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 Singlish on Facebook

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

DECLARATION

I declare that I was solely responsible for the data collection, data analysis, and dissertation writing. Any contributions made by other parties have also been appropriately acknowledged with references to relevant works of literature.

……..

B.A.A.V Karunathilake.

DEDICATION

This thesis is dedicated to my supervisor Head of the Department of Information and System Science senior lecturer Dr. M. Shafraz and my support system including my family and friends whose words of encouragement and push further to succeed in creating this thesis. Also, I would like to thank Ms. Dulanjali Weerasekara and Ms. Shehani Joseph for providing me with thoughtful insights about the topic and the domain. I would also like to thank Mr. Piumal Fernando (lawyer) who has provided me with the necessary information related to the law.

I will always appreciate all the support that has been given by any means by anyone, especially in helping me develop my thinking and technological skills. Finally, dedicated this thesis to all the academic staff who taught me and pursued my true passion over the past four years.

ACKNOWLEDGEMENT

Foremost, It is great to work with and I would like to express my sincere gratitude to my supervisor, Head of the Department of Information and System Science senior lecturer Dr.M.Shafraz for the continuous support of my final year research project for his guidance, and motivation, and immense knowledge.

I thank my colleagues for all the sleepless nights and discussions that succeeded in this thesis. At last but not least, my heartiest gratitude goes to my entire support system for their encouragement and appreciation.

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 dark side of freedom of speech. Darkside of freedom of expression has led to threatening, abusing, harassing, offending, and defaming individuals or entities. This study addresses the negative impacts that hate speech and hate crimes have on Sri Lankans and the way of mitigate hate speech using Singlish. We seek to understand the emotional, social, and psychological impact these incidents have on individuals and communities by looking at real-life experiences and from their perspectives.

The study also emphasizes the importance of 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 Speech detection in Singlish, Hate Crime

Contents

List Of Figures

List Of Tables

CHAPTER 1 - INTRODUCTION

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.

According to the Datareportal.com(Datareportal, 2024),

- "There were **12.34 million** internet users in Sri Lanka at the start of 2024 when internet penetration stood at **56.3 percent**."
- "Sri Lanka was home to **7.50 million** social media users in January 2024, equating to **34.2 percent** of the total population."
- "A total of **32.49 million** cellular mobile connections were active in Sri Lanka in early 2024, with this figure equivalent to **148.2 percent** of the total population."

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. My goal is to provide an effective solution to protect the worth and well-being of Sri Lankan internet users by creating a cutting-edge solution that can identify hate

speech in Singlish 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

Many studies 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 based on 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 critical 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, effective business advancement, and connections upkeep people interact/ network 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. Datareportal's figures show that there were 7.50 million active social media user identities in Sri Lanka in January 2024 (Datareportal, 2024). 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.

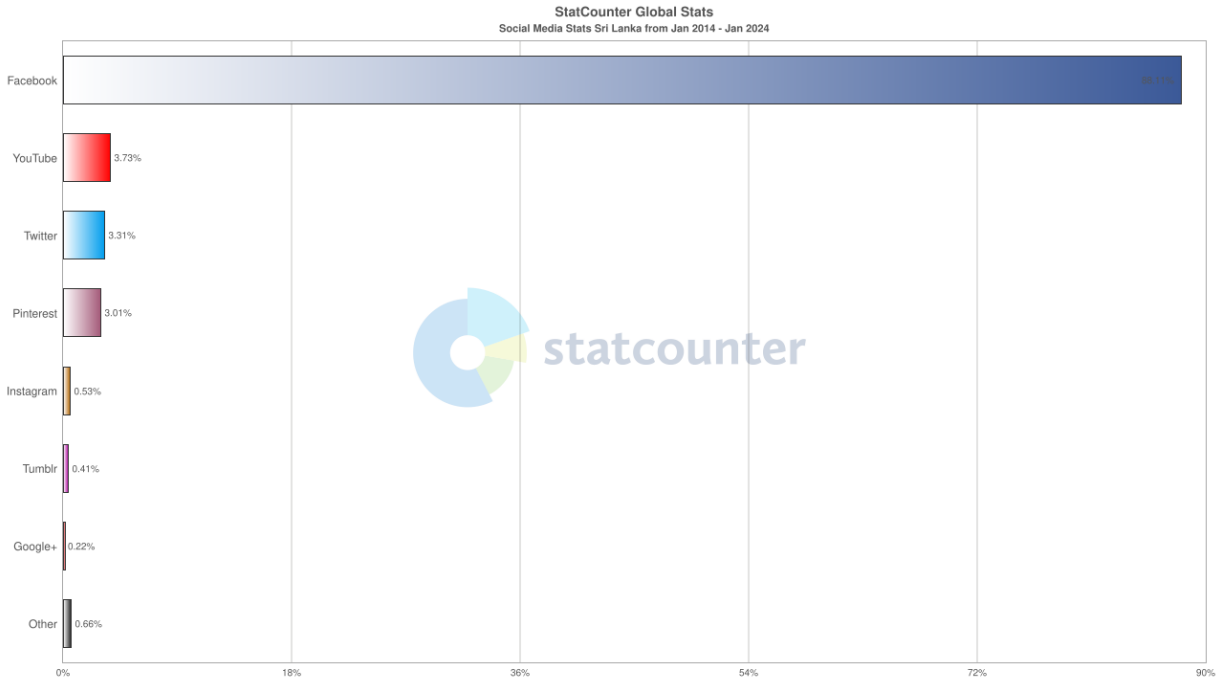


Figure 1 - 1 Statistics of Sri Lankan Social Media usage of the past decade

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(Neville Lahiru, 2024). 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.

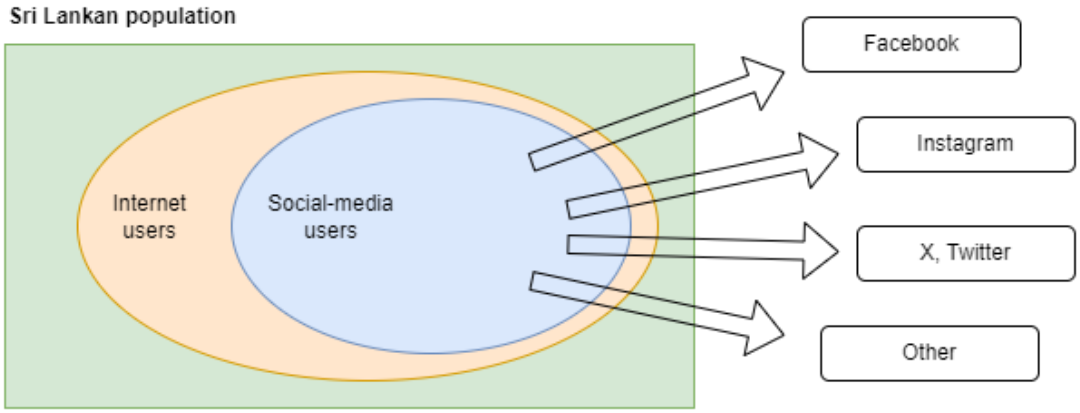


Figure 1 - 2 Sri Lankan Population Segmentation

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 an individual 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 language, and 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 as proactive action under the user protection layer or the community guidelines layer of the Facebook.

01. What is the difference between hate speech and free speech?
02. Will hate speech detection affect the freedom of speech?
03. What are methods that hate speech spread over Facebook?
04. How are cyberbullying and hate crimes related to each other?
05. 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

01. 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.

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 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, several 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 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.

05. To implement a system 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 advantages and disadvantages of the system.

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.

07. 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.

08. 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.

1.9 Outlined Solution

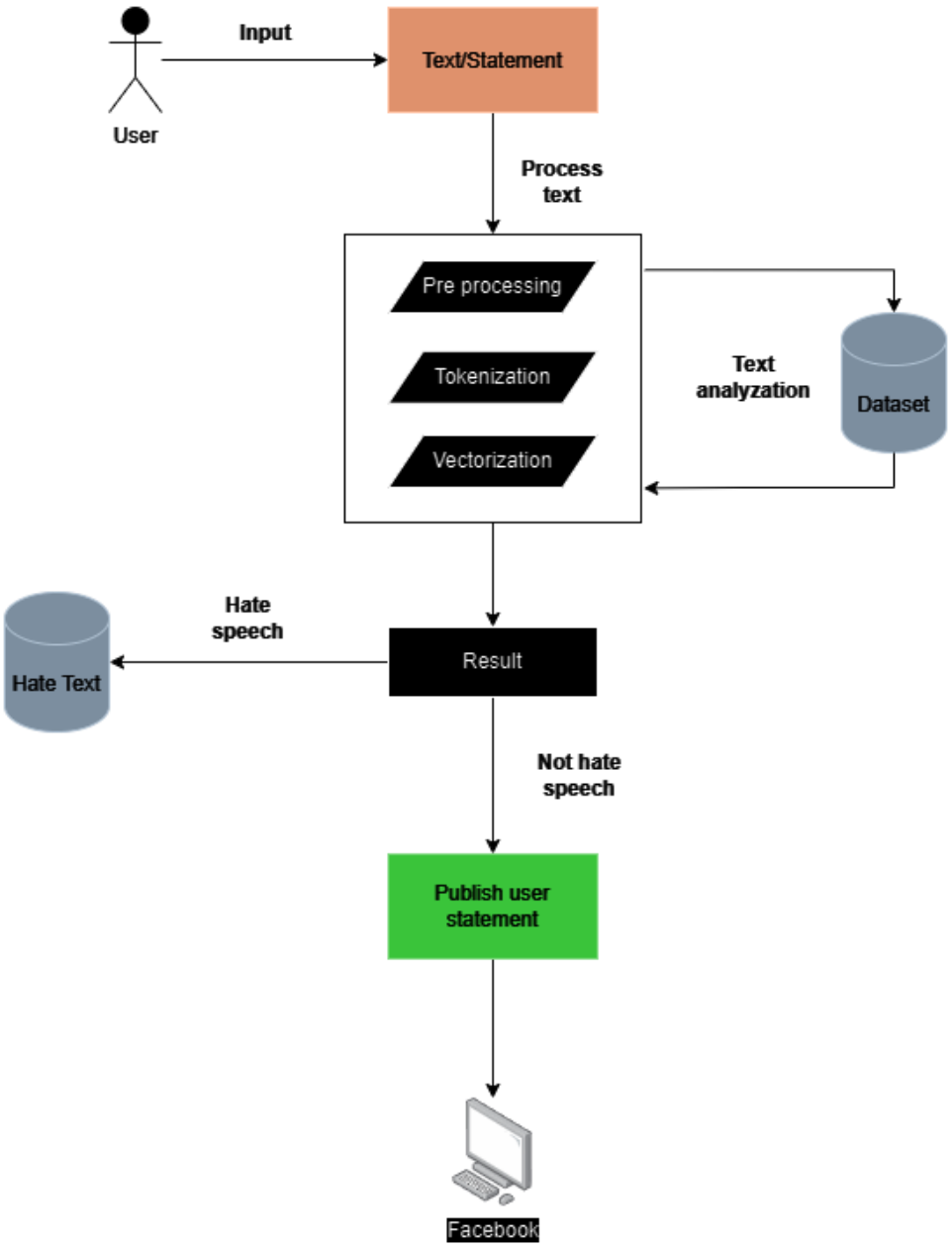


Figure 1 - 3 Outlined solution

CHAPTER 2 – LITERATURE REVIEW

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).

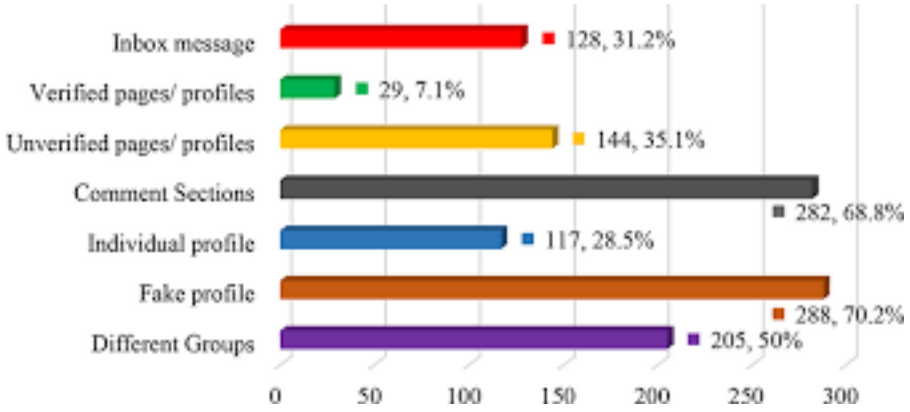


Figure 2 - 1 Hate Speech in Social Media : https://doi.org/10.1108/GKMC-01-2022-0014

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 basis for creating a hate speech detection tool 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 problem

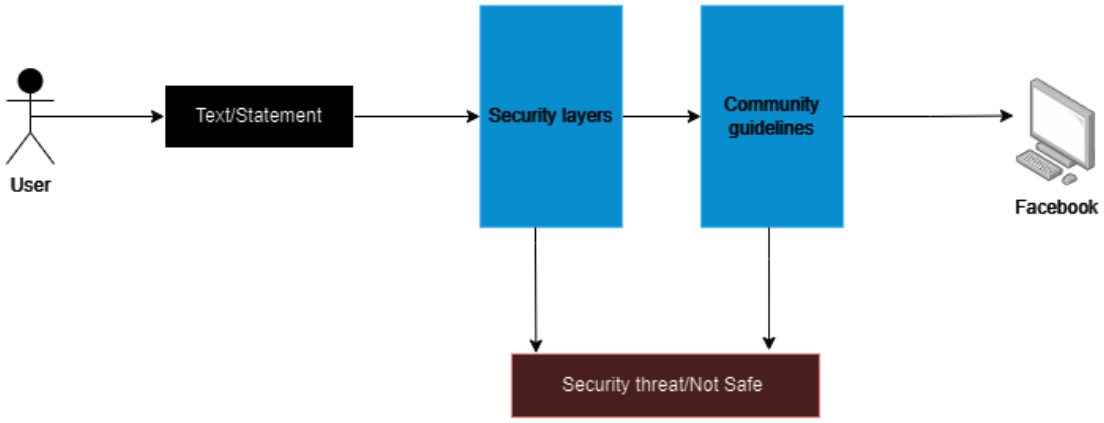


Figure 2 - 2 Existing Method (Layered)

Adhering to the current context, Singlish is widely spread in social communication. And also Singlish is not considered as a standard language. Therefore, it doesn't validate within the Facebook platform unless that content or statement is reported by another user within the platform unlikely English, Sinhala, French, Hindi, Tamil, etc. Undoubtedly it is lacking in identifying hateful or offensive statements from such languages which allows any user to tarnish, defame, abuse, or threaten using these types of community created languages and taking such loopholes and taking advantage of such in their causes.

