A logo with text and a globe

Description automatically generated with medium confidence

NSBM Green University

Faculty of Computing

Management Information Systems

Final Year Project Report

**Student Name:** B.A.A.V Karunathilake

Student ID: 20895

**Batch:** 20.3 – UGC MIS

Research topic

Detecting and countering hate speech in modernized language/bilingual language:

**Machine learning-based model to counter hate speech in Singlish on Facebook to ensure user safety and platform safety**.

Research Proposal Conducted by: B.A.A.V Karunathilake | An undergraduate: NSBM Green University | baavkarunathilake@students.nsbm.ac.lk | +94 77 159 2345 | Conducted in February 2024 | The domain of the Study: Hate speech in Sri Lanka

Contents

[Chapter 1 6](#_Toc179912419)

[1.1 Introduction 6](#_Toc179912420)

[1.2 Relevancy of The Topic 6](#_Toc179912421)

[1.2 Background of the Study 8](#_Toc179912422)

[1.2.1 Growth of the usage of social media 8](#_Toc179912423)

[1.2.2 Current context of social media-related negativities 8](#_Toc179912424)

[1.2.3 Current barriers that have provided for the safety of users of social media 9](#_Toc179912425)

[1.3 Problem Statement 9](#_Toc179912426)

[1.4 Research Question 10](#_Toc179912427)

[1.5 Motivation 10](#_Toc179912428)

[1.6 Aim 11](#_Toc179912429)

[1.7 Objectives of Study 11](#_Toc179912430)

[1.8 Challenges and Limitations 14](#_Toc179912431)

[Chapter 2 16](#_Toc179912432)

[2.1 Literature Review 16](#_Toc179912433)

[2.1.1 Definition of Hate Speech 16](#_Toc179912434)

[2.1.2 Definition of Free Speech 17](#_Toc179912435)

[2.1.3 Hate Speech vs Free Speech 17](#_Toc179912436)

[2.1.4 Hate Speech in Social Media 17](#_Toc179912437)

[2.1.5 Hate Speech Detection 18](#_Toc179912438)

[2.1.6 Bilingual and Multilingual Language Hate Speech 18](#_Toc179912439)

[2.2 Previous Studies related to the topic 19](#_Toc179912440)

[2.3 Identification of the Gap 21](#_Toc179912441)

[2.4 Current context of the subject 22](#_Toc179912442)

[2.5 Significance of the Study 23](#_Toc179912443)

[2.6 Reflection 24](#_Toc179912444)

[Chapter 3 26](#_Toc179912445)

[References 27](#_Toc179912446)

**Abstract**

In Sri Lanka, the Internet and social media are platforms that are widely open to almost everyone. Usage of Internet users is caped 14.6 million and social media is increasing day by day and up to date it is 7.2 million. People write, post, comment, and share their thoughts on these platforms, which is considered freedom of speaking of humans. Free speech has opened doors for everyone to speak up and react. Hate speech, cyberbullying, and online harassment have taken place due to the freedom of speech. Darkside of freedom of expression has led to threaten, abuse, harass, offend, and defaming individuals or entities. This study addresses the negative impacts that hate speech and hate crimes have on Sri Lankans. We seek to understand the emotional, social, and psychological impact these incidents exact on individuals and communities by looking at real-life experiences and perspectives.

The study also emphasizes how important it is to have efficient hate speech detection technologies on social media platforms. Such types of solutions are essential for preventing hate speech from spreading further since internet platforms are becoming more and more like breeding grounds for it. These technologies can help mitigate the negative effects of hate speech by promptly identifying and eliminating it, providing a more comfortable and secure environment on the internet.

In this study, I want to draw attention to the critical need for proactive steps toward preventing hate speech and hate crimes in Sri Lanka, highlighting the vital role that technology plays in preserving social harmony and well-being.

**KEYWORDS**

Hate Speech, Hate crimes, Natural Language Processing

# Chapter 1

# 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.

# 1.2 Relevancy of The Topic

Multiple researches have been conducted over the past years on detecting hate speech on several social media platforms. Identifying hate speech was considered a critical role in shaping public discourse. Millions of users from various backgrounds, multiple religions, and multiple races are using Facebook and the content they engage in is more likely they reflect societal sentiments, including hate speech.

