Multilingual Android Forensic Tool for Cyberbullying Investigations on Facebook and Instagram

Komuthu Damya Mabulage – CB010187

Submitted to the

Computing School

in partial fulfillment of the requirements for the Degree of

Bachelor of Science in

Cyber Security (Hons)

Supervised By:

Mr. Jude Mayuran

Staffordshire University

June 2025, Colombo

Abstract

This research project is focused on developing a CLI-based Android forensic tool for social media applications. This research project will mainly focus on two widely used social media platforms in Sri Lanka, which are Facebook and Instagram. This tool can extract data from the above-mentioned social media platforms, process the data with NPL for analyzing in Sinhala and English, detect hate speech leading to cyberbullying and cybercrimes in both languages with a keyword filtering feature, and finally provide a detailed report that can be used as digital evidence in cybercrime investigations in Sri Lanka. This is helpful in cybercrime investigations in Sri Lanka, which is vastly spreading throughout the years, victimizing the youth mainly. Moreover, this research is mostly focused on cyberbullying and harsh speech on social media, which are quite common cybercrimes in Sri Lanka at present.

Declaration

I declare that, to the best of my knowledge and belief, this paper does not contain any previously written or published material of my own or any other person, with the only exception of situations where proper reference is provided within the text. Moreover, it does not include any previously submitted material for a degree or diploma at any university without acknowledgment.

Signature of Candidate : - 

Date : - 15th June 2025

Name of Candidate : - Komuthu Damya Mabulage

Signature of Supervisor : - 

Date : - 15th June 2025

Name of Supervisor : - Mr. Jude Mayuran

Acknowledgements

First, I would like to extend my heartfelt appreciation to my supervisor, Mr. Jude Mayuran, for his amazing guidance and support throughout the preparation of this project proposal. Moreover, I'm thankful for the invaluable advice from my lecturers at APIIT, and I'm grateful for all the invaluable support of my family and friends, understanding and motivating me frequently. Last, but not least, I would like to express my gratitude to all those who helped and encouraged me so far, both professionals and friends.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Full Form
AI	Artificial Intelligence
ADB	Android Debug Bridge
BERT	Bidirectional Encoder Representations from Transformers
CLI	Command Line Interface
CNN	Convolutional Neural Network
DL	Deep Learning
GAN	Generative Adversarial Network
GUI	Graphical User Interface
ML	Machine Learning
NLP	Natural Language Processing
PDF	Portable Document Format
RNN	Recurrent Neural Network
SRS	Software Requirements Specification
SSIM	Structural Similarity Index
SL	Sri Lanka
SLCERT	Sri Lanka Computer Emergency Readiness Team
NCPA	National Child Protection Authority
CCID	Criminal Investigation Department

CHAPTER 01 : INTRODUCTION

1.1. Chapter Overview

This chapter outlines the research background by defining the reason that there is an actual requirement for a forensic tool which is capable of dealing with cyberbullying in specific digital domain of Sri Lanka. This examines the way mobile phones and social media are currently a fundamental component of everyday life, especially for young adults, and how this has contributed to an increase in cybercrimes such as online harassment. Moreover, the chapter also discusses the shortcomings of current technologies, such as their inability to support Sinhala and code-mixed content, and explains how this issue creates a barrier for local investigations. Furthermore, it explains problem statement of this research, the inspiration for this project, and the existing approaches that have already been done in the field. Especially, it describes the research gaps, overall goal, main objectives, and obstacles associated with developing a tool which is both intelligent yet basic for ordinary law enforcement to use.

1.2. Introduction

At present technology has rapidly transformed the day-to-day life of people, all around the world including Sri Lanka. As a developing country, Sri Lanka has quickly embraced this digital shift, especially with the growth of smartphones and internet services. Among mobile devices, Android smartphones dominate the market, making them the most commonly used mobile phone operating system in the country. However, by now these mobile phones have become essential, not only for communication and entertainment but also for education, business, and public services almost everything.

After the COVID-19 pandemic outbreak the digital transition of the country in which the schools, workplaces, and government services shifted to online, even younger generations became heavily dependent on electronic devices. With that the usage of social media platforms such as Facebook, Instagram, WhatsApp, and TikTok dramatically increased during this time. In fact, Facebook continues to be the most widely used social media platform in Sri Lanka, with millions of active users, especially

among those aged 18 to 35. While social media has played a positive role in connecting people, it has also opened the doors towards new risks most notably, cyberbullying.

