Chapter 1:

1. Introduction

The use of social media and the internet has increased significantly in Sri Lanka in recent years, following worldwide trends toward increased digital connectivity. The nation's communication and interaction dynamics have experienced an important transformation due to the growing number of people using online platforms. But in addition to this digital transformation, hate speech, hate crimes, and cyberbullying on social media are widespread problems in Sri Lanka and many other countries.

Despite the fact that these instances are extremely widespread, there is a worrying tendency to minimize or ignore the importance of them. While some people consider hate speech and cyberbullying to be just forms of free speech, others blame a lack of regulations or even dismiss them as harmless internet jokes. The truth is far from that, however, those who are subjected to this kind of online abuse frequently experience severe mental suffering as well as negative social consequences such as depression and suicide.

This study aims to address the urgent need for a proactive monitoring system customized for the Sri Lankan setting in light of these critical concerns. Our goal is to give an effective solution to protect the worth and well-being of Sri Lankan internet users by creating a cutting-edge tool that can identify hate speech and cyberbullying on social media platforms. Through this research, we hope to strengthen the resilience and security of people and communities in the realm of the internet and develop an inclusive and respectful online culture.

2. Background of the study

Social media offers fresh platforms for communication, information exchange, and self-expression, and it has quickly become a part of Sri Lankan day to day life. But despite its advantages, abuse has resulted in alarming problems. Misinformation, addiction, cyberbullying, and hate crimes have become important issues that affect both individuals and communities. Gaining an understanding of the motivations behind social media use in Sri Lanka is essential to appreciating its influence on society. For a variety of purposes, such as news access, your business advancement, and connections upkeep people interact with these platforms. Examining these driving forces indicates the ways in which social media shapes public opinion and behavior. Creating a safe environment for all social media users by mitigating the harmful forces that are reaching through social media.

2.1. Growth of the usage of social media.

Sri Lankans have been using social media with increasing frequency in recent years, with platforms such as Facebook, Instagram, WhatsApp, and X growing in popularity. This growth may be related to factors like an increased young population, more accessible internet, and the widespread availability of connection methods.

According to GWI and data.io, there are more than 6.85 million social media users who are above 18 years and 37.2 percent of them are females and 67.8 percent are males.

Social media has become an essential part of daily life, working as a main platform for entertainment, communication, and information exchange.

2.2. Current context of social media-related negativities

Social media-related problems like hate speech, hate crimes, and cyberbullying have become major concerns in the contemporary Sri Lankan setting, having a profound effect on both individuals and communities. Unfortunately, the ease with which information can be posted on social media platforms like Facebook and Twitter has led to the dissemination of harmful content, which has resulted in instances of harassment, threats, and discrimination. Hate crime victims frequently experience severe psychological suffering, anxiety about their safety, and social exclusion. Furthermore, the increase of hate speech and cyberbullying worsens tensions among communities and threatens the harmony of society.

2.3. Current safety measures that have been provided for users of social media platforms

A growing number of Sri Lankans can benefit from the positive effects of enhanced online experience protection provided by the Sri Lankan police force established the Cybercrime Division. In addition, community groups and social activists are essential for supporting victims and encouraging users proper online behavior. By means of awareness programs and advocacy initiatives, they enable people to securely utilize social media and foster an online environment that values inclusion and respect. When combined, these programs offer a thorough strategy for preserving the dignity and general well-being of Sri Lankan social media users.

3. Problem statement

as the major issue related to hate speech is the increasing number of victims. Victims who are not able to defend themselves from their insecurities are more likely to be isolated from society. This leads to escalated psychological issues such as depression and attempts to suicide.

This has become the negative effect of hate speech and in Sri Lanka, most of the hate speech attempts are done on Facebook mostly as a criticism related to a person, organization, or related to a certain community based on their characteristics.

Among the 6.5 million Sri Lankan Facebook users there are hate-spreading individuals and communities, people who follow those individuals based on their emotional experiences or for their satisfaction. Also, there are people who wish to contribute to social well-being and build up a helpful and more harmonized environment on the Facebook platform. Such individuals and or communities have been discouraged because of the hate spreaders.

Hate speech has been discussed and controlled or tried to control over other countries and This study identifies the problem and the necessity of hate speech as serious as life-threatening which must be addressed and controlled over hate speech on Facebook detecting and eliminating such content.