Previous studies have paved the path to identify and eliminate hate speech and ensure the flat form safety and user safety. Also due to previous studies, researchers were able to discover many identification models, mechanisms and methodologies including creating recent data sets for future studies, getting into machine learning, linguistical understanding, and the continuous improvements of identification and detecting algorithms.

By following the traditional machine learning approaches such as logistic regression and SVM are the more frequently used and with the increase of the size and the complexity of the usage, and the data set. Researchers have discovered that deep learning approaches such as LSTM and BERT will be more effective on handling the above-mentioned challenges.

In a country like Sri Lanka, there are multiple languages that people use such as Sinhala(mother language), English, and Tamil as the formal and general conversation language or as the written mode. In recent years people have discovered a communication linguistic mode known to be ‘Singlish’ which is a mix of Sinhala and English languages and this has widespread people use this language for their everyday conversations. This has not been identified as a formal language. Because of the widespread usage of this language, the importance of multi-lingual hate speech identification has arisen and identifying and ensuring user safety and also identifying native hate speech has always been a threat to all social media users.

The relevance of the current study is to identify the hate speech from free speech to the discourse of the subject and eliminate the hate speech which is discouraging content or defaming content posted to social media with the objective of ensuring user safety and platform safety.

The study extends to using the multilingual approach to identify Singlish which is a challenge posed by code-switching, where people use mixed languages. Previous studies have laid the groundwork for hate speech detection and the current study aims to fill the gap in the context of bilingual or multilingual which will lead to novel advances in the domain, which are more important to regions like Sri Lanka.

# 1.2 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.

## 1.2.1 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.

## 1.2.2 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 in hate speech and cyberbullying worsens tensions among communities and threatens the harmony of society.

## 1.2.3 Current barriers that have provided for the safety of users of social media

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 by the Cybercrime Division. In addition, community groups and social activists are essential for supporting victims and encouraging users to engage in proper online behavior. Awareness programs and advocacy initiatives 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.

# 1.3 Problem Statement

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.

# 1.4 Research Question

How to validate hate speech over freedom of expression and build a detection model to identify hate speech which are generated using Singlish language which is a bilingual variation of Sinhala and English, to increase the platform safety of Facebook and improve the safety of the users of Facebook by creating a hate-free environment which has not been covered yet under the user protection layer of the Facebook.

1. What is the difference between hate speech and free speech?
2. Will hate speech detection affect the freedom of speech?
3. What are methods that hate speech spread over Facebook?
4. How are cyberbullying and hate crimes related to each other?
5. What are the current hate speech detection methodologies?

# 1.5 Motivation

As per the description in 1.2, social media is accessed by millions of people and there is a clear requirement to ensure the safety of the users as well as the platform. Social media has played a major role in contributing to society in both positive and negative ways, it is a must to have an effective social media platform and social media activists to make social media more effective for society.

Hence, the current study is focused on ensuring Facebook user safety and platform safety against hate speech that occurs in the Singlish language.

Therefore, the motivation of this study is to,

Improve the Facebook user safety, providing and pro-active safety measure to encounter and eliminate hate speech without posting. And also to pave a path to eliminate hate speech that has been created using bilingual and multilingual languages.

# 1.6 Aim

The current research project’s aim is to identify the gap between hate speech detection and provide a solution to bridge the gap between the current context and ensure Facebook user safety and platform safety

# 1.7 Objectives of Study

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

6.55 million users have been recorded for January 2023. 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.

1. **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 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.

1. **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.

1. **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 demonstrate how common hate speech is on social media, especially when it comes to targeting minorities (Samaratunge and Hattotuwa, 2014). 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 Social media companies have also implemented content moderation guidelines and hate speech detection and removal capabilities (*Hate speech and Hate Crimes*, 2023). However, because of Sri Lanka's complicated sociopolitical environment, difficulties continue to arise in properly implementing these policies.

1. **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.

1. **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.

1. **To Create a proper dataset for Singlish hate speech detection,**

Creating a proper ultimate model requires a large number of data to train itself and also to test its efficiency on how it performs. Data set plays a vital role in the machine learning model creation and also to increase the efficiency and the effectiveness of the existing models.

Will create a hate speech-related Singlish data set and researchers who are willing to follow the current domain and for cross-domain research purposes will be able to use this data set.

1. **To create a System that can be used for multilingual hate speech detection purposes**

As per the description of 05, the Tool will enable to identification of Singlish and eliminate hate speech.

