It is simpler to communicate with the people from different geographical locations because of the language translation software [5]. It's beneficial since it makes it possible for people who speak different languages to communicate with one another. There are some well-known translators such as the Uni-code Converter [7], Google Transliteration IME [8], and the converter listed in [9] are notable converters. By allowing users to enter words based on their English phonetics (pronunciation), these converters produce equivalent Sinhala words with the appropriate Sinhala fonts.

There have also been suggestions for dictionaries that can give both English and Sinhala definitions for words, in addition to translation software and converters. The Madura dictionary [10], which contains over 230,000 definitions in both English and Sinhala, serves as an example.

Parents want to ensure that their children are protected online [11]. This SEES translator translates English to Sinhala and enables parents to view what their children are doing online. It reveals anything that could not be suitable for children. This tool assists children online in the same way that parents do in real life. It's like walking with them as they browse the internet. Also, this feature has also been proposed for hate speech detection in online user comments [12].

It discusses the approach to working on large-scale data sets, and online safety. It is possible to utilize particular algorithms to find offensive words from a phase by matching those words. The Aho-Corasick algorithm is an efficient string-matching algorithm, it can quickly locate and highlight particular words or phrases [13]. The

Aho-Corasick Algorithm is also used for recognizing instances of malware in the field of cyber security [14]. It is particularly suited for finding well-known malware signatures in files, network traffic, or system memory since it can quickly search for several patterns at once.

## 3. Methods

A lot of attention in this project is given to the object-oriented methodology which is a foundation in the field of software development, it offers a structured and systematic approach to building reliable systems. The basic idea behind this concept is to group data and functions into separate "objects." These objects have both the responsibilities of data carriers and action executors, functioning as the fundamental building blocks of the software architecture. This approach is widely applied when creating secure software programs. The authors use the MERN (MongoDB, Express.js, React, and Node.js) stack for developing this application, which is an advanced and widely used technology stack that naturally combines with the concepts of object-oriented methodology.

Using the MERN stack technique, the SEES Translator and inappropriate content identification system is developed throughout every level of the project's development. It provides assistance with the overall structure of the software's front-end and back-end development. The project team used this method to put it simply ensure that the software is well organized and user-friendly, as well as maintain balance and a clean environment.

The entire translation procedure is focused on the user's experience, with accuracy and simplicity being its top priorities. The user of that application has access to two different input boxes. One for the input Sinhala phase to translations for English, and the other for the output of the translation; it can also use vice versa (Fig1).

The sinhala or English meaning of a term is displayed right away as the user enters it in the proper field. When using this application, users can view the results of their translation quickly because of this application's quick response, which enhances the translation process of this application.

The authors have effectively used the Google Translation service API, provided available through the Rapid API platform, for the process of translation between Sinhala and English within this project. The system delivers dynamic and accurate translation capabilities by utilizing this powerful technology. It also enables users to easily translate text between English and Sinhala quickly and efficiently.

The researchers used the Google Translation API, facilitated by the Rapid API, for the translation process. This integration not only increases the software's flexibility but also demonstrates the project's efforts to use cutting-edge technologies for accurate and efficient language translation.

The authors have implemented a content filtering system that has been designed to recognize the translation of abusive or unsuitable words employing the Aho-Corasick Algorithm. It was created by Alfred V. Aho and Margaret J. Corasick in 1975. It is an extremely powerful string-matching algorithm that is widely recognized

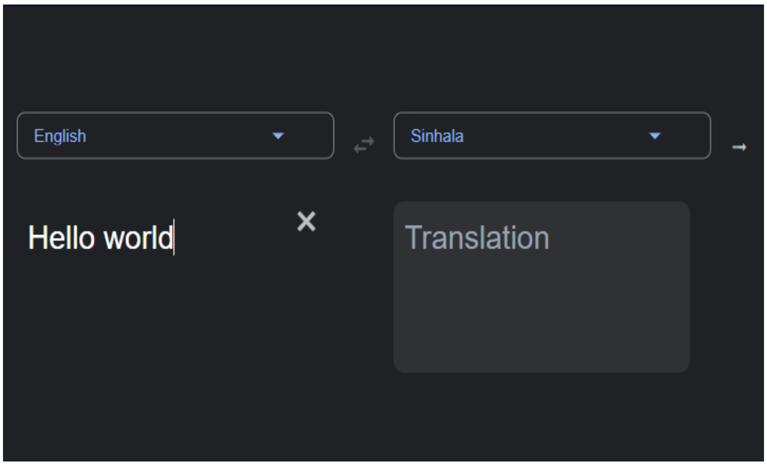

English
Sinhala
Hello world

**Fig 1: Input Boxes**

for its excellent performance in pattern matching. It builds a customized data structure known as a Trie to search for several patterns in a given text at the same time.