4. Research questions

01. What is the difference between hate speech and free speech?

02. What are methods that hate speech spread over Facebook?

03. What are the negative effects of hate speech?

04. How are cyberbullying and hate crimes related to each other?

05. What are the positive impacts of filtering hate speech?

06. Why do people use hate speech? And what are their motives?

07. What are the current hate speech detection methodologies?

08. Will hate speech detection affect to the freedom of speech?

5. Objectives of the study

01. To study the Sri Lankan Facebook usage and socio-impacts.

6.55 million users have been recorded for 2023 January. The community of Facebook has spread all across the island and it has increased drastically in recent years. This has affected both positively and negatively the society and the person's life. Facebook has become a threat to individuals' lives because it has become an addiction where people have lost their minds and given up on their work. Some have been isolated and become cyber victims through the fraud, harassment, and online crimes that have taken place through Facebook.

Facebook is a platform that we can utilize to build a productive environment where people can use it to make their lives easier by creating online helping communities, helping and protecting the users, also to spread news and updating certain situations in real-time, business creation, and as a marketing platform, etc. Facebook has its own both negative and positive impacts based on the user and their community.

02. To study about are the hate crimes, hate speech, and cybercrimes that occur because of Facebook.

Facebook's extensive user base and significant influence over online discourse have led to several instances of hate crimes, hate speech, and cybercrimes worldwide. For example, the platform has come under fire for aiding in the spreading of hate speech that calls for the violence against marginalized groups. Facebook has also been used to plan and carry out hate crimes, including the live-streaming of violent assaults. Additionally, phishing schemes and identity theft have become increasingly common on the network, taking advantage of users' weaknesses with regard to their personal data.

03. To study mitigation techniques and regulation steps that are taken to control hate speech in the world.

The negative consequences of hate speech have been restrained globally via the use of different mitigating strategies and legal procedures in reaction to its spread online. Using artificial intelligence systems and content moderation algorithms is one popular method for quickly identifying and eliminating hate speech. For instance, abusive content is now automatically detected and removed by Twitter and YouTube using pre-established criteria. The identified content is then reviewed by human moderators. In addition, a number of nations have proposed laws to hold online platforms responsible for allowing hate speech. As an example, the Digital Services Act proposed by the European Union imposes strict guidelines mandating that internet companies promptly delete any unlawful information, including hate speech, or risk paying severe penalties. Furthermore, the goal of awareness campaigns and educational programs is to provide users with the knowledge and skills necessary to identify and properly report hate speech, promoting an attitude of good citizenship and responsible online conduct. Through a blend of technological, legal, and instructional approaches, global stakeholders strive to establish online spaces that are safer, more welcoming, and devoid of hate speech's deleterious impacts.

04. To study the Sri Lankan context of hate speech crimes and their controllability of it.

Hate speech has always been a problem in Sri Lanka, frequently increasing tensions between different ethnic and religious groups. Studies (Samaratunge and Hattotuwa, 2014) demonstrate how common hate speech is on social media, especially when it comes to targeting minorities. Also, Sri Lankan police have made a separate division named the Cybercrime Division to address cyber-related issues. Different mitigating techniques have been put into place to address this problem. The International Covenant on Civil and Political Rights Act, for example, was passed by the Sri Lankan government and makes hate speech and incitement to violence illegal (*Hate speech and Hate Crimes*, 2023) Social media companies have also implemented content moderation guidelines and hate speech detection and removal capabilities. However, because of Sri Lanka's complicated sociopolitical environment, difficulties continue to arise in properly implementing these policies.

05. To implement a tool for detecting Facebook hate speech.

Developing a Facebook hate speech detection tool is crucial for creating a more respectful and secure online community. Because of the platform's extensive reach and power, the tool can quickly detect and delete offensive content, shielding users from the negative repercussions of hate speech. The tool contributes to the development of a more welcoming online community where people can express themselves without worrying about harassment or discrimination by encouraging a culture of respect and tolerance. In the end, purchasing such technology shows Facebook's dedication to maintaining user security and welfare on its network.

06. To study the pros and cons of the tool.

Creating a hate speech detection tool for Facebook has benefits and drawbacks. Positively, by quickly detecting and eliminating harmful content, this technology could significantly enhance user safety and promote a more welcoming and joyful online community. Furthermore, it would show a dedication to maintaining platform integrity and community standards, earning the trust of users and stakeholders.

But it's important to take into account any possible downsides. Accuracy issues with automated detection algorithms might result in excessive censorship and the suppression of legitimate speech. Furthermore, creating and maintaining such a tool is heavy on resources and presents difficult ethical dilemmas relating to verbal freedom and censorship. Despite these difficulties, platforms looking to encourage safety and civility online should consider the potential rewards of putting in place a hate speech detection technology.