This will enable to identification of multilingual and bilingual hate speech by changing parameters and providing necessary datasets to train and test. Therefore this can be used for any similar context that is discussed in the study.

Ultimately this will be developed to identify and eliminate the Singlish hate speech on Facebook.

# 1.8 Challenges and Limitations

* A limited time frame due to the external dependencies will negatively impact the whole research process including making the suggested system and the report that has been created using the findings.
* As a fresh undergraduate and has not been exposed into the process of creating a research or a research project. Therefore having a lack of experience will also be a disadvantage and will take time to figure out certain steps of the research process.
* Lack of experience will cause for many issues,
  + Choosing a wide scope of research
  + Choosing the wrong methodology
  + Finding and spending time on creating inappropriate data sets.

And many more.

* Facebook is a personal profile and therefore collecting data and their user experience will be confidential and hard to extract.
* Finding individuals who are supportive of providing data and information for such studies
* Fewer data sets are available to refer.
* Fewer similar studies that have been conducted
* To find out what is the most effective data modelling method, and what are the more effective methods to follow due to the lack of knowledge.

# Chapter 2

# 2.1 Literature Review

Under this chapter, this provides the definition of hate speech and how it differentiates from free speech by using the existing studies and reports that are publicly available. The importance of the definition of both the terms is required to be clarified to proceed with the study.

Hate speech has a wide scope, where it has appeared in the real world as well as the digital world. Therefore, identifying hateful content and filtering hate from social media platforms has been a critical consideration when it comes to platform security. To examine the existing models and methods to bridge the gap between hate speech detection and hate speech spreading using multilingual language/s for the purpose of getting existing findings and available resources to the current context of the study to ensure the safety of the platform and its users.

## 2.1.1 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), (Mondal, Silva and Benevenuto, 2017).

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 humour can be considered as hate speech

## 2.1.2 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 of humans and is written as a human right (*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).

## 2.1.3 Hate Speech vs Free Speech

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) (Mondal, Silva and Benevenuto, 2017). 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, the Federal Bureau of Investigation named this as hate crimes for further investigations (*Hate speech and Hate Crimes*, 2023).

In Sri Lanka, hate speech has been taken under consideration and added 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 the 2012 General Assembly 66/290 the basic right of an individual is to live in freedom and dignity without getting subjected to poverty and despair. This applies to all communities and individuals with equal rights as humans (Patabendige, 2023).

## 2.1.4 Hate Speech in Social Media

The internet has been the tool that has made globalization possible and which has given access to the world, ever since hate speech has been planted by certain users and created internet as a tool to defame and spread hate against communities and groups of people. By the end of 2000, there were 17.1 million websites and 1.1 billion domains were recorded up to date (DigitalSilk, 2024). In 2004 there was an attempt record to identify web pages that contained hateful content, Spreading racism, and extremism.

A Study that was conducted in 2015 incorporated with UNESCO (Iginio Gagliardone, Danit Gal, Thiago Alves, Gabriela Martinez, 2016) has shown that this has been a growing problem within digital communities and their users and platforms like Facebook are not reactive to hate speech unless their users have reported it. There was no other option provided to the users to ensure their safety.

Hate speech can be identified directly and indirectly where a person could post against another person, or to a community and there could be a third party who is not a part of the conversation but has encountered with the hate content which is known to be hate spreading (Jain and Sharma, 2022).

## 2.1.5 Hate Speech Detection

Over the period number of hate speech and hate content posts has increased drastically. Gender-based (male and female), transgender communities, minorities, and religious-based debates are known to be the root cause for the nowadays hate spreading content.

Due to the rapidly increasing instances of hate crime cases, people have conducted many studies over the period by testing and getting the newest technologies and methods to capture hate speech. Studies are conducted based on manual keyword picking to use machine learning to identify hate speech (Kavatagi and Rachh, 2021; Jain and Sharma, 2022).

The complexity of the domain has increased due to the information accessibility and technological literacy of the people. There fore to detect hate speech researchers had to build their own strategies based on the contextual perspective of the approach (Castaño-Pulgarín *et al.*, 2021). Text based hate speech has no emotions or expressions therefore understanding the meaning and the context is considered as a challenge.