Using this algorithm, the authors created a content detection system that can identify the translation of terms that are offensive or inappropriate. The authors have been able to greatly improve the effectiveness of this content filtering feature by using this algorithm. The text to be translated by the user is the input to this offensive word detection algorithm.

The Aho-Corasick Algorithm is used in a variety of industries, including content filtering, intrusion detection systems, antivirus software, and natural language processing. It is a popular option in cybersecurity because of its effectiveness in multiple pattern matching, which allows it to quickly find harmful code or keywords in network data. This Algorithm also guarantees great performance even when used with large datasets according to its linear time complexity. This quality is especially useful in data mining, where the algorithm can quickly search through huge quantities of data for particular patterns or phrases.

Furthermore, this algorithm’s effectiveness in pattern matching and versatility in a range of applications are both highlighted by the authors' use of it for content filtering. The algorithm's success in identifying abusive language highlights the importance of keeping an online environment that is safer and more respectful.

This translation platform provides English-to-Sinhala and Sinhala-to-English translation services in the context of content filtering and translation. The program gives users the option to enter content in any language, and it makes sure that translations keep a polite and non-offensive tone. Consider a situation where someone adds the phrase "This is horrible, and I hate it." The Aho-Corasick Algorithm is used in this instance by the platform's translation system to find potentially offensive words in the input.

The terms "horrible" and "hate" are accurately detected by the system when it is combined with a list of offensive keywords and their Sinhala equivalents. The input is quickly searched for these keywords and their translations, ensuring a thorough filtering process.

The pattern matching procedure starts by adding identified inappropriate list of bad words (Fig2) into the algorithm; it takes inappropriate words and separate characters one by one, then creates the Trie structure as (Fig3).

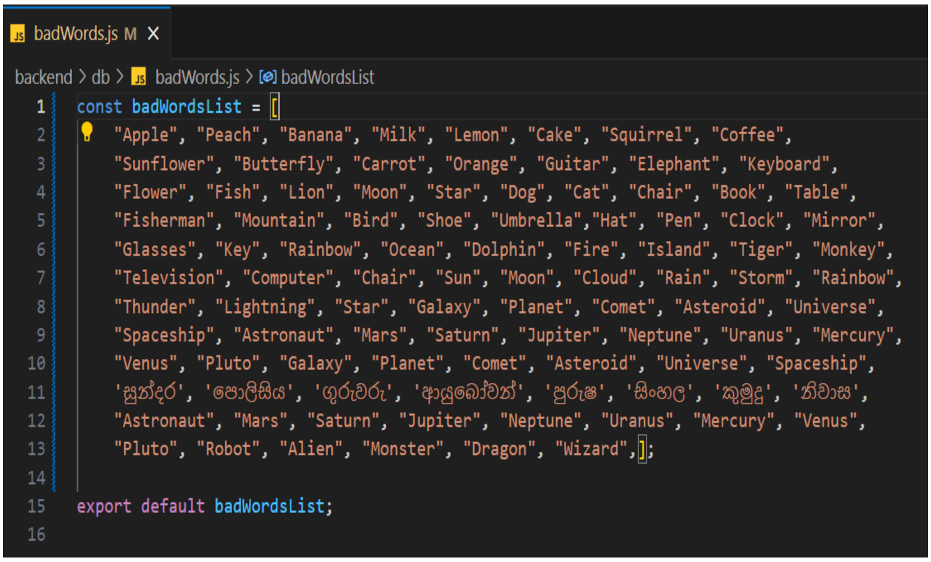
According to (Fig 4) the algorithm was given the pattern of “APPLE”, this algorithm begins with the root element and adds characters of “APPLE” one by one below with a reference to the next character. Using the dark greycolour circle indicates the final state if the search string comes to the final state, it shows that the pattern has occurred in the input string.