Figure 5 explains the flow of the existing process and the two main layers of Facebook that validate their security measures and cross walking of the community guidelines which is identified as an inappropriate action. Others will be posted since they pass through these security and guideline layers. In the current case, a user must meet the security requirements as well as the requirements that are mentioned in the community guidelines.

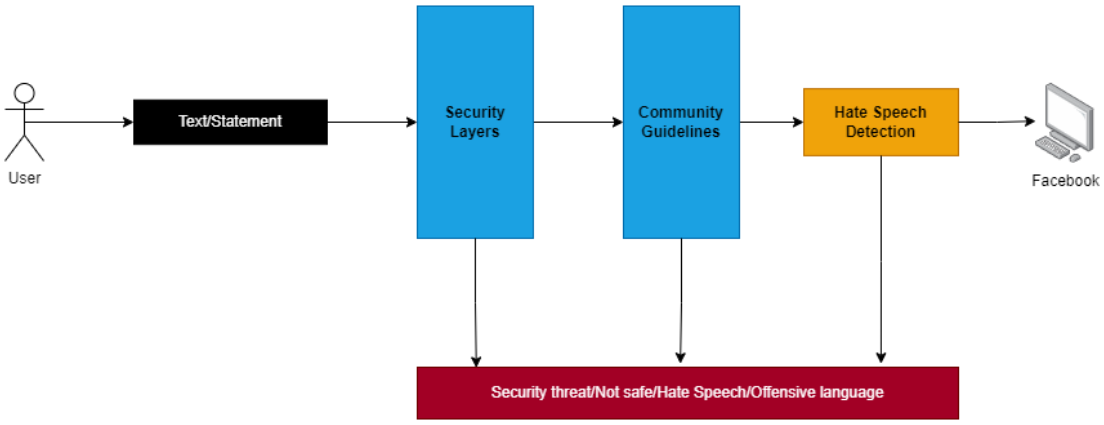


Figure 2 - 3 Suggested Solution (Layered)

On the other hand, the Suggested solution has the ability to understand hate speech content posted in Singlish and is able to encounter those as other security violations, community guideline violations, and hateful and offensive content. As per the Figure 6 only the hate speech identification method has been implemented where it addresses the main issue mentioned within the current study since hate speech is identified as a major community thread.

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.

MBERT 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.

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.

2. **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.

3. **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.

4. **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.

5. **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

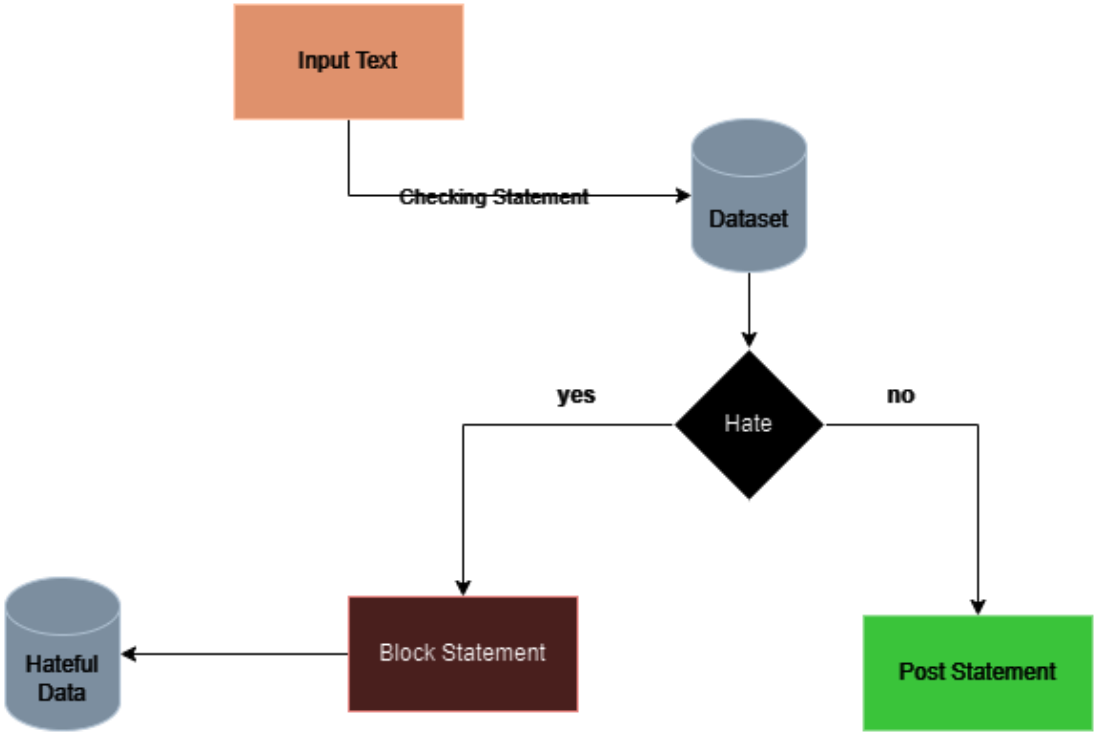


Figure 3 - 1 Overview of the suggested solution

The above illustration shows the solution from the primary level of perspective each has its own complex one or more processes that have happened from the background. The given idea is once a user releases a text input to the Facebook platform before it gets posted, it pre pre-processed and analyzed using the model that is fed by the given data set. The model itself decides whether the statement is hateful or not. If the provided statement is identified as hate speech then it will be flagged and recorded that the statement contains hateful content and is not best for the platform. If the statement is flagged as non hate content it will be posted. This process takes place in real which will not posted until it checks with the model. And for the ones who posted hate content will go along with the Facebook security and community guidelines protocols and within the current study, it only discussed about the hate speech model to increase user safety.

The suggested solution has a machine learning model to train itself with the given data and will be able to feed more data to encounter complex scenarios and word bindings. Offensive terms and offensive terms included statements have been considered as they must be removed from the platform. Hateful content and

offensive statements will be flagged and recoded to another data set which collects the ultimate dataset containing only hateful content. This data set can be used for the existing model to increase its performance, precision, and accuracy or can be used to build a better model using advanced algorithms and train the model using this collection of data. Balance of the using dataset is mandatory since the model is unaware of the language and Singlish has never been used before to capture from libraries. Therefore it requires having both hateful and non hateful statements in a similar ratio to get better results.

3.3 Research Framework

The overall study has been carried out using the deductive research framework using the quantitative approach. The entire process has been conducted using four cycles. Information and relevance, Ideate, implementation, and evaluation are the 4 cycles that are used to create a novel solution and relevant artifacts from the study.

Ultimate artifact is the solution that addresses the issues of hate speech detection in Singlish. As a complementary output will be the data set and the findings that are up to date in this current domain.

3.4 Research Philosophy

The Study aims to build a solution to detect hate speech in Singlish on Facebook to ensure the platform's safety and increase user safety. The Ultimate objective is to create a harmonious environment by reducing hate from such platforms due to the immensely growing social media platforms on the internet. The study requires collecting comments and text based posting that posted by users using Singlish on Facebook to create a collection of data. These texts/ statements need to be categorized or labeled as hate speech, offensive, or neither. Neither stands for statements that don’t carry any hateful or offensive words or expressions. In order to detect hate speech, the model needed to be trained using all three labels. Even though Singlish is widely used in online communication, data availability of the Singlish language is the biggest challenge because data has not been created for research purposes. With the context of the study, a quantitative approach was suggested to continue. Since the problem is not hypothetical or theory based, it is required to conduct a comprehensive

pre study to get the previous studies and their knowledge to get inspiration and insight into the current study. Since the model uses text/ statements as the inputs, which classifies the statements it requires a number of inputs to train rather than considering the quality based characters of the statements which made the study to take the quantitative approach.

In the current study, continues the discourse of the freedom of expression and hate speech and also the neediness to detect and remove hate speech from social media platforms such as Facebook. [model]

When collecting data, in order to protect the privacy of the users basic ethics and considerations have been followed throughout the process.

3.5 Research Execution Mode

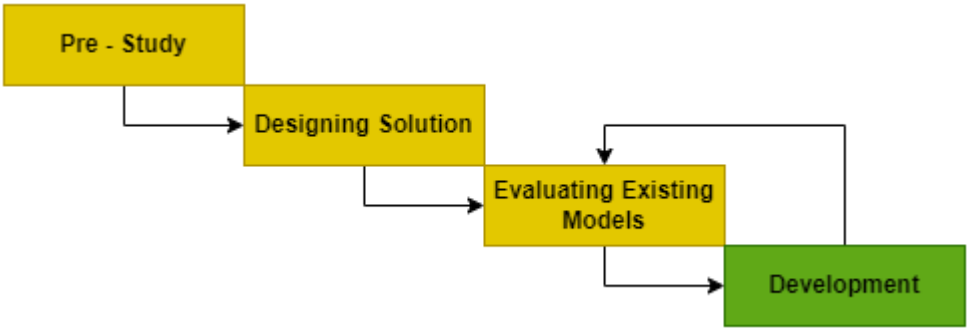


Figure 3 - 2 Execution methodology

The study was executed as per the above illustrated diagram in Figure 8. This has been a process that is more like a waterfall from pre-study to implementation of the system.

Pre-Study

Pre-study is the period of getting into the domain and familiarizing with the artifacts and products that have been built up to date. Such as previous studies, models, datasets, statistical information, support articles, etc. Pre-study helps to identify and get to know about the problem and also to identify the gaps and design a creative solution to bridge those gaps. This stage also involves an in-depth exploration of the key issues related to the research problem, examining various sources of information and understanding the methodologies and approaches previously used. By thoroughly analyzing the existing data and research, the pre-study phase sets the foundation for

the entire project. It enables the researcher to map out potential solutions while ensuring that the final design will effectively target the problem areas and contribute to the field by filling the identified gaps. Ultimately, this phase is crucial for shaping the direction of the research and ensuring its relevance and impact.

Designing Solution

After a comprehensive pre-study, the next step is to design a solution that addresses the problem statement and fills the necessary gaps identified during the research phase. The design must be targeted, focusing on the specific issues highlighted, and must also be feasible within the study's scope, resources, and time constraints. It is essential for the solution to be both practical and realistic, ensuring that it can be developed efficiently and deliver the expected outcomes. The design should aim to resolve the identified challenges while being adaptable to potential obstacles. Ultimately, this stage paves the way for achieving the research objectives by proposing a clear and achievable approach.

Evaluating Existing Models

After getting the requirements and outlined solution, It is required to identify the most suitable models and methods for developing the solution effectively. Therefore, with the inspirations of the pre-study documents and their respective deliverables, it is a process of selecting and implementing the best most suitable models and technologies to develop the solution.

Therefore, it requires building multiple solutions using multiple models using the data set/ data sets. Each model has its own characteristics that help in achieving the best results.

Models are evaluated using available measurements such as precision, accuracy, and f-score.

Development

Development of the solution has needed to be created from scratch because Singlish has not been used earlier in making hate speech detection systems, sentiment analysis or any other text classification as an individual language.

Data to create the dataset, which is required for the development was created manually using the authentic comments that were posted by real users to some other individuals, organizations and influencing characters(Actors, Social-Media influencers, Celebrities, Politicians).

Development has been divided into three phases because of the same reason that has occurred in creating the dataset.

Figure 3 - 3 Model development phases

Phase 1 – Manually labeling hate and non-hate data

An appropriate data set has not been created to use for these clarification model-based systems. Therefore, manual labeling is used for identification purposes. Models are unaware of the context of the textual inputs because Singlish is not a language that has been used for hate speech detection.

Identifying the linguistic characteristics of a statement requires having pre-trained models or pre-defined labels to identify the class of the statement.

Hate speech	Offensive	Neither	State
1	0	0	1
0	1	0	1
0	0	1	0
1	1	0	1

Table 3 - 1 Labels of the dataset

Manual labeling has been used as above in Table 1. Hate speech, Offensive, Neither, and overall state are the labels that have been used to flag against the text based statements.

1 = ‘Yes’ ,

0 = ‘No’.

Phase 2 – Using Machine learning model/ models to test and train using the dataset

After collecting data and training a model using manual labeling are able to train and test more machine learning models to facilitate the advanced requirements that are not covered in Phase 1.

Machine learning models can be trained and intelligent to use supervised and unsupervised learning to learn continuously to fulfill the task. In this context that is detecting hate speech on Facebook using Singlish language.

Created and gathered labeled data are sufficient to train traditional models to get a positive prediction. More data will lead to a more accurate, precise result when detecting hate speech.

Getting into the machine learning approach will provide a more effective method to detect hate speech and the model itself is able to understand the contexts and the statements.