Cyberbullying has become one of the most common forms of cybercrime in Sri Lanka, often targeting teenagers and young adults. Offenders use hate speech, emotional manipulation, and even share personal content to harass victims online. According to local statistics, a considerable number of cyberbullying cases are unreported, especially in rural areas, due to fear and lack of trust in the legal systems. Even with national cybercrime units like SLCERT and CCID in place, the lack of localized tools makes it difficult for authorities to effectively investigate these incidents especially when this content is written in Sinhala and in Sinhala-English code-mixed language.

Furthermore, this research addresses that gap by proposing a multilingual Android forensic tool designed specifically for the Sri Lankan context. Unlike most existing forensic solutions that are costly and access to local language is limited, this tool focuses on extracting and analyzing social media content from Facebook and Instagram on Android devices. It integrates natural language processing (NLP) techniques to detect cyberbullying in both Sinhala Unicode and code-mixed content, and it generates detailed, structured reports that can be used in digital investigations. The tool is cost-effective, simple to use, and designed to assist Sri Lankan law enforcement in handling the growing threat of cyberbullying with language-aware intelligence and forensic accuracy.

1.3. Problem Background

At present the rate of cybercrimes in Sri Lanka is on the rise, in which only about 30% of cybercrimes are officially reported while the rest 70% are the unreported cybercrime cases, most commonly in rural areas of the country like Monaragala (Heshan Maduranga, 2024). There are three main institutions in Sri Lanka to carry out cybercrime investigations namely Sri Lanka Computer Emergency Readiness Team (SLCERT), Computer Crime Investigation Division (CCID) and National Child Protection Authority (NCPA)(Sampath, 2023). These units carry out investigations using various forensic tools and these forensic tools are expensive.

Moreover, most forensic tools are not quite efficient with the Sri Lankan mother tongue, Sinhala which is a complex language to analyze. Additionally, at present the new generation do not use the Sinhala letters to type the message in Sinhala, they use English letters to type Sinhala words according to the pronunciation of the Sinhala word, mixing both languages which is even more complex (Muthuthanthri and Smith, 2024).

1.4. Problem Statement

The inability of existing forensic tools to detect and analyze Sinhala and Sinhala-English code-mixed cyberbullying content on Android-based social media platforms severely limits the effectiveness of cybercrime investigations in Sri Lanka. Most tools are expensive, foreign-developed, and languages are incompatible, creating a critical gap in local forensic capability. Furthermore, most exciting approaches talk about the technical and the practical side of the tools so accordingly all the tools focus on detecting but do not focus on generating reports based on the analysis reports.

1.5. Problem Definition

This situation up brings a need for a digital forensic tool to analyze, filter and detect harsh speech, cyberbullying attempts in multilingual languages like Sinhala and Sinhala-English code-mixed language with low budget specifically customized for Sri Lanka. Furthermore, this tool must be capable of extracting data from social media platforms, analyzing Sinhala and code-mixed language inputs, detecting cyberbullying patterns, and producing clear, usable evidence for law enforcement investigations.

1.6. Motivation

The motivation behind this research roots from the growing digital threat to Sri Lankan youth and the lack of tools available to effectively fight against it. The identified gaps in existing systems where language limitations and high costs prevent effective

investigations. As Sinhala language which is a unique and specialized to Sri Lanka is not widely supported in most NLP models, and even Sinhala-English transliterated language forms are rarely understood, a specialized solution is required to suit the unique Sri Lankan context. This project also aims to reduce the burden on law enforcement officers by providing a CLI-based system which is user-friendly and adaptable to their investigative workflows.

1.7. Existing Work

Citation	Summary	Limitation	Contribution
(Fernando and Deng, 2023)Cellebrite (2025)	A leading forensic tool for data extraction from mobile devices, including social media apps like Facebook and Instagram.	Lacks support for Sinhala or code-mixed language analysis; high cost.	Useful for initial data extraction from Android devices.
Oxygen Forensics (2025)	This tool offers detailed extraction of multimedia, chat logs, and metadata from social media platforms.	Not optimized for NLP or detection of hate speech in local languages.	Help in retrieving raw evidence efficiently.
Samarasinghe et al. (2020)	This implements CNN and FastText models to detect hate speech in Sinhala Unicode text.	Model performance is limited by small and imbalanced dataset.	Demonstrated the feasibility of machine learning for Sinhala text classification.
Muthuthanthri & Smith (2024)	This approach applies BERT for detecting hate speech in Sinhala-English code-mixed content on Facebook.	Model accuracy drops with minor and sarcastic expressions.	Showed strong potential of transformer models for code-mixed content.

