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 5](#_Toc178516354)

[1.1 Introduction 5](#_Toc178516355)

[1.2 Relevancy of The Topic 5](#_Toc178516356)

[1.2 Background of the Study 7](#_Toc178516357)

[1.2.1 Growth of the usage of social media 7](#_Toc178516358)

[1.2.2 Current context of social media-related negativities 7](#_Toc178516359)

[1.2.3 Current barriers that have provided for the safety of users of social media 8](#_Toc178516360)

[1.3 Problem Statement 8](#_Toc178516361)

[1.4 Research Question 9](#_Toc178516362)

[1.5 Motivation 9](#_Toc178516363)

[1.6 Aim 10](#_Toc178516364)

[1.7 Objectives of Study 10](#_Toc178516365)

[1.8 Challenges and Limitations 13](#_Toc178516366)

[Chapter 2 15](#_Toc178516367)

[2.1 Literature Review 15](#_Toc178516368)

[2.1.1 Definition of Hate Speech 15](#_Toc178516369)

[2.1.2 Definition of Free Speech 15](#_Toc178516370)

[2.1.3 Hate Speech in Social Media 16](#_Toc178516371)

[2.1.4 Hate Speech Detection 16](#_Toc178516372)

[2.1.5 Bilingual and Multilingual Language Processing 16](#_Toc178516373)

[2.2 Followed Methods of Detecting Hate Speech in Social Media 16](#_Toc178516374)

[2.3 Identification of the Gap 17](#_Toc178516375)

[2.4 Previous Studies of related to the topic 18](#_Toc178516376)

[2.5 Significance of the Study 19](#_Toc178516377)

[Chapter 3 20](#_Toc178516378)

**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 of 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 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.

# 1.3 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.

# 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. What are methods that hate speech spread over Facebook?
3. What are the negative effects of hate speech?
4. How are cyberbullying and hate crimes related to each other?
5. What are the positive impacts of filtering hate speech?
6. Why do people use hate speech? And what are their motives?
7. What are the current hate speech detection methodologies?
8. Will hate speech detection affect to the freedom of speech?

# 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 which 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 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.

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 (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.

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).

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).

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 this has 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 on the 2012 General Assembly 66/290 the basic right of a 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).

## 2.1.3 Hate Speech in Social Media

## 2.1.4 Hate Speech Detection

## 2.1.5 Bilingual and Multilingual Language Processing

# 2.2 Followed Methods of Detecting Hate Speech in Social Media

# 2.3 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 has 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 analytical report by distributed over the years which signifies the importance of the hate speech detection.

Most common and widely spread method is the keyword approach where it users terms from a 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 has been using 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, 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 in face book is not only English and Sinhala. The user created language known as the 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 are spread using Singlish and English on Facebook will be help to decrease the number of hate speech cases and hate crime incidents.

# 2.4 Previous Studies related to the topic

# 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.

# Chapter 3