In (Fig 4), the valid transitions are represented by blue colour lines whereas failure transitions are represented by green colour lines. Failure transitions facilitate effective back tracking, allowing the system to find patterns across the input text without unnecessary reprocessing so that the effectiveness of the algorithm is improved.

After that using the Trie, the algorithm then starts the scan process for the input text, starting from the root node, and smoothly traverses down character by character comparing to the pattern. If the next node contains the letter that is required to fulfil the pattern algorithm traverses to that child node. If the node does not contain the required value using the failure link it will travers back to the root node.

The algorithm keeps track of the place and context of each match it finds inside the input text and provides the pattern found in the input text. This will repeat until the end of the input string; after that, it will return the detected bad words from the sentence and indicate that it found some bad words in it.

In addition, the authors make sure that user privacy is protected and that the content detection procedure does not inappropriately access or expose personal data. With a more comprehensive and flexible option at their disposal, this content filtering system is still strong and configurable, ensuring that the system can check for any occurrences of inappropriate language with a suitable approach.



```
const badWordsList = [
    "Apple", "Peach", "Banana", "Milk", "Lemon", "Cake", "Squirrel", "Coffee",
    "Sunflower", "Butterfly", "Carrot", "Orange", "Guitar", "Elephant", "Keyboard",
    "Flower", "Fish", "Lion", "Moon", "Star", "Dog", "Cat", "Chair", "Book", "Table",
    "Fisherman", "Mountain", "Bird", "Shoe", "Umbrella","Hat", "Pen", "Clock", "Mirror",
    "Glasses", "Key", "Rainbow", "Ocean", "Dolphin", "Fire", "Island", "Tiger", "Monkey",
    "Television", "Computer", "Chair", "Sun", "Moon", "Cloud", "Rain", "Storm", "Rainbow",
    "Thunder", "Lightning", "Star", "Galaxy", "Planet", "Comet", "Asteroid", "Universe",
    "Spaceship", "Astronaut", "Mars", "Saturn", "Jupiter", "Neptune", "Uranus", "Mercury",
    "Venus", "Pluto", "Galaxy", "Planet", "Comet", "Asteroid", "Universe", "Spaceship",
    'සුන්දර', 'පොලිසිය', 'ගුරුවරු', 'ආයුබෝවන්', 'පුරුෂ', 'සිංහල', 'කුමුදු', 'නිවාස',
    "Astronaut", "Mars", "Saturn", "Jupiter", "Neptune", "Uranus", "Mercury", "Venus",
    "Pluto", "Robot", "Alien", "Monster", "Dragon", "Wizard",];

export default badWordsList;