## 2.1.6 Bilingual and Multilingual Language Hate Speech

Hate speech detection on social media platforms like Facebook has become a widely spreading research domain with the identification as a threat to humanity and a concern that affects human life. Each study encounters issues and provides a solution to mitigate each. With new trends and with human nature they find a way to counter each solution that the researchers have suggested in their studies.

Multilingual and Bilingual communication is a community based language model that people have created over time to create an effective communication method. Singlish is a language that the people living in Sri Lanka are currently using. Multilingual and bilingual languages are used to spread hate due to they are novel and identifying such languages with mixed characteristics is not possible as same as standard languages such as English, French, Sinhala, etc. Many studies have been conducted to capitalize on the issue of spreading hate speech in bilingual or multilingual languages, each study has identified single or more approaches to address this issue.

# 2.2 Previous Studies related to the topic

Multilingual and bilingual hate speech detection provides a broader scope due to the usage of the number of languages. This widens the scope and increases the complexity depending on the languages and the number of languages. Hate speech detection in a multilingual setting often requires a combination of machine learning, linguistic analysis, and cultural context. Previous studies related to the subject topic that have specifically targeted complexity, cultural nuances, and code switching between languages are expressed

The characteristic of multilingual and bi-lingual languages is there are no certain boundaries or limits to control over their user abilities for those languages have a broader scope in changing their state and code switching between languages. The “meta learning” technique is an approach where its users are able to use it to build a rapid solution with a limited number of data (Prasad *et al.*, 2023). The author has conducted a test and has gained positive results using Bengali and English as the reference languages.

Code mixed languages that are frequently used in such scenarios. Identifying linguistic characters in multilingual and bilingual contexts requires a large number of data. Connecting two models and bridging their different capabilities on lexicons and pre-trainability is another successful approach to identifying hate speech (Pamungkas *et al.*, 2022; Prasad *et al.*, 2023). Pre-trained models are very much effective in capitalizing on the challenges that occur due to code switching and code-mixed language linguistical characteristics.

Novel approaches such as deep learning techniques such as neural network are used to increase the performance and the accuracy of detecting hate speech detection in code mixed content. HSDH is a model that was built to identify hate speech using ‘Hinglish’ which is a mix of Hindi and English languages. A very large number of data was available when making the model which was an advantage when conducting the study (Kumar Kaliyar *et al.*, 2023). Using the deep learning approach has strengthened the study.

A similar Study has been conducted to create a connection and integrate the linguistic characters of languages (Hashmi *et al.*, 2024). A thorough study was conducted on each language to identify linguistic insights in each language and applying them to a detection model will enable to understanding the relationships between the languages. This will increase the performance of the system and its accuracy due to the identification of the patterns of using hate speech in each language or getting cues on the correlation between the two languages.

The resource requirement of such systems varies from one another. The resource requirement depends on the complexity of the system, complexity and interdependencies between data, number of data, number of technologies/ algorithms/ models, etc. Therefore handling such scenarios in a sustainable method is another approach to countering hate speech without creating and regression on other stakeholders/ factors. MLHS-CGCapNet is a lightweight model that is able to handle identify hate speech complex scenarios related to hate speech detection (Kousar *et al.*, 2024). Remarkably this handles multilingual environments without keeping or acquiring large computational resources.

Transformer based models are another successful approach and models like BERT are an advanced approach compared to other studies. The author has been able to enhance and leverage the BERT’s deep contextual understanding and detection capabilities in code mixed settings. Bert has to be pre-trained against the data set in order to handle the complexities of code mixing in languages when detecting hate speech (Deepasree Varma *et al.*, 2022).

# 2.3 Identification of the Gap

A previously conducted study related to hate speech (Brown, 2018) identified that hate speech can be in both physical and online methods. Since online hate speech identification has become a worldwide issue, researchers have conducted many studies on detection and mitigation methods. Tontodimamma (Tontodimamma *et al.*, 2021) took 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 by distributed over the years which signifies the importance of hate speech detection.

The most common and widely spread method is the keyword approach which users 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 platform (Pereira-Kohatsu *et al.*, 2019). These technologies have laid the base ideation of creating a hate speech detection tool which 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, a 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 Sinhala terms that basically can be identified as a combination of Sinhala and English languages.

Using machine learning 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.

# 2.4 Current context of the subject

# 2.5 Significance of the Study

The study addresses the issue of safety among Sri Lankan Facebook users. And also to demotivate and discourage the people who tend to commit those hate crimes. Since Facebook is a widespread platform where over 6.5 million users were recorded and because of a certain set of people their social well-being has become doubtful due to hate crimes.