Phase 3 – Using advanced machine learning algorithms

Advanced machine learning algorithms enable us to understand many complex scenarios. Such as code-switching, natural language processing, contextual identification, paraphrasing, etc.

To train a successful model it requires an enormous number of data. The system that was built in the previous phase has gathered a number of data related to hate speech that can be used to implement advanced machine learning algorithms such as deep learning, neural networks, etc.

Each of these phases contains challenges that are native to the respective phases. Those challenges will be addressed either as a novel solution or as an advancement in the following phase.

Phase 3 has been considered as the ultimate solution for detecting hate speech in the Singlish language using novel and advanced algorithms and libraries.

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 that is posted in Singlish from Facebook. Since the study is focusses on addressing the hate speech issue on Facebook. 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.

Since Singlish is a user created language that has never been used on such causes, therefore to create a usable data set, it requires creating a dataset manually to use in this system in order to identify and detect hate speech.

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.

The dataset is required to be inserted beforehand to train the model and to understand the above mentioned linguistic landscapes and characteristics of the language. Labeling needs to be provided to identify the behavior of the given text based content. In this case, it has been given manually and directed manually because of the Singlish language.

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. 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.

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

By testing and training the given data will be analyzed against the given labels.

3.8 Ethical Considerations

Data Privacy

Privacy of the collected data will be upheld with the author, and no user, username, or personal identification mechanism was used or is being used. Only the statements that were posted to the social media platforms were taken to examine the model. Furthermore, the research strictly adheres to anonymization principles to ensure that no personal data is linked to any individual. The focus is entirely on the content, ensuring that no user profiles or identifiable information are involved. All collected data is handled in compliance with ethical guidelines, and access to it is limited solely to the author for the purpose of model development. By maintaining this approach, the research not only protects individual privacy but also ensures that the study remains objective, focusing purely on the linguistic aspects of hate speech detection without any bias toward the users themselves.

Bias and Fairness

Three models identified contexts in three different perspectives are used to get multiple perspectives and provide a more comprehensive view of the posted content. This approach ensures a broader evaluation of each statement, allowing for more nuanced understanding. By considering multiple angles, the analysis becomes more thorough, reducing the risk of misinterpretation and enhancing the overall robustness of the classification process. This will increase the fairness of the data classification, ensuring that no single model biases the outcome. As a result, decisions will be more balanced and objective, leading to a more reliable outcome. This method not only promotes fairness but also positively impacts both accuracy and precision, making the system more effective at detecting hate speech while minimizing false positives or negatives.

.**Impact on Freedom of Expression**

Hate speech is bounded by the freedom of expression. A careful balance is required to distinguish between valid criticism and genuinely hateful contexts, as misidentifying a legitimate critique as hate speech could unjustly limit an individual's right to free speech. It is important to ensure that the model does not overreach by flagging or censoring comments that may be strong in opinion but not inherently hateful. Classification may affect the freedom of expression by mistakenly identifying a valid criticism as a hateful context. To address this, the model can be recalibrated using necessary parameters that better account for tone, context, and intent. By fine-tuning these settings, the system can better balance the ratio between freedom of speech and hate speech detection. This approach will allow for a more accurate distinction, ensuring that while harmful content is effectively identified, genuine discourse and criticism are protected.

3.9 System Evaluation Model

To Evaluate the system it needs to be fit for the requirement. Therefore, it needs to be run and checked multiple times. First, build the detection model using different machine learning models. It must ensure that it fits with the given requirements with the functionalities. Then provide the same data set to train and test for each model. The dataset should be the same and it should consist of both hate and non hateful data also it is required to have them on a balanced ratio. Then all the modules should be evaluated with common measurements such as precision, accuracy, f1-score, etc. These scores will be based on the given data set and identification with the given labels and features of statements. Then to evaluate with an unbiased random check should be conducted by entering multiple statements into every model and recording its real accuracy in practical scenarios such as short text, long text, relevant statements, hate, and non hate text, etc. With this examination and with the feedback of the result will be able to find the best model that fits for the job of identifying hate speech in Singlish on Facebook.

CHAPTER 4 – RESULTS

4.1 Chapter Overview

Under chapter Four, discusses a number of machine learning models and their performance, fit to the job, and each of their outcomes and relevant findings that were considered by the author and recommended by previous studies. Data set and their usage, performance, level of accuracy, level of precision, and level of F1-score are taken as to measure each model. External tests have been conducted to examine the fit for the purpose of each model under this chapter.

A considerable finding that has been discovered was not a single model is able to perfect fit for the job but has performed well in the cause.

4.2 Data collection related to hate speech and non-hate speech

	Count(No.)	Percentage(%)
Hate	60	33.33
Offensive	34	18.88
Neither	86	47.77
	180	100

Table 4 - 1 Summary of the dataset

The data collection method has been discussed in the 3.6 Data Collection Method. 100 statements are collected and identified as that within those statements are consists with hate spreading expression. 50 statements are identified as that they contain offensive statements among the data set. 100 statements are identified as neutral cases which are not identified as hateful or offensive content.

A total of 250 statements were gathered from under each case (hate, offensive, neither). The current study has considered that offensive words/ statements are

considered hateful content and must be eliminated for the purpose of ensuring the platform's safety and creating a safe environment within the Facebook platform.

Hateful content = Hate statement + Offensive statements

Therefore, the study continued and classification was based on hateful and neither statements to train each model.

4.3 Labels and Description

4.3.1 Hate

The label (column) ‘Hate’ contains hateful statements that are used to defame, tarnish, or false and aggressive statements towards another individual, organization, or community.

These statements are filtered sarcastic and also from healthy criticism. Each of these statements provides either hateful, false information to defame that subjected person organization, or community.

4.3.2 Offensive

The label (column) ‘Offensive’ indicates the statements that contain offensive words or multiple words within the statements.

Offensive is considered to be a must-eliminated set of words from such platforms therefore all the offensive statements are considered to be must-eliminated and flagged as hate content.

1 is used to indicate that the statement contains offensive content and 0 is used to indicate that this contains no offensive content.

4.3.3 Neither

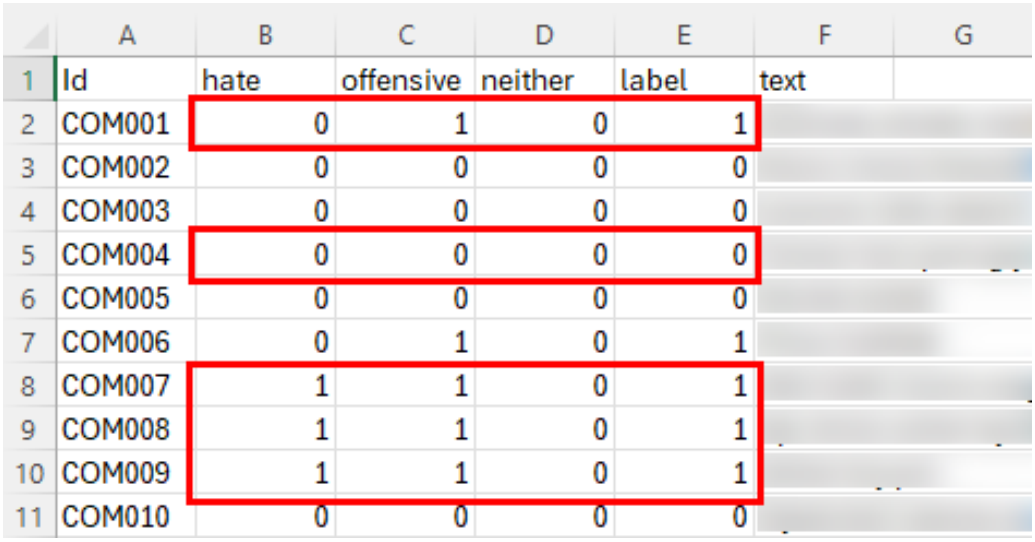
The label (column) ‘Neither’ contains neutral statements that do not have hate or offensive content within the statements.

1 is used to indicate that the statement contains no hate or not containing any offensive content

4.3.4 Label

The label is the column in the dataset that indicates whether that particular statement is hate speech or not considering other columns.

1 is used to indicate that the statement contains hate and 0 is used to indicate that this contains no hate.



	A	B	C	D	E	F	G
1	Id	hate	offensive	neither	label	text	
2	COM001	0	1	0	1		
3	COM002	0	0	0	0		
4	COM003	0	0	0	0		
5	COM004	0	0	0	0		
6	COM005	0	0	0	0		
7	COM006	0	1	0	1		
8	COM007	1	1	0	1		
9	COM008	1	1	0	1		
10	COM009	1	1	0	1		
11	COM010	0	0	0	0		

Figure 4 - 1 Dataset labels

4.3.5 Text

Under the 'Text' column, is the collection of statements that are collected from Facebook that are posted as comments in several types of accounts and content.

4.5 Results and Findings of the Study

The study is conducted to build a system to detect hate speech in Singlish on Facebook to create a harmonious platform and ensure the safety of its users when developing the system multiple machine learning models have been considered and built prototypes to test against the practical scenarios.

Both Traditional and deep learning models were used to test and find the best performing classification model to build the detection system.

- Linear Models
    - Logistic Regression

    - Naïve Bayes
- Tree-Based Models
    - Decision Tree
    - Random Forest
- Support Vector machines
    - SVM
- Neural Networks
    - Deep learning
    - Transformers - MBERT

Each model has its own capabilities and limitations in detecting hate speech.

Model	Capabilities
Logistic Regression	• Logistic regression is often used for classification tasks. • Ability to feature prediction. • Ability to work with smaller data sets. • Useful for binary classifications.
Naïve Bayes	• Efficient computational power. • Performs well in smaller datasets. • Best for real-time tasks due to the processing speed. • Effectively working with TF-IDF, BoW
Decision Tree	• Easy to understand and visualize, less complex. • Better at making decisions • Better performing with medium and larger datasets.
Random Forest	• Better o handling unstructured data. • Using multiple decision trees to avoid overfitting. • Ranking important features.

SVM	• Effective for high-dimensional spaces. • Working effectively with smaller datasets. • Best for simple languages.
Deep Learning	• Ability to capture sequential patterns and contextual relationships. • Automatically learning features. • Works well with larger datasets
MBERT	• Bidirectional. • Model is pre-trained and can be fine-tuned using datasets. • Multilingual support. • Contextual understanding.

Table 4 - 2 Capabilities of each model

Model	Limitations
Logistic Regression	• Only considers the linear relationship between features. • Dataset needs to be balanced to get a better performance.
Naïve Bayes	• Unable to identify patterns and dependencies within the features. • Limited context understanding.
Decision Tree	• Poor generalization in complex tasks. • Overfitting on small or imbalanced datasets. • Struggle to capture context and relationship within the features.
Random Forest	• Larger the dataset, Slower the process. • Need pruning and tuning for better performance. • Struggles to understand context and the

	sequence of the features.
SVM	• Doesn’t provide probabilities. • Computational resources increases with the size of the dataset. • Less suited for contextually heavy dataset.
Deep Learning	• Requires large number of labeled datasets to learn. • Computational are higher comparatively. • Predictions reasoning cannot be seen. • Not effective and suitable for low resource settings.
MBERT	• Requires high computational power. • Difficult to use in real-time. • Complexity is high

Table 4 - 3 Limitations of each model

Above Table 3 and Table 4 provides the information related to the capabilities and limitations respectively.

Each of the models performs differently. The same data set is used to train each model and based on the given data, it has provided multiple performance indicators to measure the model's suitability towards the study’s objective.

It is required to find the optimum level to split the dataset to train and test the model. Therefore, each and every model will be trained under 50%-50%, 40%-60%, 30%-70%, 20%-80%, and 10%-90%.

- **Naïve Bayes**

Naïve Bayes is based on probability and is great for text classification tasks. It assumes that the presence of a word is independent of others, which can be a limitation.

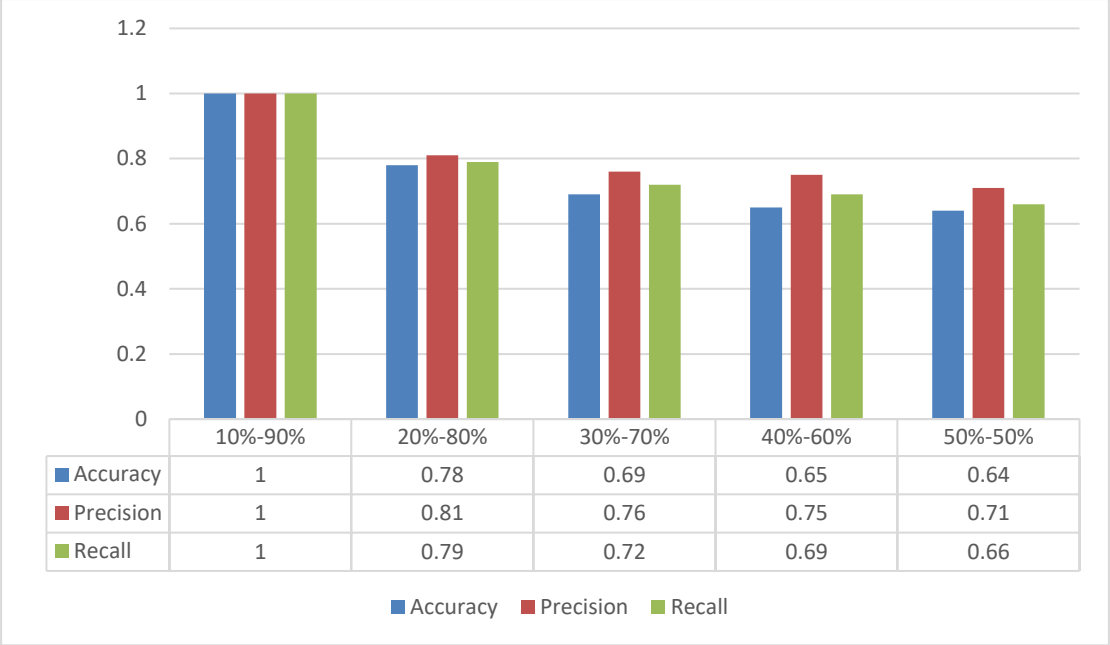


Figure 4 - 2 Naive Bayes: Measurements

- **Logistic regression**

Logistic Regression is a basic model that predicts whether something belongs to one class or another, In this case, hate speech or not. It's simple and easy to implement, but it might not capture more complex patterns in language.