```

**Fig 2: Bad words list**

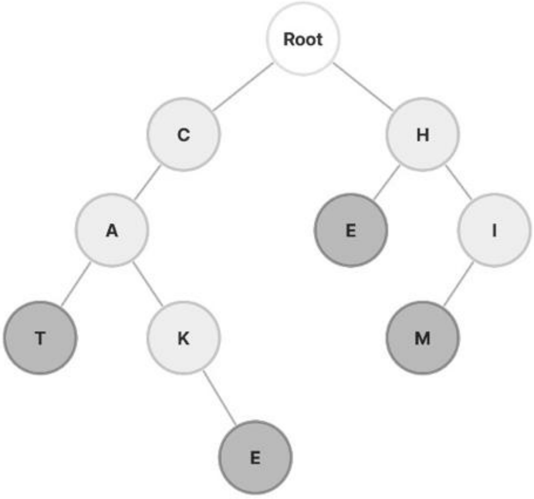


**Fig3:Aho-Corasick Trie structure**

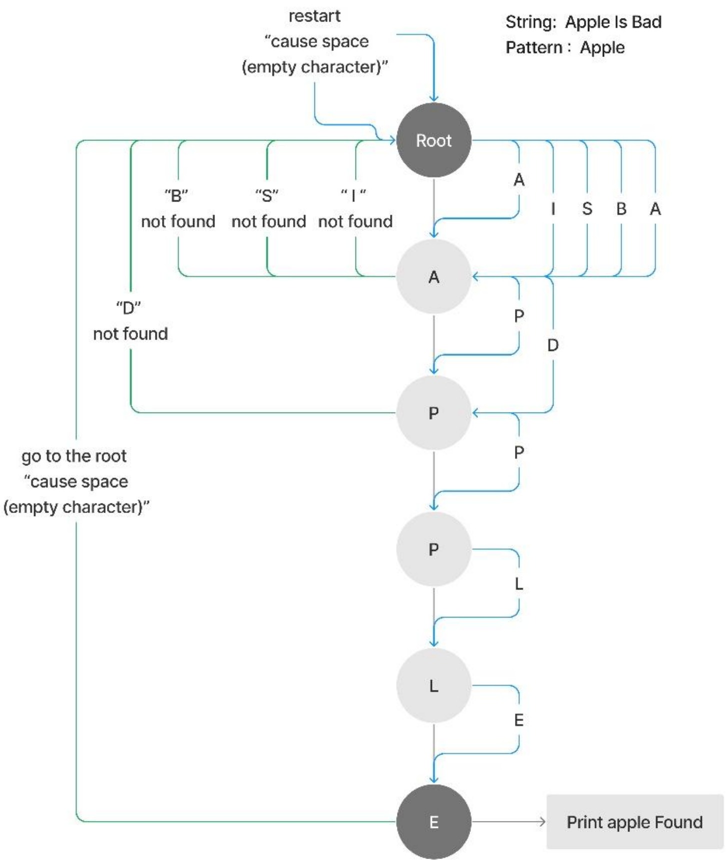


**Fig 4: The Aho-Corasick algorithm**

## 4. Results

The research findings demonstrate the system's seamless ability to detect inappropriate words in both English and Sinhala, even when there are no spaces between the input text. This is made possible by the Aho-Corasick algorithm, which is a highly adaptable and efficient algorithm for content filtering. Also, the Aho-Corasick algorithm uses less memory than other content filtering algorithms, making it memory-efficient. This is especially advantageous for devices or systems with limited resources or memory.

One of the key advantages of the Aho-Corasick algorithm is its ability to handle real-time modifications to the filtering pattern. This means that the system can be easily updated to include new inappropriate words or phrases or to remove words that are no longer considered to be inappropriate.

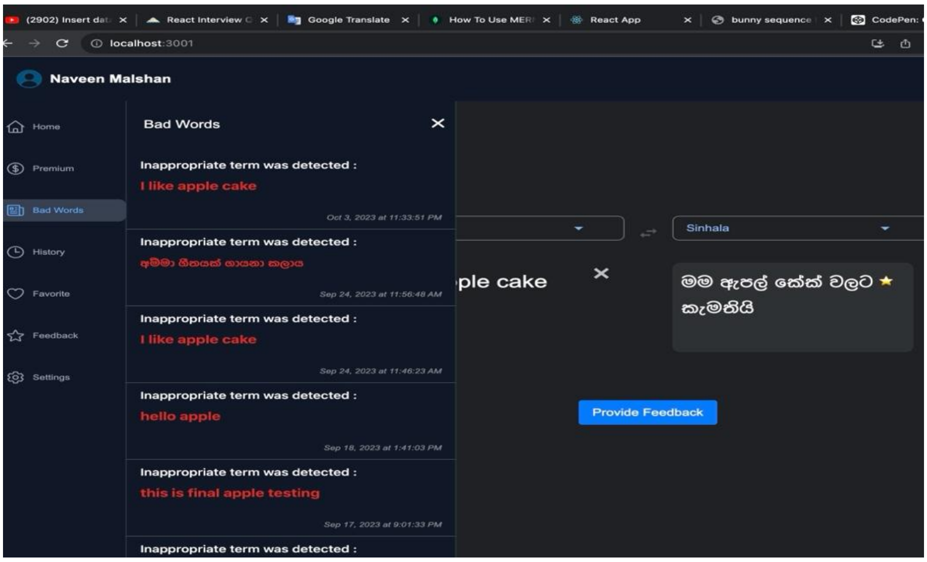
The algorithm can also be readily expanded to incorporate unique word lists or dictionaries. This adaptability enables system administrators or content moderators to customize the content filtering process to meet particular circumstances or requirements, improving its accuracy and relevance.

In addition to its real-time adaptability, the Aho-Corasick algorithm is also very efficient at processing content. This allows the system to perform real-time content analysis and filtering at a high speed. This is important for platforms and applications that can handle large volumes of user-generated content.