Ensure the safety of Facebook users – Cyberbullying and hate speech have been a critical issue and have created a major impact on the users. Due to this issue, users have neglected to use the platform and omit the threats, and harassment that are coming through Facebook. The study has identified the many reasons and provided a solution to enhance and ensure the safety of Sri Lankan Facebook users.

Detect and identify several methods of users getting offended on Facebook – There are several ways of communicating and exchanging information on Facebook. Direct chats, Content sharing, Feed uploads, groups, Facebook pages, and many more. People do send text messages, Post content, and share content with or without captions and people do comment a lot on their content and as well as the other peoples’ content. The study helps to identify the text-based content that is added to Facebook as a status or comment using the suggested tool.

Identify the hate speakers – With the pooling system, we are able to store data about hate speakers and will be able to identify their behavior on social media and track their comments, and if necessary to ground certain users also from Facebook itself can trigger an alert on such users get the necessary actions regarding them.

Minimize the rate of hate speech-related issues - Due to the large number of users on Facebook, number of crimes and issues are also high comparatively. Therefore, by implementing the suggested number of crimes will be decreased and the user safety ratio will increase.

# 2.6 Reflection

In this study and with the provided content I as the author, have come to an understanding of the critical importance of defining and differentiating terms like hate speech, free speech, and the respective relationship between one another. These terms are not only central to the broader conversation surrounding online communication but also vital in understanding the role they play on social media platforms like Facebook, where much of this hate speech occurs using Singlish. This understanding is particularly relevant in Sri Lanka.

Hate speech is generally understood as any form of expression that incites violence, discrimination, or hatred against a group or individual based on attributes like race, ethnicity, religion, or gender. It is harmful because it creates divisions within society, often escalating conflicts and marginalizing vulnerable communities. This makes the detection of hate speech crucial, especially on social media platforms where it can spread quickly and reach large audiences. On the other hand, free speech refers to the right of individuals to express their opinions without censorship or punishment, even if those opinions are unpopular or offensive to some. Free speech is a cornerstone of democratic societies and is protected by laws in many countries, allowing people to openly debate ideas and express their thoughts. However, it has limits, especially when it crosses into hate speech, which causes harm and threatens public safety. The challenge is finding a balance between protecting free speech and preventing the spread of hate.

In exploring the literature, it became clear that understanding the line between hate speech and free speech is essential. While everyone has the right to express their views, this right does not extend to speech that encourages harm or violence. The relationship between hate speech and free speech is a topic of ongoing debate, especially on social media platforms where millions of users post content daily. Platforms like Facebook have policies in place to remove hate speech, but these policies are not always perfect. The vast amount of content and the speed at which it is shared make it difficult to monitor and moderate effectively.

Through my review of studies on multilingual and bilingual hate speech detection, I have learned that identifying hate speech in code-mixed languages like Singlish requires more than just applying existing models designed for standard languages like English. It requires a deep understanding of the language's variations and the context in which it is used. The flexibility of Singlish allows users to express themselves in ways that are not easily captured by traditional models, which is why the development of language-specific datasets is so important.

In a multilingual context, it is required to identify each of the language characteristics and their behavior. Such as context, environment, cultural effect, etc.

However, the use of Singlish presents a unique challenge for hate speech detection. Due to its flexibility and lack of formal structure, Singlish can vary greatly in how it is written, making it difficult for automated systems to detect hate speech. Additionally, because it is a relatively informal and community-created language, there are very few datasets available for Singlish, which makes training machine learning models for hate speech detection even more difficult. The lack of comprehensive Singlish datasets makes it difficult to build accurate hate speech detection models, which means that much of the hate speech in this language may go unnoticed or unaddressed. This is particularly concerning in Sri Lanka, where Singlish is commonly used on platforms like Facebook to spread hate speech. As Singlish continues to evolve, the need for more sophisticated models that can detect hate speech in this unique language is becoming increasingly clear.

Various approaches have been followed to address the issue of hate speech, especially in multilingual or code-mixed environments. Multiple studies and their respective researchers have experimented with different models and techniques to improve the accuracy of detection systems while overcoming the challenges posed by linguistic diversity and other informal factors.