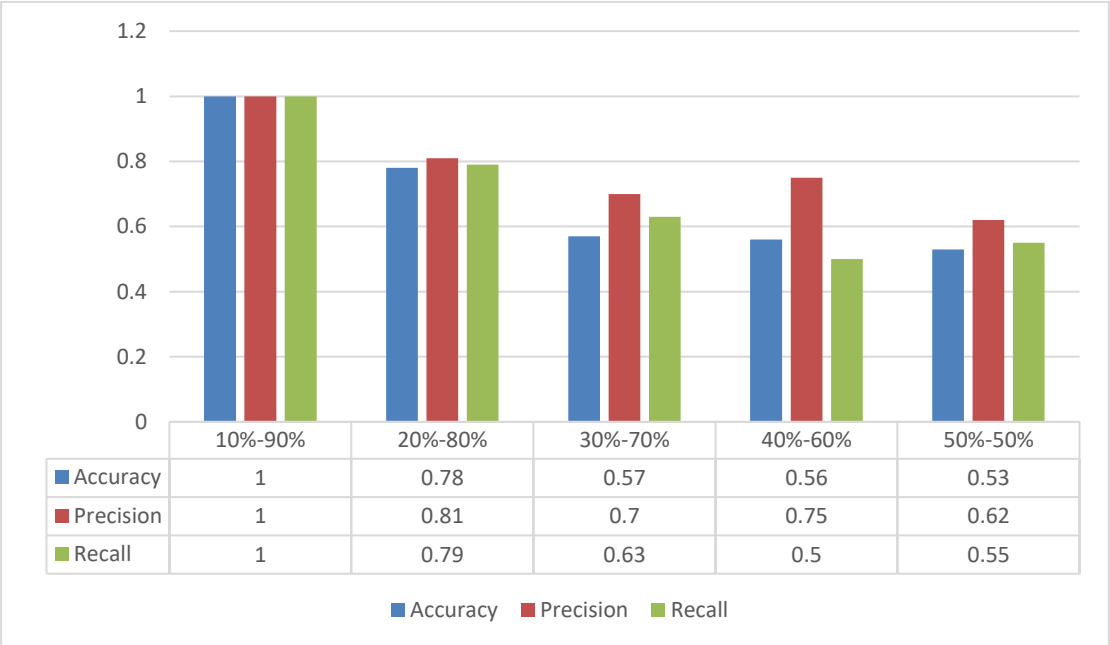
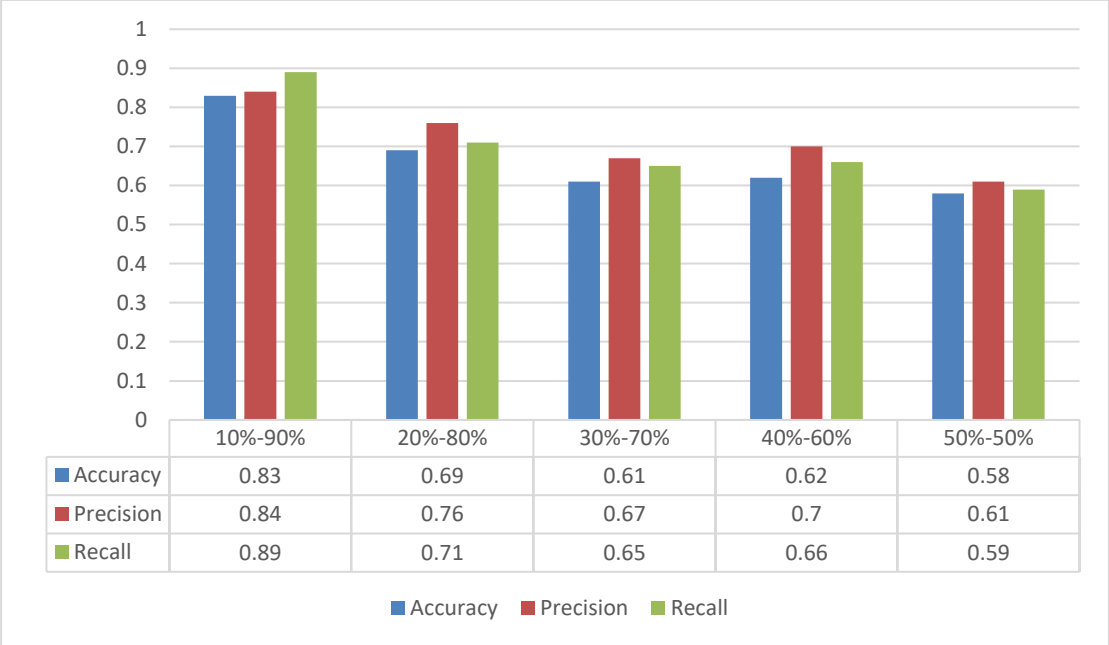


Figure 4 - 3 Logistic Regression: Measurements

- **Decision Tree**

A decision tree is a simple model that works by splitting data based on certain features. For hate speech detection, it breaks down the text into simpler decisions. It's easy to understand and visualize.

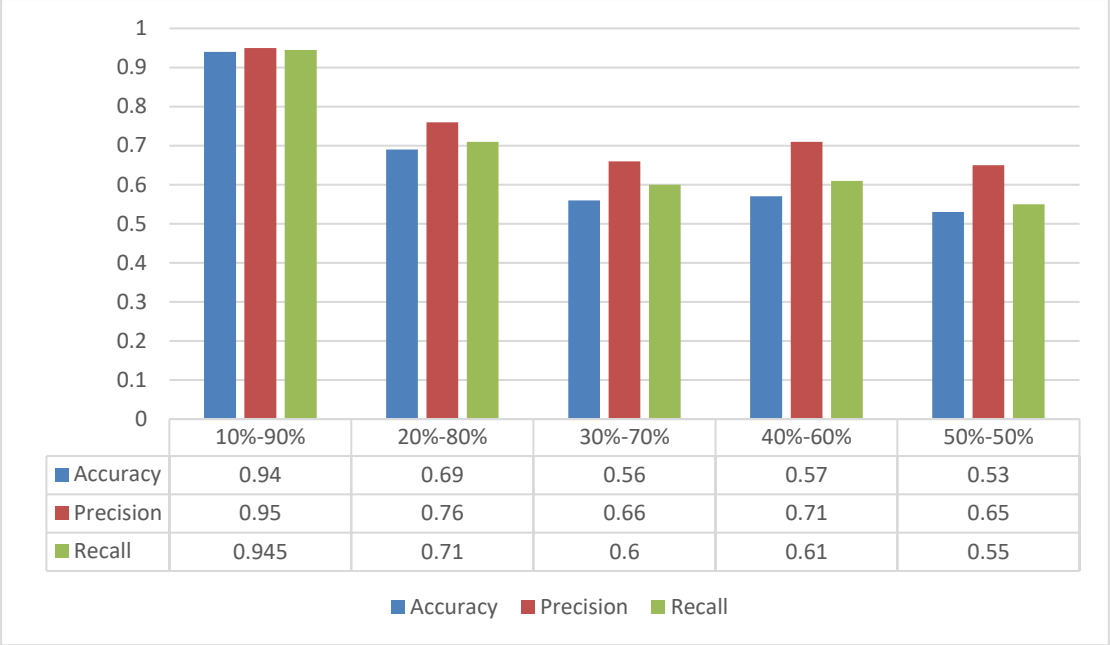


	10%-90%	20%-80%	30%-70%	40%-60%	50%-50%
Accuracy	0.83	0.69	0.61	0.62	0.58
Precision	0.84	0.76	0.67	0.7	0.61
Recall	0.89	0.71	0.65	0.66	0.59

Figure 4 - 4 Decision Tree: Measurements

- **Random Forest**

Random Forest is like a collection of decision trees. It combines the results of multiple trees to improve accuracy. It’s good at handling large datasets and reducing errors from individual trees, making it more accurate for tasks like hate speech detection.

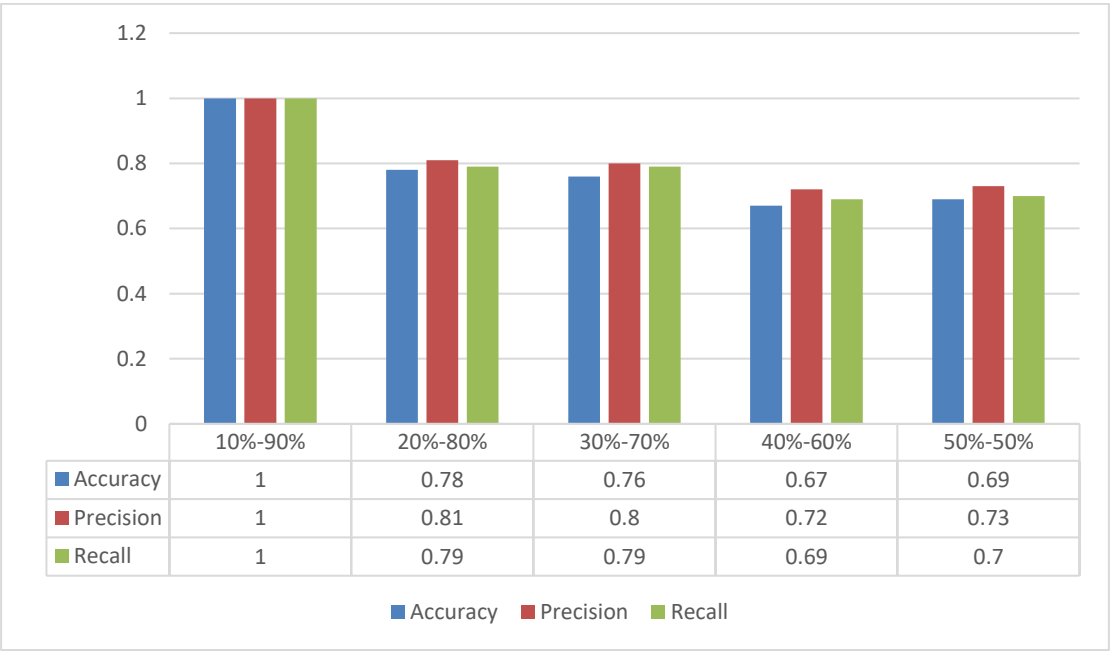


	10%-90%	20%-80%	30%-70%	40%-60%	50%-50%
Accuracy	0.94	0.69	0.56	0.57	0.53
Precision	0.95	0.76	0.66	0.71	0.65
Recall	0.945	0.71	0.6	0.61	0.55

Figure 4 - 5 Random Forest: Measurements

- **SVM**

SVM works by finding the best boundary between different classes (like hate speech and non-hate speech). It's good at handling smaller datasets and can be effective for detecting hate speech, though it may struggle with very large or complex data.

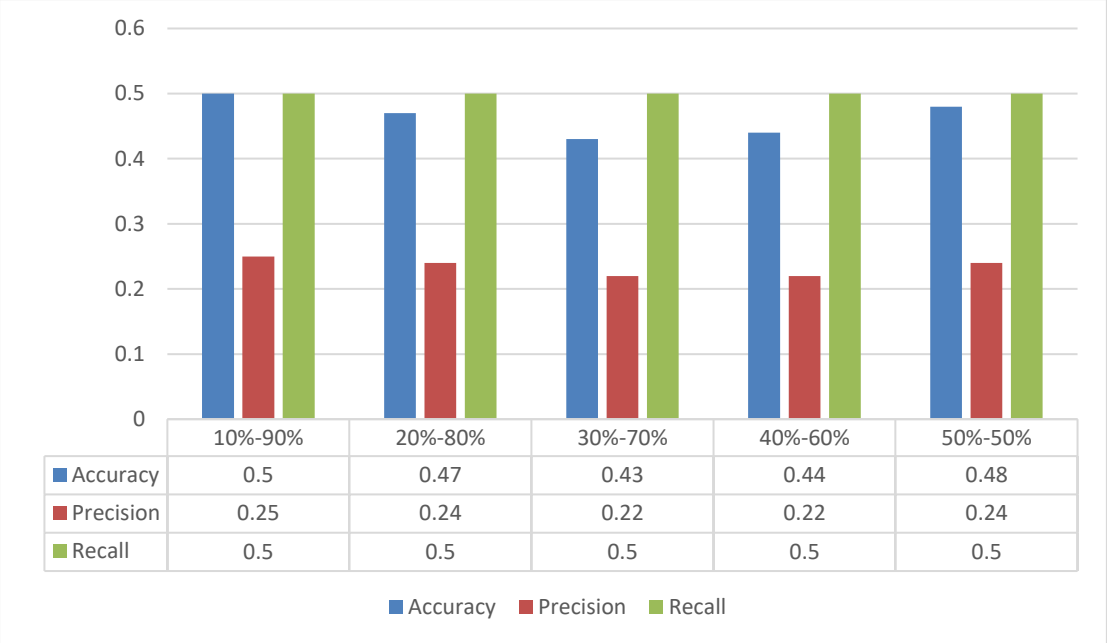


	10%-90%	20%-80%	30%-70%	40%-60%	50%-50%
Accuracy	1	0.78	0.76	0.67	0.69
Precision	1	0.81	0.8	0.72	0.73
Recall	1	0.79	0.79	0.69	0.7

Figure 4 - 6 SVM: Measurements

- **Deep Learning**

Deep learning models are inspired by the human brain. They work well for large datasets because they can learn complex patterns, even hidden ones. For hate speech detection, deep learning can capture difficult-to-detect patterns in language, but it needs a lot of data and processing power to perform well.