The system's ability to detect inappropriate words in both English (Fig5) and Sinhala (Fig6) makes it a valuable tool for a variety of digital platforms, such as social media, networks, forums, and messaging programs. applications. For example, it could be used to filter user-generated content on social media platforms, online forums, and messaging applications. It could also be used to protect children from exposure to inappropriate language in educational settings and online games.

The system for detecting inappropriate words in English and Sinhala text has the potential to improve the safety and security of online communities, protect children from exposure to inappropriate language, reduce the amount of hate speech and cyberbullying online, and improve the quality of user-generated content. The system's adaptability, efficiency, and multilingual capabilities make it a valuable tool for a variety of applications.

The Aho-Corasick algorithm also benefits from ongoing optimizations and improvements because it is widely applied and extensively studied. This guarantees that the content filtering system stays current with the most recent developments in the industry, offering trustworthy and effective filtering capabilities.



Bad Words

Inappropriate term was detected :
I like apple cake
Oct 3, 2023 at 11:33:51 PM

Inappropriate term was detected :
අම්මා ගීතයක් ගායනා කලාය
Sep 24, 2023 at 11:56:48 AM

Inappropriate term was detected :
I like apple cake
Sep 24, 2023 at 11:46:23 AM

Inappropriate term was detected :
hello apple
Sep 18, 2023 at 1:41:03 PM

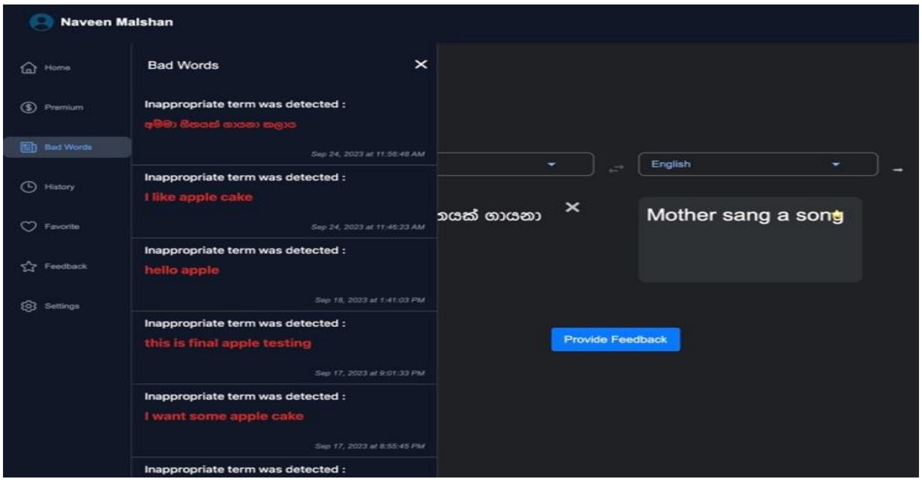
Inappropriate term was detected :
this is final apple testing
Sep 17, 2023 at 9:01:33 PM

Inappropriate term was detected :

ple cake

මම ඇපල් කේක් වලට කැමතියි

**Fig 5: English translation to Sinhala**



Bad Words

Inappropriate term was detected :
අම්මා ගීතයක් ගායනා කලාය
Sep 24, 2023 at 11:56:48 AM

Inappropriate term was detected :
I like apple cake
Sep 24, 2023 at 11:46:23 AM

Inappropriate term was detected :
hello apple
Sep 18, 2023 at 1:41:03 PM

Inappropriate term was detected :
this is final apple testing
Sep 17, 2023 at 9:01:33 PM

Inappropriate term was detected :
I want some apple cake
Sep 17, 2023 at 8:55:45 PM

Inappropriate term was detected :

තයක් ගායනා

Mother sang a song

**Fig 6: Sinhala Translation to English**

## 5. Discussion

This paper is an effort to close the language gap between English and Sinhala, this research article presents a ground-breaking strategy aimed at providing safe and secure language translation, particularly for children. In this system, (Sinhala-English English-Sinhala), is a full-featured set of language processing and translation tools created with the primary objective of facilitating responsible and secure communication between these two languages.