The HSDH, for instance, focuses on hate speech detection in code-mixed data, particularly Hinglish (Hindi and English). This research uses deep neural networks to better capture the nuanced patterns in such contexts. The strength of this approach lies in its ability to identify hate speech within sentences that combine words from multiple languages.

MLHS-CGCapNet, introduces a lightweight model designed for multilingual hate speech detection. This model focuses on reducing computational complexity while maintaining accuracy across multiple languages. The importance of this approach is that it enables the detection of hate speech on resource-constrained devices. Given the widespread use of mobile devices in Sri Lanka, an approach like MLHS-CGCapNet is highly relevant. Its lightweight nature makes it more practical for real-time hate speech detection on platforms like Facebook.

BERT has been highly effective in detecting hate speech due to its deep contextual understanding of language, capturing both the meaning and sentiment of words in different contexts. In multilingual or code-mixed environments, pre-trained models are particularly useful because they can leverage knowledge from large amounts of previously learned multilingual data.

[My comment]

# Chapter 3 - Methodology

## 3.1 Chapter Overview

Chapter methodology covers the entire suggested solution and the implementation of the solution. It provides the necessary justifications on how the provided solution will address the subject problem of detecting hate speech in Singlish on Facebook to ensure user safety and platform safety.

Furthermore, with the comprehensive literature review that has been conducted, the Implementation approach for the solution has changed, and the implementation of the solution has been conducted in stages wise.

### 3.1.1 Objectives

5 main objectives have been placed to conduct this chapter. These objectives overlay when creating the methodology and selecting the appropriate frameworks that help to build an effective model to detect hate speech in Singlish.

1. **Create an accurate data collection/data set for hate speech detection in bi-lingual language Singlish**

To make an effective model data is required for any further workings. English and Sinhala as separate languages, have multiple sources to get data sets for hate speech. As per the subject of the study, Singlish is a bilingual language format that is written in English letters to provide the Sinhala meaning terms and as a new-found language, it is lacking with the datasets to work with. To get along with the study it is required to have a sufficient dataset including non-hate and hate speech content to lay the foundation for this study and also for future research in the same domain.

1. **Discover data preprocessing techniques**

Data pre-processing is the major process of data classification and language-based detection systems. The pre processing identifies the language, text normalization, and tokenization of the text input.

* Language Identification – identifies the language of the text input.
* Text normalization- removes the noises, punctuations and converts to a uniform lowercase term and standardizes the given text input
* Tokenization – removes the stop words such as “the”, “a”, “is”, and “am”, etc. And creates individual tokens for each and every word in the given statement.
* Vectorizing – which makes the tokenized words into machine-understandable language for machine learning purposes.

1. **Compare and contrast the existing machine learning models to optimize the gap between hate speech detection systems**

The previous chapter (Chapter 2) and conducting the pre-study related to hate speech detection have given the most efficient and effective solutions and findings relevant to the subject. By inheriting and getting inspiration from the existing studies and their discoveries helps to develop a better solution to identify the potential threats and risks along with the opportunities that create an advantage to make a better, optimized, effective, and efficient bilingual hate speech detection system.

1. **Identify limitations and strengths to fine-tune the language model**

Using traditional techniques/algorithms to process with novel techniques such as deep learning approaches are able to evaluate one another to find strengths and weaknesses in between.

Inheriting the strengths and omitting the weaknesses in the existing model will help to create a novel solution with better accuracy.

Pre-trained modes can capture more accurate linguistic features on models that are currently available and fine-tuned considering the limitations and the strengths.

1. **Evaluating the developed novel model**

Precision, Accuracy, and recall are the three main matrices to evaluate and test against the model to identify the limits of the trained model. Effectiveness of the system and fulfills the initial requirement of identifying hate speech using bilingual language.

## 3.2 Suggested Solution

## 3.3 Research Framework

## 3.4 Research Philosophy

## 3.5 Research Execution Mode

## 3.6 Data Collection Method

In this study, a quantitative approach will be used to implement on detection of hate speech in Singlish. The primary source of data gathering will be extracting hateful and non-hateful text, text-based content which are posted in Singlish from potential social media platforms such as Facebook, x, YouTube, TikTok, etc. By following this approach will allow to get natural expression through extraction and different forms of the terms that are used to interpret the same idea. As the secondary method of collecting data is getting the existing dataset which has been taken into making the previous studies under the same domain area. This approach will bridge the gaps between both primary and secondary approaches to create a robust and more effective machine-learning model to detect accurate hateful content and will create a sufficient and up-to-date data set for further enhancements.