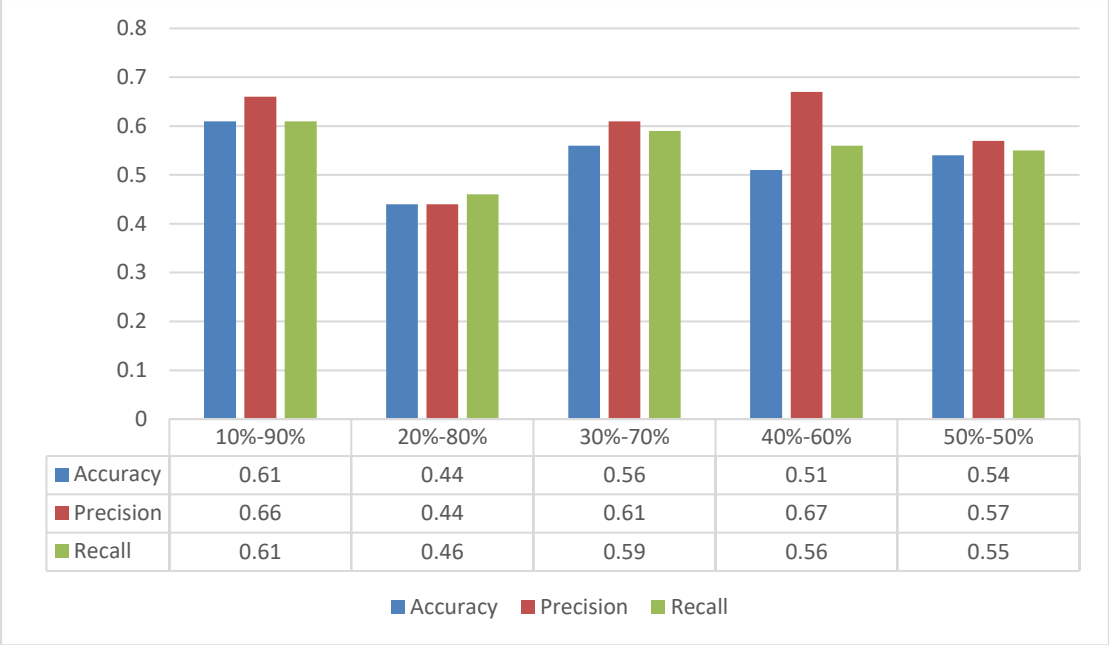


	10%-90%	20%-80%	30%-70%	40%-60%	50%-50%
Accuracy	0.5	0.47	0.43	0.44	0.48
Precision	0.25	0.24	0.22	0.22	0.24
Recall	0.5	0.5	0.5	0.5	0.5

Figure 4 - 7 Deep Learning: Measurements

- **MBERT**

MBERT is a powerful language model from deep learning. It understands the meaning of words in context, making it very good for analyzing sentences with complex meanings, such as identifying subtle hate speech. MBERT requires a lot of data and computing power but provides high accuracy.



	10%-90%	20%-80%	30%-70%	40%-60%	50%-50%
Accuracy	0.61	0.44	0.56	0.51	0.54
Precision	0.66	0.44	0.61	0.67	0.57
Recall	0.61	0.46	0.59	0.56	0.55

Figure 4 - 8 MBERT: Measurements

In machine learning, the recommended split for training and testing data is usually 80% for training and 20% for testing. This helps models learn from a large portion of the data while still having enough data to test performance. However, in this case, I found that for some models 10%-90% split (10% for testing, and 90% for training) showed better performance. This could be because of the size of the dataset used for detecting hate speech in Singlish. Smaller datasets may benefit from using more data for training to help the model learn better, which can lead to improved results in testing.

Most models indicate higher values in 10%-90% and 20%-80% data splits, except for Deep learning and MBERT models. Deep learning indicates its highest performance 10%-90% and 50%-50% split which is with a slight difference. On the other hand, MBERT indicates its highest performance in a 10%-90% and 30%-70% data split.

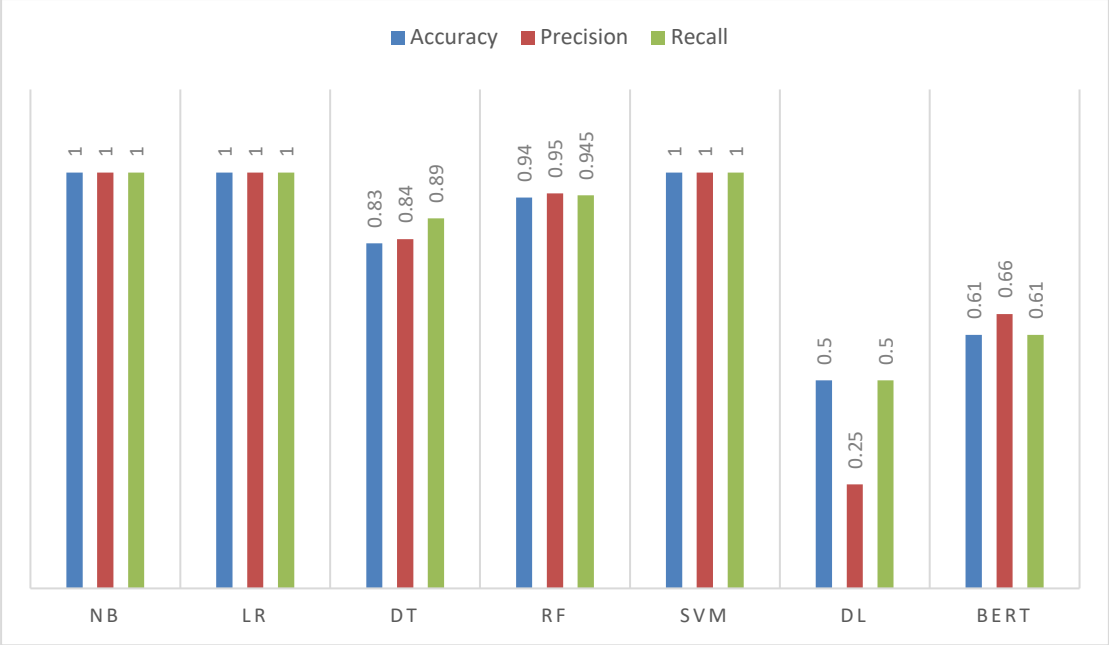


Figure 4 - 9 Summary of models: Measurements

4.5.1 Testing and Evaluation

15 statements that are collected from Facebook will be entered as user inputs as the test. The same test will be conducted across multiple models Naive Bayes, Logistic Regression, Decision Tree, Random Forest, SVM, Deep Learning, and MBERT to assess their performance at their optimal data splitting levels. This test aims to evaluate the practical usability of each model, focusing on the task of detecting hate speech in Singlish on Facebook.

This evaluation seeks to determine the most suitable model by comparing them at their top two optimal split levels. These splitting levels represent how the dataset is divided into training and testing sets, which significantly affects the model's ability to learn and generalize. By exploring multiple split ratios, the goal is to identify which models perform best, regardless of specific metrics discussed in previous sections. This ensures a focus on their practical performance in real-world applications, beyond theoretical accuracy.

The test will provide insights into how each model handles the unique challenges of hate speech detection in Singlish, such as slang, mixed languages, and informal tone. Evaluating all models under consistent conditions and comparing their results will highlight which models achieve not only high accuracy but also maintain robustness

in real-world contexts. Furthermore, by examining various split ratios, the evaluation will determine the point at which each model strikes the best balance between precision, recall, and overall effectiveness.

This approach ensures that the final model selected is not only statistically sound but also capable of detecting hate speech effectively in the dynamic, context-rich environment of social media. The evaluation will guide the choice of a model that is both theoretically strong and practically applicable for hate speech detection.

In this study, it considers hate speech and offensive language. Due to the ethical constraints test data will not be discussed or will not be added to this report and test results and a summary will be shown within the report.

The Test includes 15 statements which are mixed with hateful, offensive, and neutral statements.

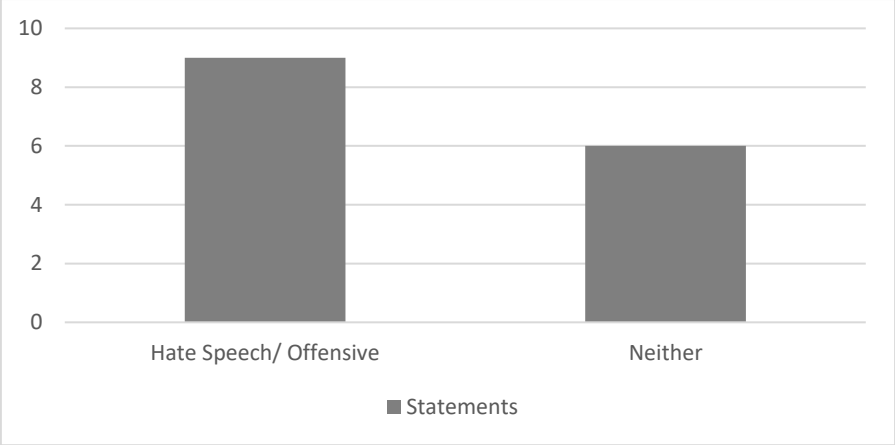


Figure 4 - 10 Test Statement Summary

For each model, 2 test rounds were carried out and total test rounds were conducted in the testing phase to find the best model which suits to achieve the goal of the study. Figure 20 displays the summary of the results.

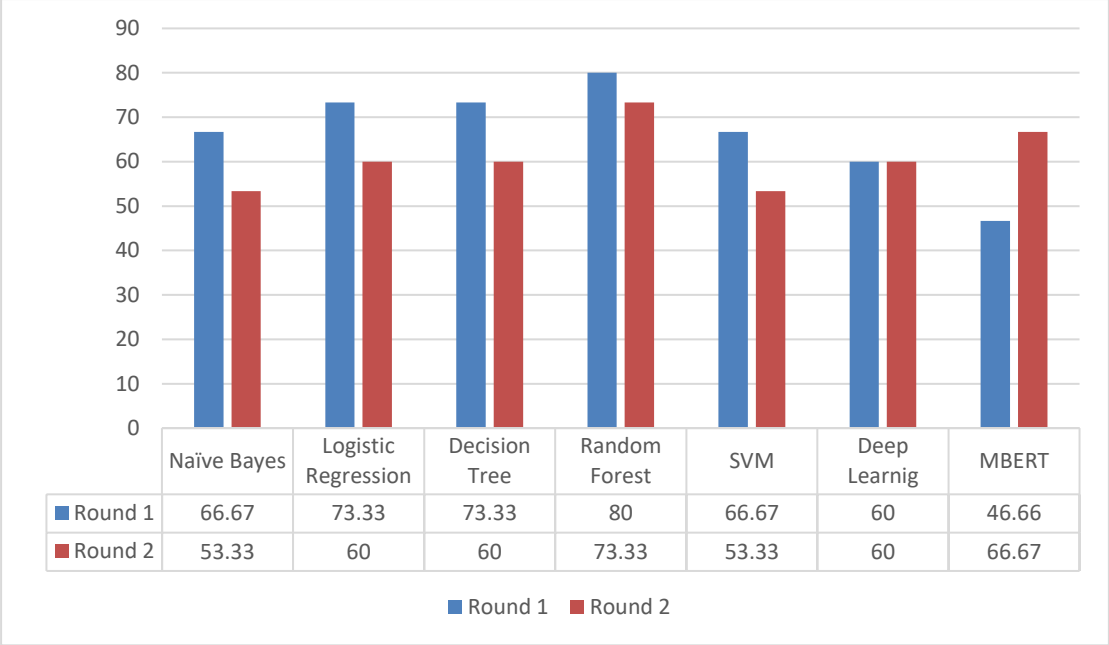


Figure 4 - 11 Summary of test results

4.5.2 Results

Results over 7 models show the ability to predict and classify hate speech in Singlish. By going through the data, it shows that as the suggested 10%-90% data split has given better results compared to other splits.

Random forest model given the highest classification rate of 80.00 in a 10%-90% split while, Decision Tree and Logistic regression models got a 73.33 classification rate in a 10%-90% split. Thirdly, Naïve Bayes and SVM got a rate of 66.67 in 10%-90%. On the other hand, Deep learning and MBERT took the lowest rate of classification 60.00 and 66.67 in a 30%-70% split (MBERT).