6. Summary of limitations

Social media are acting differently in many people's lives. Some kept them private and exposed the life of social media only to a certain set of people to that individual. Some kept it public where anyone can reach out and interact with their profiles and content. But not everyone is interested in sharing their experience throughout the journey of social media.

Therefore collecting Data is very hard from the people to satisfy the content in the study. Since there is a large number of users on Facebook, the identified pool is very large and it is very hard to select the perfect sample to continue the study. People are less likely to fill out a questionnaire and engage with the study continuously. Being ready to help individuals are rare with the current lifestyle. Since the study is focusing on the Sri Lankan community much of the data are not available for training a model. Also, Singlish is a user created language and there is no proper language guidelines to follow to recreate data.

Above mentioned limitations are the major concerns that the author focuses on while conducting the study.

Chapter 2:

1. **Literature review**

2. **Conceptual framework**

1. Literature review

Under this chapter, provides the definition of hate speech and how it differentiates from free speech and the previous methods and studies conducted on a similar domain to investigate the gap of the current study to approach a system-based module or a solution to bridge the gap between the existing and the current context of the study.

Definition of hate speech

Hate speech cannot be clearly defined due to its variance of usage among the platforms, users, and contexts(Howard, 2019). Hate speech is considered as defaming, harming, or causing a threat or harassment on an individual or for a group of individuals based on characteristics such as religion, gender, race, disabilities, nationalities, or wealth (Tontodimamma et al., 2021).

MacAvaney (MacAvaney et al., 2019) mentioned about four types of definitions for hate speech,

1. Hate speech is to incite violence or hate
2. Hate speech is to attack or diminish
3. Hate speech has specific targets
4. Whether humor can be considered hate speech

Definition of free speech

Freedom of expression, freedom of speech, and many other terms are used to identify the term free speech. Free speech has become a fundamental right from human rights (Universal Declaration of Human Rights, no date). The right has been granted from the 19th of the International Covenant on Civil and Political Rights, adopted in 1966 (Freedom of Opinion and Expression, 2024).

Free speech is invoked as the communication and expression exchanged with other parties with the moral of communicating (Howard, 2019).

Hate speech has been covered by the freedom of expression most of the time since it has been amended to the country's jurisdictions (Fino, 2020). Since hate speech has no legal definition this has been neglected and this has been taken into consideration with the escalation of the number of cases and victims by the hate speech and for the identification of such instances

FBI named these as hate crimes for further investigations (Hate speech and Hate Crimes, 2023).

In Sri Lanka, hate speech has been taken under consideration and added that as a challenge in the digital age due to the increasing number of social media users including Facebook. According to the Human Security Handbook 2016, promotion for human security which was addressed in 2012 General Assembly 66/290 the basic right of an individual is to live in freedom and dignity without being subjected to poverty and despair. This applies to all the communities and individuals with equal rights as humans (Patabendige, 2023).

Related work and Identification of the gap

Previously conducted study related to hate speech (Brown, 2018) identified that the hate speech can be in both physical and online methods. Since the online hate speech identification has become a worldwide issue, researchers have conducted many studies on detection and mitigation methods. Tontodimamma (Tontodimamma et al., 2021) has taken the thirty years of details related to hate speech and created a study of the yearly hate speech related publication and created an analytical report distributed over the years which signifies the importance of hate speech detection.

The most common and widely spread method is the keyword approach where it uses terms from an ontology or dictionary to identify the potential keywords of hate speech related content (MacAvaney et al., 2019). Counter messaging is also an approach to address the individuals or the accounts that are directly or indirectly spreading hate speech among those platforms (Samaratunge and Hattotuwa, 2014; Hattotuwa and Wickremesinhe, 2023). Deep learning has been considered as the more prominent ML technology where it can be trained itself to achieve the specified goals. Such technologies have been used to identify inflammatory language and hate speech by using four different deep-learning models (Gaurav et al., 2023). Neural network is another prominent technology that has been used to detect hate speech in comparative platforms (Pereira-Kohatsu et al., 2019). These technologies have laid the base ideation of creating a hate speech detection tool that is contextually related to Sri Lanka. According to the Sri Lankan stats (Social media stats Sri Lanka, 2024) of social media Facebook is the widespread and most dominating platform compared to other platforms. Due to the larger number of users, the number of hate speech cases and hate crimes have been reported on Facebook. The loophole with the Facebook community standards and the hate speech is the language most people use on Facebook is not only English and Sinhala. The user created language known as Singlish, which uses English letters to pronounce or write the Sinhala terms that basically can be identified as a combination of Sinhala and English languages.