## 3.7 Data Analysis Techniques

Considering the linguistic landscape of the study, to analyze the collected data to generate an accurate outcome number of analyzing steps are needed to be followed within the model.

At first, collected data are needed to be pre-processed which is the process of cleaning the data before analyzing. Therefore eliminating noises, removing stop words, hashtags, special characters, URLs, symbols emojis, numerical values, and symbols from the data. And making them uniform data by making each and every character to lowercase. After normalizing the uniformed terms are needed to run through a process of tokenization where it splits the text slack into individual word tokens. BERT will be the most relevant technique to use to split and tokenize the text. Tokenized texts are more humanized and to make them machine-understandable, tokenized data are required to vectorize. Which is a process that converts the tokenized text into numerical values which are machine understandable for further operations. Techniques such as mBERT. and BERT for contextual embedding and Word2Vec, and GloVe for word embedding are being used.

Vectorized data are now can be analyzed to generate accurate output running through the model.

## 3.8 Ethical Considerations

**Data Privacy**

Privacy of the collected data will be upheld with the author and no user, user name, or person identification mechanism wasn’t used or is not being used. Only the statements that were posted to the social media platforms were taken to examine the model.

**Bias and Fairness**

Three models identified contexts in three different perspectives are used to get multiple perspectives and get a better view of the posted content. This will increase the fairness of the data classification and will be unbiased when making the decisions. This will affect to accuracy and precision as well.

**Impact on Freedom of Expression**

Hate speech is bounded by the freedom of expressions. Balance is required to identify the contextual aspects of criticism and hateful contexts. Classification may affect to the freedom of expression by identifying a valid criticism as a hateful context. By using the necessary parameters model can be recalibrated to balance the ratio of freedom of speech and as well as the hate speech

# Chapter 4

## 4.1 Model Evaluation

## 4.2 Model Features and Description

## 4.3 UML Modeling

## 4.4 Technical Considerations

## 4.5 Output and Artifacts

## 4.6 Model Construction

# Chapter 5

## 5.1 Discussion

## 5.2 Conclusion

## 5.3 Future Work

# References

Brown, A. (2018) ‘What is so special about online (as compared to offline) hate speech?’, *Ethnicities*, 18(3), pp. 297–326. Available at: https://doi.org/10.1177/1468796817709846.

Castaño-Pulgarín, S.A. *et al.* (2021) ‘Internet, social media and online hate speech. Systematic review’, *Aggression and Violent Behavior*, 58(March). Available at: https://doi.org/10.1016/j.avb.2021.101608.

Deepasree Varma, P. *et al.* (2022) ‘Hate Speech detection in English and Malayalam Code-Mixed Text using BERT embedding’, *Proceedings of International Conference on Computing, Communication, Security and Intelligent Systems, IC3SIS 2022*, pp. 1–6. Available at: https://doi.org/10.1109/IC3SIS54991.2022.9885339.

DigitalSilk (2024) *How Many Websites Are There In 2024?* Available at: https://www.digitalsilk.com/digital-trends/how-many-websites-are-there/#:~:text=As of 2024%2C there are,are actively maintained and visited.

Fino, A. (2020) ‘Defining Hate Speech’, *Journal of International Criminal Justice*, 18(1), pp. 31–57. Available at: https://doi.org/10.1093/jicj/mqaa023.

*Freedom of Opinion and Expression* (2024). Available at: https://www.dagdok.org/un-by-subject/human-rights/freedom-of-opinion-and-expression/#:~:text=These principles constitute the foundation,entered into force in 1976. (Accessed: 28 February 2024).

Gaurav, A. *et al.* (2023) ‘Deep Learning Based Hate Speech Detection on Twitter’, *IEEE International Conference on Consumer Electronics - Berlin, ICCE-Berlin*, pp. 1–6. Available at: https://doi.org/10.1109/ICCE-Berlin58801.2023.10375620.

Hashmi, E. *et al.* (2024) ‘Enhancing Multilingual Hate Speech Detection: From Language-Specific Insights to Cross-Linguistic Integration’, *IEEE Access*, 12(August), pp. 121507–121537. Available at: https://doi.org/10.1109/ACCESS.2024.3452987.