TEST EVALUATION TABLE

Case	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Number	Percentage
Actual	0	1	0	1	1	1	0	1	0	0	1	0	1	1	1		
Naïve Bayes	p	f	p	p	p	f	p	p	f	f	p	p	f	p	p	10	66.67
	p	f	p	p	f	f	p	p	f	f	p	f	f	p	p	8	53.33
Logistic Regression	p	p	p	p	p	f	p	p	f	f	p	p	f	p	p	11	73.33
	p	f	p	p	p	f	p	p	f	f	p	f	f	p	p	9	60
Decision Tree	p	p	p	p	p	p	f	p	f	p	f	f	p	p	p	11	73.33
	p	p	f	p	p	p	f	p	f	f	f	f	p	p	p	9	60
Random Forest	p	p	p	p	p	p	p	p	f	p	f	f	p	p	p	12	80
	p	p	p	p	p	p	f	p	f	f	p	f	p	p	p	11	73.33
SVM	p	f	p	p	p	f	p	p	f	f	p	p	f	p	p	10	66.67
	p	f	p	p	f	f	p	p	f	f	p	f	f	p	p	8	53.33
Deep Learnig	f	p	f	p	p	p	f	p	f	f	p	f	p	p	p	9	60
	f	p	f	p	p	p	f	p	f	f	p	f	p	p	p	9	60
transformers	P	f	P	f	f	f	P	P	f	P	f	P	f	P	f	7	46.66
	f	P	f	P	P	P	f	P	f	f	P	p	P	P	P	10	66.67

Figure 4 - 12 Test evaluation table

P = Pass, F = Fail

The table shown in Figure 21 shows the exact performance of each of the models and under the number column, it displays the number of occurrences that are classified correctly (Hate, Offensive, Neither). Random Forest has been able to classify 12 statements correctly out of 15 statements which is the highest number of correct classifications.

	TPR	FPR	TNR
Naïve Bayes	0.75	0.42	0.57
Logistic Regression	0.77	0.33	0.66
Decision Tree	0.72	0.25	0.75
Random Forest	0.8	0.2	0.8
SVM	0.75	0.42	0.57

Table 4 - 4 Confusion matrix

From the Test evaluation table, more data has been harvested to get a better idea of each model's performance. Even though the Logistic regression and Decision Tree models achieved 73.33 rates previously, the True Positive Rate (TPR) of each model is different, and the logistic regression model has a rate of 0.77 which is higher than

the decision tree model's TPR. Random Forest has gained the highest TPR among the highest rated models.

On the other hand, the Decision tree model has achieved a 0.75 True Negative Rate (TNR) which is only less than the Random forest model.

To get a better result and find the most suitable model, accuracy and precision have been calculated based on the test performance.

	Accuracy	Precision
Naïve Bayes	0.66	0.66
Logistic Regression	0.73	0.77
Decision Tree	0.73	0.88
Random Forest	0.8	0.88
SVM	0.66	0.66

Table 4 - 5 Accuracy and precision

Above table 6 displays the accuracy and precision of filtered models. The decision tree and the Random forest models are taken at the highest precision rate of 0.88. Logistic regression and Decision tree models have the second highest accuracy rate of 0.73. The Random forest model has been able to get the highest accuracy of 0.8.

It is evident that for detecting hate speech in Singlish logistic regression, decision tree, and random forest models are more suitable compared to the other models. Among these three models, the **Random forest** has shown better performance with training, testing, and adapting to practical scenarios.

4.4 Technical Considerations

Functional and non-functional requirements are being identified in the **Designing Solution** stage and **Evaluating Existing Models** stage which is discussed under the 3.5 Research Execution Mode in Chapter 3. Technological considerations depend on the functional and non-functional that are considered to be within the system. Other than these considerations, a few more criteria needed to be considered during the planning such as time span, budget, and maintenance.

Within this system, the implementation budget was not considered and the other two were considered. The main functional requirements are,

- The user should be able to insert a text input,
- A model which understands Singlish,
- Identify hate speech accurately,
- Provide a precise result,
- Pool hateful statements for future work,
- UI to conduct a demonstration.

The main non functional requirements are,

- Increased accuracy,
- Increased precision,
- Realtime detection/ process speed,
- Performance,
- Ability to update/ advance

The system itself is a detection model that is trained to detect hate speech in Singlish on Facebook.

To address all these requirements **Python** was used to build the model. **Flask** has been used to build the user interface and display the model performance. Programming languages like Java, C, C++, and C# have their ways of attending to such tasks but Python has some unique characteristics that make Python more suitable for creating this system. Simplicity, understandability, readability, community support, and extended libraries with a rich ecosystem were those characteristics. Python has the ability to create advanced and complex projects using simple codes and structures that are easy to understand and use.

On the other hand, Flask is the web framework of Python that is widely used to build interfaces, web applications, microservices, etc. Flask is selected because it is compatible with and easy to integrate the model and user interface, has concise code to increase performance, and is simple to edit.

4.4.1 Used technologies, Tools, and frameworks

- Programming Language/ s: Python
- Other languages: HTML, CSS
- User Interface: Python and HTML
- UI Framework: Flask
- IDE: PyCharm Enterprise Edition
- Browser: Chrome
- Version Control: Git Hub (repository)/ GitHub Desktop
- Libraries: Pandas, Scikit-learn
- Algorithms: Random forest with TFID vectorizer
- Resource: Facebook
- other: Excel, Miro

Technical considerations related to the functional and non functional requirements are listed above.

4.4.2 Usage of Python and Flask

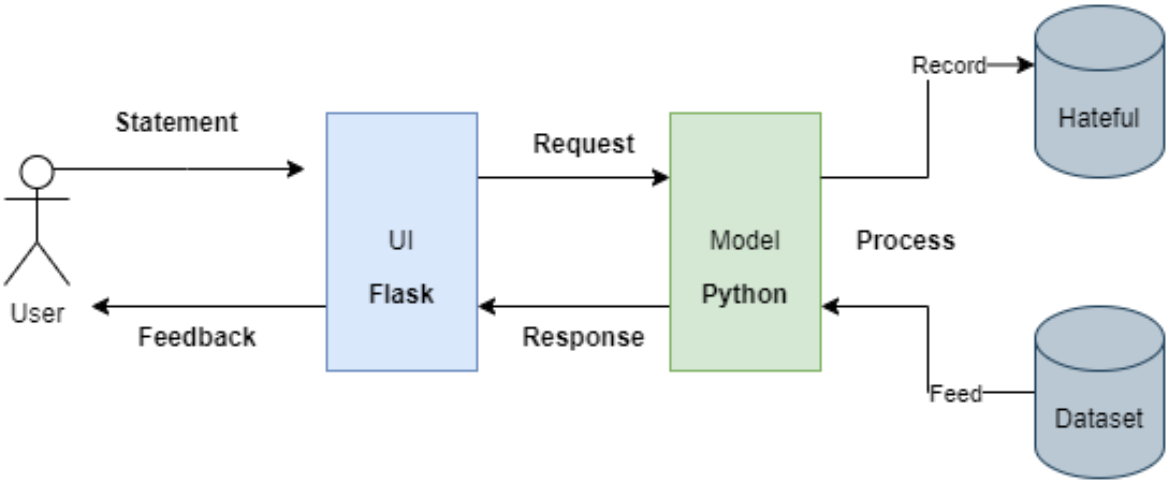


Figure 4 - 13 Integration of Python and Flask

As per the above illustration, python and Flask have an inter-related dependency within the system. The user directly inserts text, and statements using the web interface which has been built using the Flask web framework, and then the request is sent to the model to pre-process and analyze the word or statement and return a response to the user interface and from the interface itself generates a feedback to the user on what has happened to the request.

In the latter part, we could identify the data layer according to the layered architecture and there are two data sources that are uni-directional at this point of the implementation. Hateful is responsible for holding hate texts and statements that are identified using the system. The dataset that is used to train and test the model feeds the model continuously for an accurate result.

4.4.3 Folder Structure

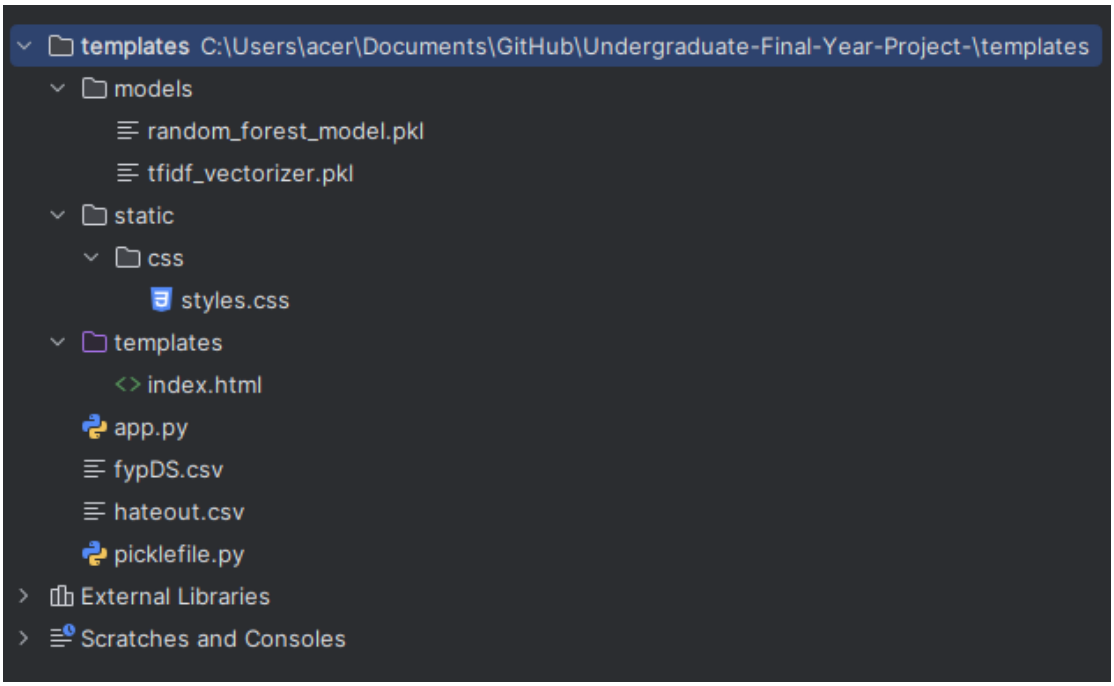


Figure 4 - 14 Folder structure

This is the folder structure that was used to build the system. The code is modularized into sections to increase the maintainability and ease the code.

App.py

It is the Python code that handles the backend of the system. Two APIs were created to submit the comment and route back to the index interface.

```
@app.route('/')
def index():
    return render_template('index.html')

@app.route('/submit', methods=['POST'])
def submit_comment():
    data = request.json
```

```
    user_comment = data.get('comment', '')

    user_comment_tfidf = vectorizer.transform([user_comment])
    prediction = model.predict(user_comment_tfidf)

    if prediction[0] == 1:
        return jsonify({'status': 'blocked'})
    else:
        return jsonify({'status': 'allowed'})
```

this handles the submit action and the validation process using the model.

Picklefile.py

Text preprocessing and model training are done inside the picklefile. Users can change parameters and update the model using this file without causing damage to the rest of the file. Once you run the file two outputs will be saved on the model directory. Which are random_forest_model.pkl and tfidf_vectorizer.pkl.

```
data = pd.read_csv("fypDS.csv")

X = data['text']
y = data['label']
X = X.fillna('')

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.1,
random_state=42)

vectorizer = TfidfVectorizer()

X_train_tfidf = vectorizer.fit_transform(X_train)

model = RandomForestClassifier(random_state=42)

model.fit(X_train_tfidf, y_train)
```

```
with open("models/random_forest_model.pkl", "wb") as model_file:
    pickle.dump(model, model_file)


with open("models/tfidf_vectorizer.pkl", "wb") as vectorizer_file:
    pickle.dump(vectorizer, vectorizer_file)


print("Model and vectorizer have been saved successfully.")
```

Models

The models directory holds the trained model and the vectorizer holds the model when the user runs the app.py.

Once the picklefile.py runs the newest state of the model and the vectorizer are saved to the models directory.

Static

The CSS file that is used as the stylesheet for the index.html file is saved and called from this directory.

Templates

The index.html file is called when the system is running. Embedded script has been used to control the requests and responses. AJAX is used to submit forms without refreshing the whole web interface.

```
function submitComment(event) {
    event.preventDefault();

    const comment = document.getElementById("comment").value;

    fetch('/submit', {
        method: 'POST',
        headers: {
            'Content-Type': 'application/json',
        },
        body: JSON.stringify({ comment: comment }),
    })
    .then(response => response.json())
    .then(data => {
        if (data.status === 'blocked') {

            alert("Comment is blocked because it contains hate speech.");
        } else {

            const commentSection = document.getElementById("comment-section");
```

```
            const newComment = document.createElement("p");
            newComment.textContent = comment;
            commentSection.appendChild(newComment);
        }


        document.getElementById("comment").value = '';
    })
    .catch((error) => {
        console.error('Error:', error);
    });
}
```

4.5 Output and Artifacts

The current study is in phase 2 of development and to achieve phase 3 it requires more data to move along with. Since there is no dataset that has collected hateful statements the system itself creates a data set for future references. Once a statement is detected as hate speech it records that statement in a CSV file that is attached to the system.

All the hateful and offensive comments were recorded there.

```
if prediction[0] == 1:  # If the comment is predicted as hate speech (assuming '1' means hate speech)
    # Store the hate speech comment in the 'hateout.csv' file
    df = pd.DataFrame({'comment': [user_comment]})


    if os.path.isfile("hateout.csv"):
        df.to_csv("hateout.csv", mode='a', header=False, index=False)
    else:
        df.to_csv("hateout.csv", mode='w', header=True, index=False)
```

this is known to be the ultimate dataset of hate speech in Singlish and this will help future studies in this domain and will also help to carry out this to other studies as well.