This paper demonstrates the functions of this online application, a Sinhala-to-English and English-to-Sinhala translator, in addition to the research findings. To support effective translations, Rapid API is used by this system to handle translations from English to Sinhala and from Sinhala to English, encompassing both active and passive voices. Additionally, the present, past, and future tenses of this English-to-Sinhala translator, powered by Rapid API's translator API, are accurately translated.

Notably, the authors have added a distinctive feature to the inappropriate language detector. A nasty word library is created utilizing this function using a specified list of offensive terms. An Aho-Corasick algorithm is used by the system to determine whether a word entered by the user is appropriate. When a user under the age of 18 searches for an offensive word, the system adds the word to a list of offensive words and prevents the user from erasing it. Parents can keep an eye on this data to make sure their kids are browsing in a safe and suitable manner.

Furthermore, the authors have added a user feedback feature to improve the Caliber of translations. This gives the system’s users the ability to comment on translations that might need to be improved. With the aid of this iterative feedback loop, the authors are able to improve the precision of translations over time and deliver the best possible language translation experience.

In conclusion, this SEES translating, content filtering system, and the related web application work to promote effective translation across two linguistic domains by addressing the requirement for responsible and secure language use in addition to assisting in the removal of language barriers.

## 6. References

[1] D. P. Wijethunga, A. Nanayakkara and J. Wijayakulasooriya, "Simplified Way of Producing Sinhala Concatenative Text To Speech for Embedded Systems," in *Proceedings of Technical Sessions*, 2013.

[2] D. I. De Silva, P. K. D. A. Alahakoon, P. V. I. Udayangani, D. Kolonnage, M. H. P. Perera, and S. Thelijjagoda. "Application of Transfer based Machine Translations from Sinhala to English," In Proceedings of the 4th SLIIT Research Symposium, vol. 2, pp. 33- 36. 2008.

[3] L. Wijerathna, W.L.S.L. Somaweera, S.L. Kaduruwana, Y.V. Wijesinghe, D. I. De Silva, K. Pulasinghe, S. Thelijjagoda, "A Translator from Sinhala to English and English to Sinhala (SEES)," Proc. International Conference on Advances in ICT for Emerging Regions (ICTer), Colombo, Sri Lanka, December 2012, pp. 14-18.

[4] E. Nicoladis and F. Genesee, "A longitudinal study of pragmatic differentiation in young bilingual children," *Language Learning*, vol. 46, no. 3, pp. 439-464, 1996.

[5] A. Omar, A. Khafaga and . I. E. N. A. W. Shaalan, "The Impact of Translation Software on Improving the," *International Journal of Advanced Computer Science and Applications*, vol. 11, p. 6, 2020.

[6] D. De Silva, A. Alahakoon, I. Udayangani, V. Kumara, D. Kolonnage, H. Perera, S. Thelijjagoda, “Sinhala to English Language Translator," 4th International Conference on Information and Automation for Sustainability, Colombo, Sri Lanka, 2008, pp. 419- 424.

[7] M.Prasad,"Phonaticunicode converter,"2012.[Online].Available:https://unicode.malindaprasad.com/.

[8] i. Google, "Google Input Tool," 2011. [Online]. Available: https://www.google.com/inputtools/services/features/input-method.html. [Accessed 2023].

[9] Translate.com, "Go Global with Translate.com," 2011. [Online]. Available:https://www.translate.com/english-sinhala.

[10] M. Kulatuga, "Madura Online Dictionary," 2012. [Online]. Available: https://www.maduraonline.com/.

[11] A. Chand, "Do you speak English? Language barriers in child protection social work with minority ethnic families.," *British Journal of Social Work,* vol. 35, no. 6, pp. 807-821, 2005.

[12] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic and N. Bhamidipati, "Hate speech detection with comment embeddings," *international conference on world wide web,* pp. 29-30, 2015.

[13] S. Hasib, M. Motwani and A. Saxena, "Importance of aho-corasick string matching algorithm in real world applications," *Journal Of Computer Science And Information Technologies,* vol. 4, pp. 467-469, 2013.

[14] D. Regéciová, D. Kolář and M. Milkovič, "Pattern Matching in YARA: Improved Aho-Corasick Algorithm," *IEEE Access,* vol. 9, pp. 62857-62866, 2021.

[15] E. Perera, "Some unique features of Sinhala Language [Online]," https://www.lankalibrary.com/, 2000.