*Hate speech and Hate Crimes* (2023). Available at: https://www.ala.org/advocacy/intfreedom/hate (Accessed: 28 February 2024).

Hattotuwa, S. and Wickremesinhe, R. (2023) ‘Social Media in Sport’, *Social Media in Sport* [Preprint]. Available at: https://doi.org/10.4324/9780367766924-ress7-0.

Howard, J.W. (2019) ‘Free speech and hate speech’, *Annual Review of Political Science*, 22, pp. 93–109. Available at: https://doi.org/10.1146/annurev-polisci-051517-012343.

Iginio Gagliardone, Danit Gal, Thiago Alves, Gabriela Martinez, U. (2016) *Countering online hate speech*.

Jain, A. and Sharma, S. (2022) ‘A Survey on Identification of Hate Speech on Social Media Post’, *ICAN 2022 - 3rd International Conference on Computing, Analytics and Networks - Proceedings*, pp. 1–6. Available at: https://doi.org/10.1109/ICAN56228.2022.10007283.

Kavatagi, S. and Rachh, R. (2021) ‘A Context Aware Embedding for the Detection of Hate Speech in Social Media Networks’, *2021 International Conference on Smart Generation Computing, Communication and Networking, SMART GENCON 2021*, pp. 1–4. Available at: https://doi.org/10.1109/SMARTGENCON51891.2021.9645877.

Kousar, A. *et al.* (2024) ‘MLHS-CGCapNet: A Lightweight Model for Multilingual Hate Speech Detection’, *IEEE Access*, 12(June), pp. 106631–106644. Available at: https://doi.org/10.1109/ACCESS.2024.3434664.

Kumar Kaliyar, R. *et al.* (2023) ‘HSDH: Detection of Hate Speech on social media with an effective deep neural network for code-mixed Hinglish data’, *2023 14th International Conference on Computing Communication and Networking Technologies, ICCCNT 2023*, pp. 1–6. Available at: https://doi.org/10.1109/ICCCNT56998.2023.10306709.

MacAvaney, S. *et al.* (2019) ‘Hate speech detection: Challenges and solutions’, *PLoS ONE*, 14(8), pp. 1–16. Available at: https://doi.org/10.1371/journal.pone.0221152.

Mondal, M., Silva, L.A. and Benevenuto, F. (2017) ‘A measurement study of hate speech in social media’, *HT 2017 - Proceedings of the 28th ACM Conference on Hypertext and Social Media*, pp. 85–94. Available at: https://doi.org/10.1145/3078714.3078723.

Pamungkas, E.W. *et al.* (2022) ‘Hate Speech Detection in Code-Mixed Indonesian Social Media: Exploiting Multilingual Languages Resources’, *2022 7th International Conference on Informatics and Computing, ICIC 2022*, pp. 1–5. Available at: https://doi.org/10.1109/ICIC56845.2022.10006940.

Patabendige, C.L. (2023) *Human Security Perspectives on Hate Speech*. Available at: https://www.defence.lk/Article/view\_article/27635.

Pereira-Kohatsu, J.C. *et al.* (2019) ‘Detecting and monitoring hate speech in twitter’, *Sensors (Switzerland)*, 19(21), pp. 1–37. Available at: https://doi.org/10.3390/s19214654.

Prasad, D. *et al.* (2023) ‘Real-Time Multi-Lingual Hate and Offensive Speech Detection in Social Networks Using Meta-Learning’, *IEEE Region 10 Annual International Conference, Proceedings/TENCON*, pp. 31–35. Available at: https://doi.org/10.1109/TENCON58879.2023.10322364.

Samaratunge, S. and Hattotuwa, S. (2014) ‘Liking Violence’, (September).

*Social media stats Sri Lanka* (2024). Available at: https://gs.statcounter.com/social-media-stats/all/sri-lanka (Accessed: 28 February 2024).

Tontodimamma, A. *et al.* (2021) ‘Thirty years of research into hate speech: topics of interest and their evolution’, *Scientometrics*, 126(1), pp. 157–179. Available at: https://doi.org/10.1007/s11192-020-03737-6.

*Universal Declaration of Human Rights* (no date). Available at: https://www.un.org/en/about-us/universal-declaration-of-human-rights (Accessed: 27 February 2024).

# Appendix

## Front-End Implementation

## Model Implementation

## Output