4.6 Model Construction

In the initial phase, the model was constructed into a single module where each and every function was embedded into one single peach of code which had plenty of lines.

```
import pandas as pd
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report, accuracy_score
import os

def train_random_forest(X_train, y_train):  1 usage   Asirimath Vimukthi *
    # Fill missing values in the training data with empty strings
    X_train = X_train.fillna('')
    vectorizer = TfidfVectorizer()
    X_train_tfidf = vectorizer.fit_transform(X_train)
    model = RandomForestClassifier(random_state=42)
    model.fit(X_train_tfidf, y_train)

    return model, vectorizer

def evaluate_model(model, vectorizer, X_test, y_test):  1 usage   Asirimath Vimukthi *
    # Fill missing values in the test data with empty strings
    X_test = X_test.fillna('')
    # Transform the test data
    X_test_tfidf = vectorizer.transform(X_test)
    # Predict the labels for the test data
    y_pred = model.predict(X_test_tfidf)

    print("Accuracy:", accuracy_score(y_test, y_pred))
    print("Classification Report:\n", classification_report(y_test, y_pred))


if __name__ == "__main__":

    data = pd.read_csv("fypDS.csv")

    X = data['text']
    y = data['label']

    X_train, X_test, y_train, y_test = train_test_split( *arrays: X, y, test_size=0.1, random_state=42)

    model, vectorizer = train_random_forest(X_train, y_train)

    evaluate_model(model, vectorizer, X_test, y_test)
```

Figure 4 - 15 initial model (i)

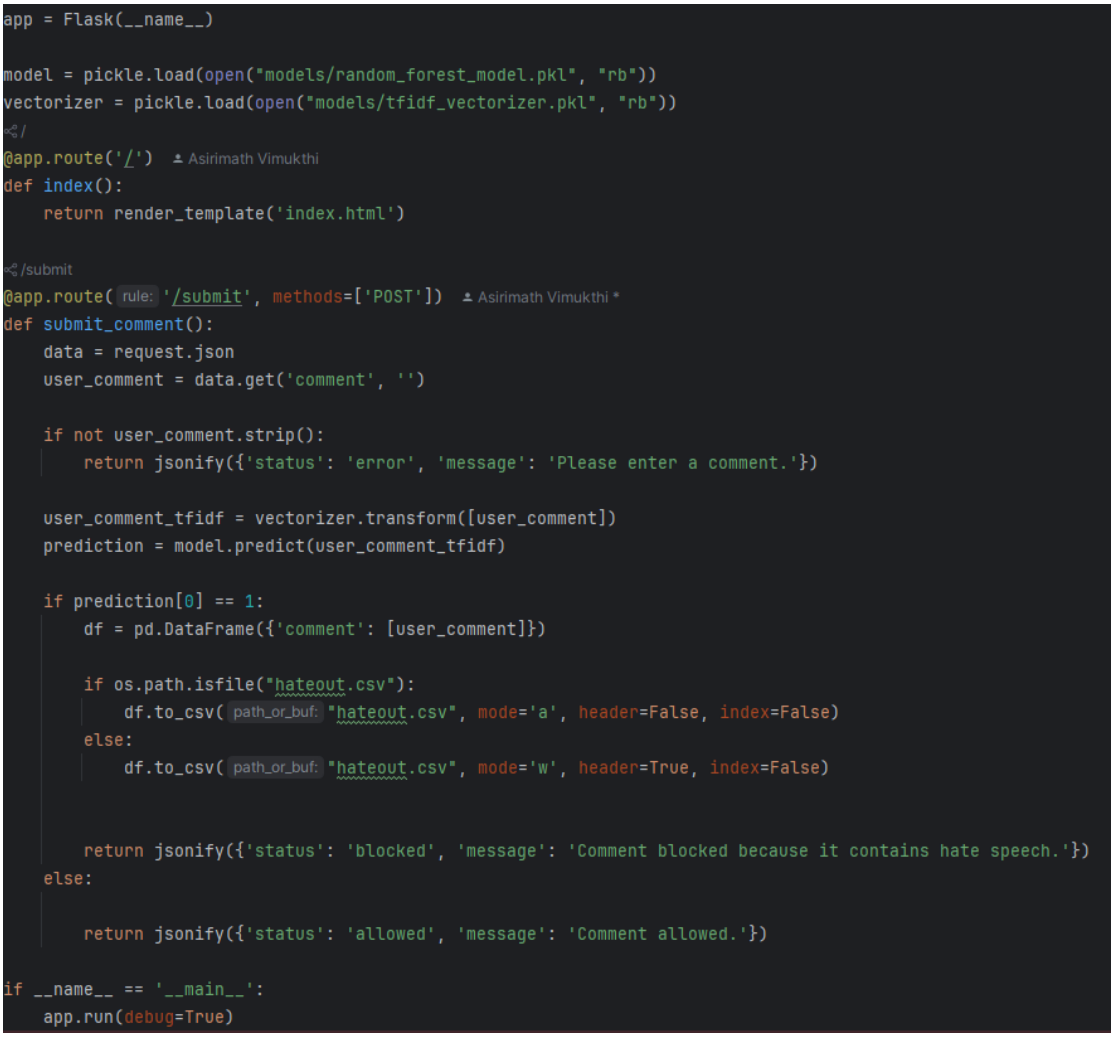
```
while True:
    user_comment = input("Enter a comment to validate (or type 'exit' to quit): ")

    if user_comment.lower() == 'exit':
        print("Exiting the program.")
        break

    user_comment_tfidf = vectorizer.transform([user_comment])

    prediction = model.predict(user_comment_tfidf)
    probabilities = model.predict_proba(user_comment_tfidf)[0]

    print(f"Probability of non-hate speech: {probabilities[0]:.4f}")
    print(f"Probability of hate speech: {probabilities[1]:.4f}")

    if prediction[0] == 1:  # Assuming '1' means hate speech
        print("This comment is classified as hate speech.")
        # Save the comment if it is flagged as hate speech
        df = pd.DataFrame({'statement': [user_comment]})
        if os.path.isfile("hateout.csv"):
            df.to_csv( path_or_buf: "hateout.csv", mode='a', header=False, index=False)
        else:
            df.to_csv( path_or_buf: "hateout.csv", mode='w', header=True, index=False)
        print(f"Flagged comment saved to hateout.csv")
    else:
        print("This comment is not classified as hate speech.")
```

Figure 4 - 16 initial model (ii)

As in the above images, the whole code carries loading data, training, testing data, evaluating the model, and saving the hate comments once they get detected.

Then the code was needed to modularize according to the folder structure that has been mentioned in 4.4.3 Folder Structure.

The code modularization is done to increase the usability and maintainability of the code and also to decrease the potential damages that can occur due to the several changes.



```
app = Flask(__name__)

model = pickle.load(open("models/random_forest_model.pkl", "rb"))
vectorizer = pickle.load(open("models/tfidf_vectorizer.pkl", "rb"))
/
@app.route('/')  Asirimath Vimukthi
def index():
    return render_template('index.html')

/submit
@app.route( rule: '/submit', methods=['POST'])  Asirimath Vimukthi *
def submit_comment():
    data = request.json
    user_comment = data.get('comment', '')

    if not user_comment.strip():
        return jsonify({'status': 'error', 'message': 'Please enter a comment.'})

    user_comment_tfidf = vectorizer.transform([user_comment])
    prediction = model.predict(user_comment_tfidf)

    if prediction[0] == 1:
        df = pd.DataFrame({'comment': [user_comment]})

        if os.path.isfile("hateout.csv"):
            df.to_csv( path_or_buf: "hateout.csv", mode='a', header=False, index=False)
        else:
            df.to_csv( path_or_buf: "hateout.csv", mode='w', header=True, index=False)


        return jsonify({'status': 'blocked', 'message': 'Comment blocked because it contains hate speech.'})
    else:

        return jsonify({'status': 'allowed', 'message': 'Comment allowed.'})

if __name__ == '__main__':
    app.run(debug=True)
```

Figure 4 - 17 app.py (i)

This image shows the modularized code that's handling the back end and also the saving hate comments to the CSV file that has been mentioned in the 4.5 Output and Artifacts. This contains the main method to run the application that runs on top of the Flask.

```
data = pd.read_csv("fypDS.csv")

X = data['text']
y = data['label']


X = X.fillna('')

X_train, X_test, y_train, y_test = train_test_split( *arrays: X, y, test_size=0.1, random_state=42)

vectorizer = TfidfVectorizer()

X_train_tfidf = vectorizer.fit_transform(X_train)

model = RandomForestClassifier(random_state=42)

model.fit(X_train_tfidf, y_train)

with open("models/random_forest_model.pkl", "wb") as model_file:
    pickle.dump(model, model_file)

with open("models/tfidf_vectorizer.pkl", "wb") as vectorizer_file:
    pickle.dump(vectorizer, vectorizer_file)

print("Model and vectorizer have been saved successfully.")
```

Figure 4 - 18 picklefile.py

This is the picklefile which holds the model and also training the model using the vectorizer is done from this module. All the changes related to the model can be done through the picklefile.py.

CHAPTER 5 – DISCUSSION AND CONCLUSION

5.1 Discussion

The purpose of this study is to improve the platform safety of Facebook and to ensure the user safety of Facebook users in Sri Lanka by providing a solution to capture and eliminate hate speech that happens using Singlish language. The literature review and pre-study have identified that hate speech is vastly growing and the number of victims is increasing day by day due to this. It has been identified as a critical issue around the world. In the current study, has identified the importance of identifying and eliminating hate speech from social media platforms because of the increase in internet usage and social media platform usage.

The existing system addresses this issue by taking a different approach. And available system does not support Singlish since it has not been identified as a language. Only the stand-alone or accredited languages are being tested and taken into consideration for these classifications. This system has multiple limitations. Such as this hate speech process starts once the user has reported and after the evaluation by the security team such terms will be blocked and necessary actions will be taken against the users who posted such. The reason behind these limitations is ‘Freedom of expression’. People who are known to be hate spreaders use this to protect themselves from getting captured.

The newly built system addresses this issue using machine learning to identify hate speech proactively where the model identifies the statement as hate speech once it is submitted and it checks in real-time before it gets posted. A limitation is the contextual identification of the test. Since this is the first system that has used Singlish to identify hate speech. This can be achieved once the current study is continued for the future with the necessary changes to build a better enhanced and advanced ML model. Also since this is the first of its own, there can be scenarios in which the model is able to identify all the hate content since in the test phase it has only given 0.80 accuracy and 0.88 precision. Conducting multiple iterations will provide more accurate results.

The new system has been developed considering two labels and 3 categories of statements. More context could be generated over the next phases to cater to the

untouched area of this domain. This system is mainly focused on Facebook since that is the most commonly used social media platform in Sri Lanka. The system has not been implemented on Facebook itself since this is a suggested system that cannot be used on actual Facebook. Therefore, to get an idea for both technical and non-technical people the system can be used using the demo user interface.

The model is evaluated with six other models from different categories to conduct fair research and also used a common test iteration at each of the model's optimum levels. These results show the validity and suitability of the selected method. These results may vary with certain environmental changes such as the size of the dataset, number of resources, availability of resources, etc.

In this discussion, it has given the overall declarations of pre-study to implementation and limitations that have been identified. Also, the challenges and solutions that are taken to overcome them.

5.2 Conclusion

In conclusion, The system is built to ensure the safety of the platform and its users from hate speech in Singlish that happens on Facebook by providing a proactive solution. The domain that has followed provided the previous studies related to hate speech using multilingual or bilingual languages and this has given the inspiration to build this system. Multi-lingual is using multiple languages and Singlish is using English to provide the sound and the meaning of the Sinhala language. The comprehensive literature review has given and paved the path to provide a successful implementation to detect hate speech in Singlish on Facebook.

5.3 Future Work

For future work, there are two important pathways to improve the effectiveness and accuracy of the hate speech detection system. The first direction involves optimizing the model to achieve a deeper contextual understanding. This can be accomplished by expanding the current dataset and integrating it with other related Singlish datasets. By doing so, the system will be able to capture more nuances and language variations specific to Singlish, such as code-switching between Sinhala and English, slang, and informal expressions. With this larger and more representative dataset, the model

could be fine-tuned to understand subtle differences in meaning, ultimately improving its ability to detect hate speech with higher precision.

The second potential approach is to rebuild the system using a more advanced machine learning model, such as MBERT or a deep learning architecture, after collecting a sufficient amount of data to create a comprehensive dataset. This would allow for the development of an unsupervised or semi-supervised hate speech detection system that is specifically trained in Singlish. A model of this nature would not rely solely on predefined labels or supervised learning, but rather would learn from the patterns within the data itself. As a result, it would have the capacity to identify hate speech more intelligently, by understanding the broader context in which certain words or phrases are used.

Such a model would be capable of recognizing complex speech patterns like sarcasm, criticism, natural expressions, and slang, which are often misunderstood by traditional models. This context-aware system could intelligently discern between harmful content and benign statements that might otherwise be flagged as hate speech. In the future, this advancement would contribute to a more robust and sophisticated hate speech detection tool, capable of effectively moderating content on platforms like Facebook, where Singlish is commonly used. This would enhance the platform's ability to reduce online hate while maintaining the freedom of natural expression.

6. 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.

Datareportal (2024) *The state of digital in Sri Lanka in 2024*. Available at: https://datareportal.com/reports/digital-2024-sri-lanka.

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.

Neville Lahiru (2024) 'We need to pay more attention to cyberbullying'. Available at: https://readme.lk/we-need-to-pay-more-attention-to-cyberbullying/.

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).