Using ML-based technology to counter the hate speech which is spread using Singlish and English on Facebook will help to decrease the number of hate speech cases and hate crime incidents.

2. Conceptual framework

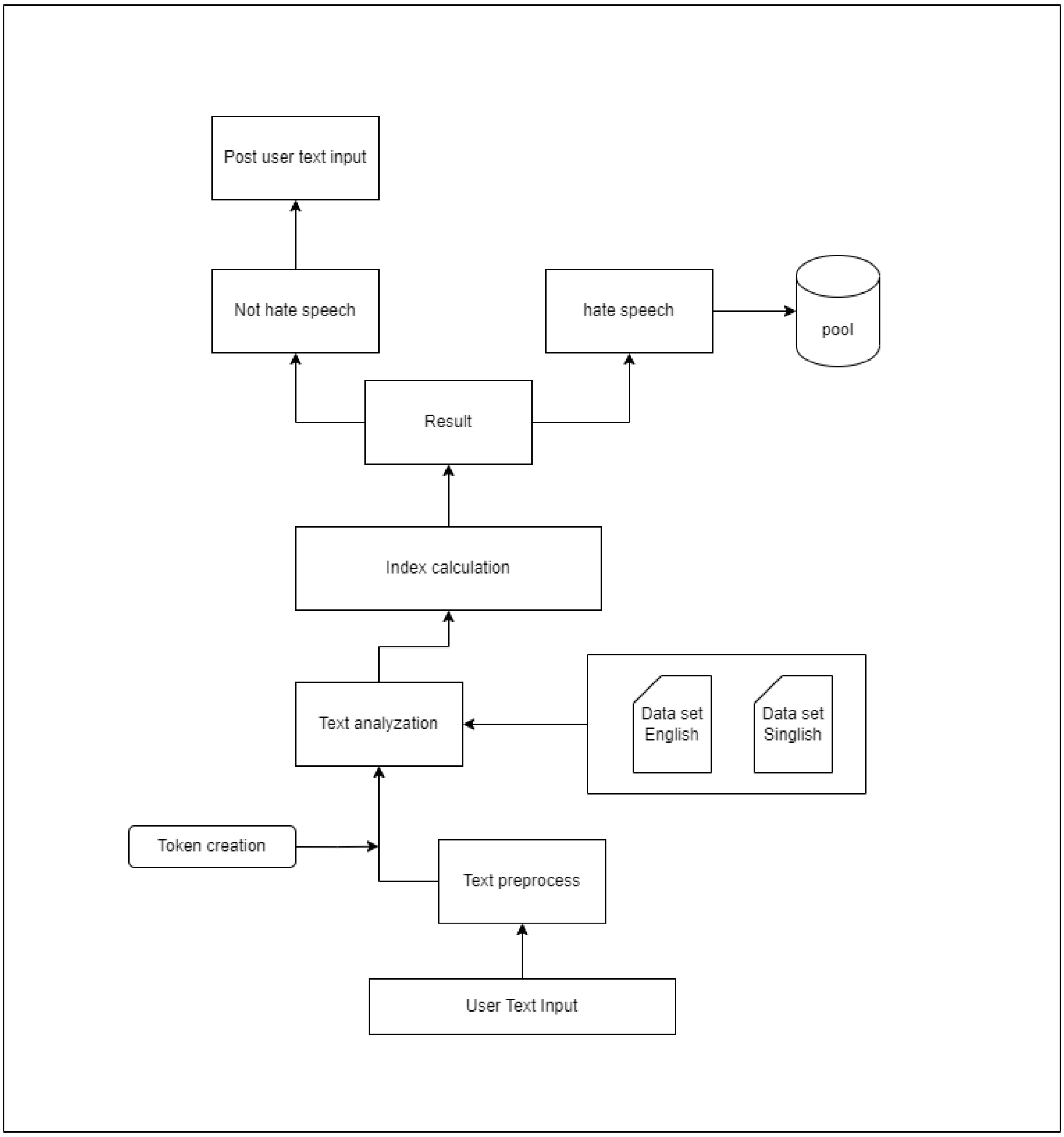

Figure 1 : Conceptual framework

Figure 1. Depicts the methodology of the hate speech detection tool with its origination to the very end till either posting the text input or pooling it without posting it on Facebook. Each stepper includes one or more processes, and each stepper needs to be completed before starting the next stepper. The user input text will pre-processed and will create a token as reference to the pre-processed text and analysed against the dataset to identify whether the user input text contains any hateful content. Such content will eliminated and will pool in a DB with the details of the user and the content.