7. APPENDIX

Front-End Implementation

Figure 7 - 1 User interface

This is the created user interface to demonstrate the functionality of the model.

```html
<!DOCTYPE html>
<html lang="en">
<head>
    <meta charset="UTF-8">
    <meta name="viewport" content="width=device-width, initial-scale=1.0">
    <title>Hate Speech Detection</title>
    <link rel="stylesheet" href="{{ url_for('static', filename='css/styles.css') }}">
    <script>

        function submitComment(event) {
            event.preventDefault();

            const comment = document.getElementById("comment").value;

            // Send the comment to the server using AJAX
            fetch('/submit', {
                method: 'POST',
                headers: {
                    'Content-Type': 'application/json',
                },
                body: JSON.stringify({ comment: comment }),
            })
            .then(response => response.json())
            .then(data => {
                if (data.status === 'blocked') {
                    // Display the popup if the comment is hate speech
                    alert("Comment is blocked because it contains hate speech.");
                } else {
                    // Add the comment to the list of comments if it's not hate speech
                    const commentSection = document.getElementById("comment-section");
                    const newComment = document.createElement("p");
                    newComment.textContent = comment;
                    commentSection.appendChild(newComment);
                }

                // Clear the input field
                document.getElementById("comment").value = '';
            })
            .catch((error) => {
                console.error('Error:', error);
            });
        }
```

Figure 7 - 2 HTML 1

HTML source behind the user interface and the AJAX-based script to connect both back end and the front end.

```
    </script>
</head>
<body>

    <h1>Demo Interface - Hate Speech Detection System</h1>

    <div id="iframe-container">
<iframe src="https://www.facebook.com/plugins/post.php?href=https%3A%2F%2Fwww.facebook.com%2Fphoto%2F%3Ffbid%3D372063196
        </div>


    <form id="comment-form" onsubmit="submitComment(event)">

        <textarea id="comment" name="comment" rows="4" cols="50" placeholder="Enter your comment"></textarea><br>
        <button type="submit">Submit</button>
        <button type="reset">Cancel</button>
    </form>


    <!-- Section to display comments -->
    <h2>Comments</h2>
    <div id="comment-section"></div>

</body>
</html>
```

Figure 7 - 3 HTML 2

Enabling users to input a text as a comment. Submit action and cancel action has provided to the user.

```css
body {
    font-family: Arial, sans-serif;
    margin: 20px;

}

textarea {
    width: 64.5%;

}

button {
    margin-top: 10px;
}

#comment-section {
    margin-top: 20px;
    border-top: 1px solid #ddd;
    padding-top: 10px;
}

#comment-section p {
    padding: 5px 0;
    border-bottom: 1px solid #ddd;
}

h1 {
    text-align: center;
}
#comment-form {
    position: relative;
    left: 273px
;

}
#iframe-container {
  display: flex;
  justify-content: center;
  align-items: center;
}
```

Figure 7 - 4 CSS

Style sheet on behalf of creating the user interface.

Model Implementation

Model implementation and evolution from the initial phase to the current phase have been discussed in the 4.6 Model Construction.

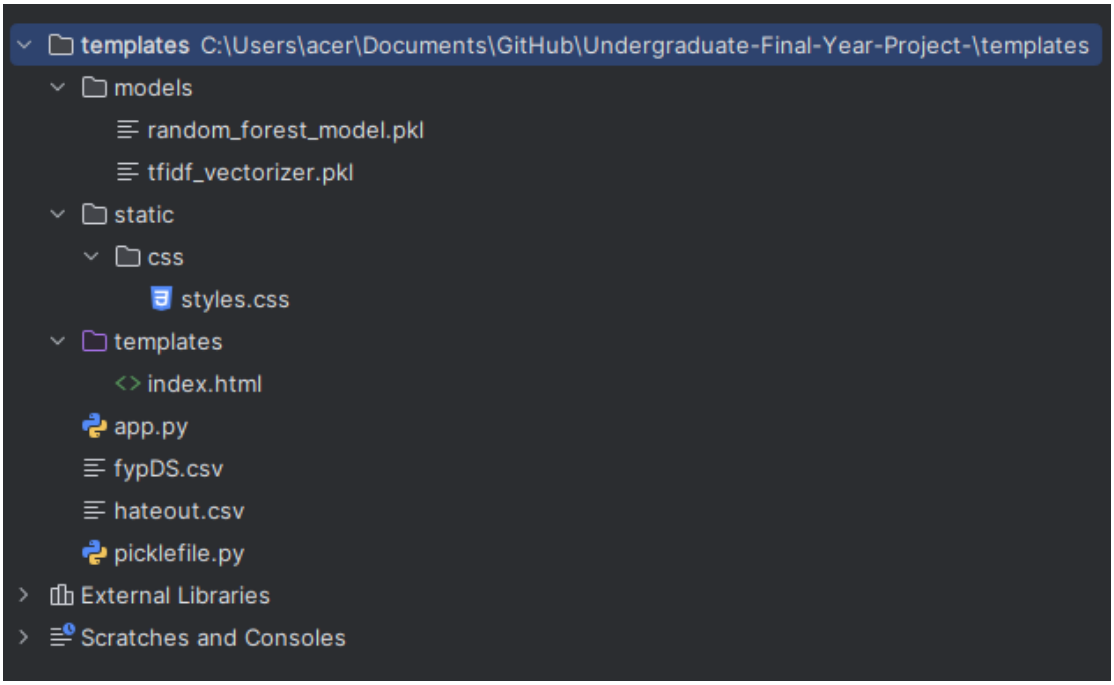


Figure 7 - 5 Structure

```python
app = Flask(__name__)

model = pickle.load(open("models/random_forest_model.pkl", "rb"))
vectorizer = pickle.load(open("models/tfidf_vectorizer.pkl", "rb"))
/
@app.route('/')  Asirimath Vimukthi
def index():
    return render_template('index.html')

/submit
@app.route( rule: '/submit', methods=['POST'])  Asirimath Vimukthi *
def submit_comment():
    data = request.json
    user_comment = data.get('comment', '')

    if not user_comment.strip():
        return jsonify({'status': 'error', 'message': 'Please enter a comment.'})

    user_comment_tfidf = vectorizer.transform([user_comment])
    prediction = model.predict(user_comment_tfidf)

    if prediction[0] == 1:
        df = pd.DataFrame({'comment': [user_comment]})

        if os.path.isfile("hateout.csv"):
            df.to_csv( path_or_buf: "hateout.csv", mode='a', header=False, index=False)
        else:
            df.to_csv( path_or_buf: "hateout.csv", mode='w', header=True, index=False)


        return jsonify({'status': 'blocked', 'message': 'Comment blocked because it contains hate speech.'})
    else:

        return jsonify({'status': 'allowed', 'message': 'Comment allowed.'})

if __name__ == '__main__':
    app.run(debug=True)
```

Figure 7 - 6 Flask app

```
data = pd.read_csv("fypDS.csv")

X = data['text']
y = data['label']


X = X.fillna('')

X_train, X_test, y_train, y_test = train_test_split( *arrays: X, y, test_size=0.1, random_state=42)

vectorizer = TfidfVectorizer()

X_train_tfidf = vectorizer.fit_transform(X_train)

model = RandomForestClassifier(random_state=42)

model.fit(X_train_tfidf, y_train)

with open("models/random_forest_model.pkl", "wb") as model_file:
    pickle.dump(model, model_file)

with open("models/tfidf_vectorizer.pkl", "wb") as vectorizer_file:
    pickle.dump(vectorizer, vectorizer_file)

print("Model and vectorizer have been saved successfully.")
```

Figure 7 - 7 Picklefile code

Output

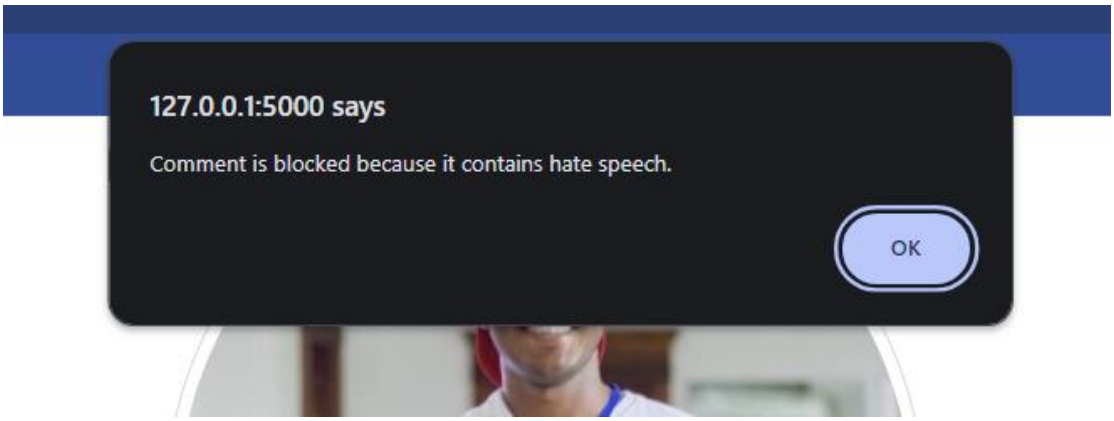


Figure 7 - 8 Hate detected notification

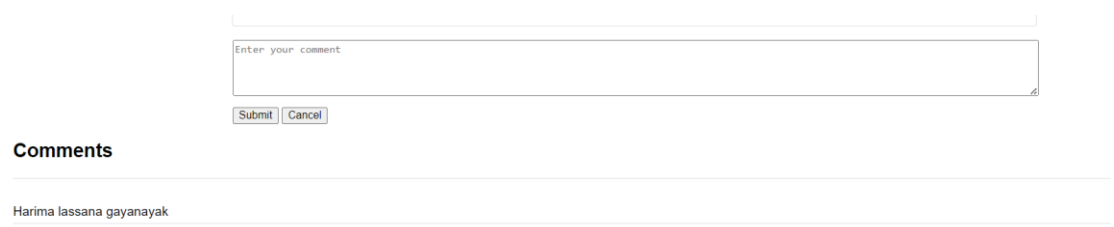


Figure 7 - 9 Posted non hate comment

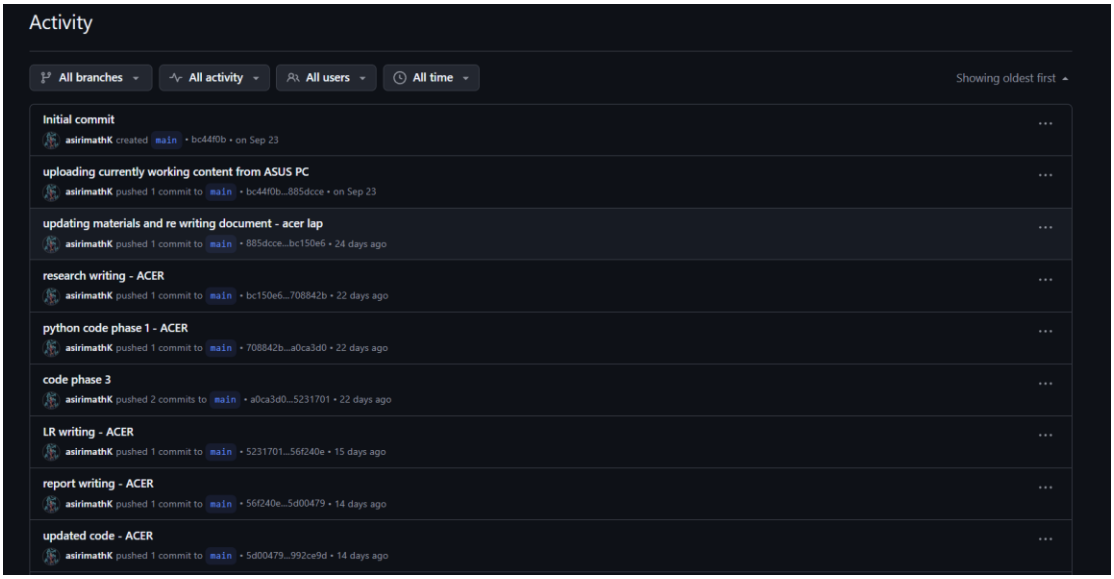


Figure 7 - 10 Git commits 1

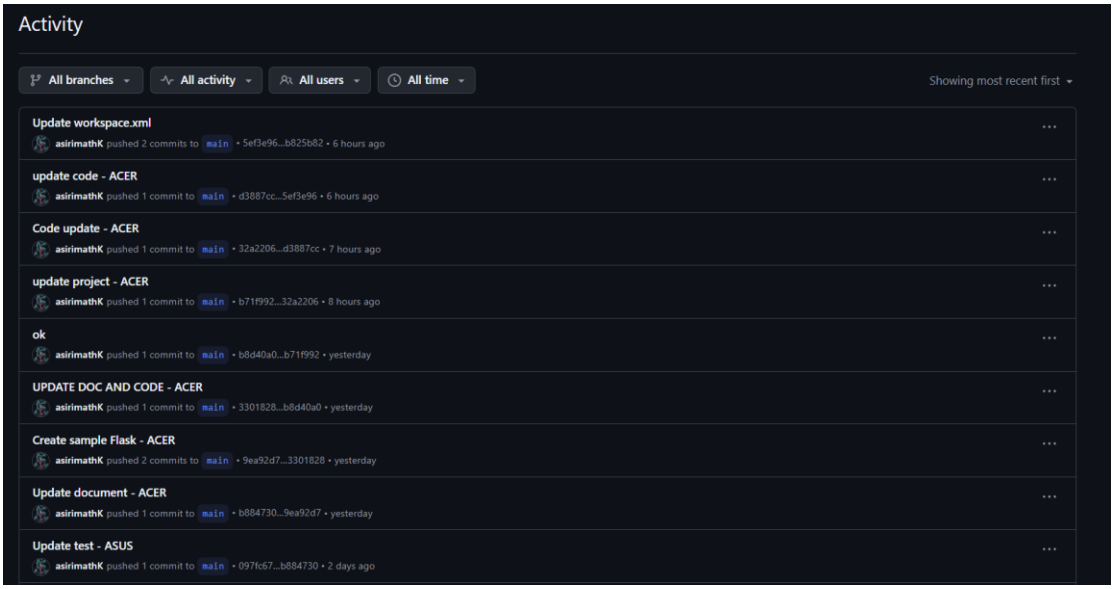


Figure 7 - 11 Git commits 2

Project Repository: https://github.com/asirimathK/Undergraduate-Final-Year-Project-.git