Ruwandika & Weerasinghe (2018)	Early work on hate speech detection using SVM for Sinhala comments.	Limited dataset and applicability to real-world forensic workflows.	Provided foundational dataset and binary classification baseline.
Abu Hweidi et al. (2023)	This approach conducts forensic investigation on social media apps with NIST methodology to extract and analyze digital evidence.	Focused more on extraction than content analysis; no multilingual processing.	Provided a structured forensic framework applicable to Facebook and Instagram investigations.
Chang & Yen (2020)	This focuses on forensic extraction of Facebook Messenger data from Android phones.	Lacks hate speech detection and language adaptability.	Useful for understanding app-specific data structures and recovery techniques.
Menahil et al. (2021)	This performs forensic analysis across multiple social media platforms including metadata recovery and message tracking.	Did not incorporate AI/ML techniques for content analysis or language-specific detection.	Outlined a broad forensic workflow for social network investigations.
Fernando & Deng (2023)	This enhances the detection accuracy for Sinhala hate speech using custom NLP techniques and improved feature extraction.	Focused on Sinhala only, no support for code-mixed or multilingual data.	Improved precision of local language hate speech classification.

Chathurangi et al. (2024)	This integrates detection of spam, bot activity, and cyberbullying on Sri Lankan social media platforms.	Rule-based detection with limited flexibility and no deep learning.	Proposed a context-aware local framework targeting multiple social media threats.

Table 1 - Chapter 1: Table of Existing Work

1.8. Research Gap Identification

a. Lack of Sinhala language support in forensic tools

Most popular social media forensic tools like Cellebrite, Oxygen Forensic Suite, and Magnet AXIOM are strong tools which are highly effective in social media forensic analysis, but these tools are not much effective in Sihala language which is the mother tongue of Sri Lanka. This creates a challenge for the investigators to carry out investigations, collect, analyze and document all the findings.

Justification – Developing the Android Forensic tool for social media applications like Facebook and Instagram with the localized feature of detecting and analyzing evidence in Sinhala which is essential for addressing cybercrimes and carrying out strong cybercrime investigations in Sri Lanka.

b. High cost & complexity of existing tools

Strong digital forensic tools are expensive to buy and maintain mostly. Moreover, some tools are complex and quite inconvenient to use. Specially in Sri Lanka, with an unstable economy, daily rising cybercrimes, newly emerging cybercrime trends and limited number of experienced cyber security, digital forensics personnel using these types of tools are inconvenient.

Justification – The proposed tool will be customized to fit the Sri Lankan context. Localized tools are comparatively not costly, and maintenance cost is low. In some cases, it will not cost at all. Moreover, simplicity of the cli based structure of the

forensic tool will be easy for the personnel with less experience to handle proposed tool.

c. Limited & imbalanced datasets

Most of the existing tools and research have the common limitation of limited dataset availability and imbalanced datasets. This limitation affects the ability of the tools to generalize and detect hate speech accurately. Sinhala language datasets are rare and limited so to raise the accuracy of the tool, availability of much enough datasets are crucial.

Justification – Expanding the available datasets will help to enhance the accuracy and the reliability of the proposed tool, providing valuable and accurate digital evidence for crime investigations.

d. Poor reporting/documentation in tools

Detailed documentation of the findings or the digital evidence collected is helpful in crime investigations. Most of the tools are focused on the performance of the tool and not much concerned about the small details like documentation of the digital evidence.

Justification - The proposed tool will have the capability to detect hate speech, extract digital evidence from social media platforms like Facebook and Instagram and create detailed reports for crime investigations including all the findings and detected cyber bullying attempts. Moreover, this approach will provide significant value to law enforcement in Sri Lanka.

1.9. Contribution to Body of Knowledge

This study elaborates a valuable contribution to the fields of digital forensics and multilingual hate speech detection by developing a localized Android forensic tool tailored for the unique Sri Lankan context. Although the existing researches and approaches focus on hate speech classification in Sinhala and code-mixed Sinhala-English content (Fernando and Deng, 2023; Muthuthanthri and Smith, 2024), this capability into a forensic framework designed to extract and analyze evidence directly

from mobile social media platforms has not being integrated. Therefore, a combination of Natural Language Processing (NLP), transformer-based architectures (BERT), and Android forensic workflows in a CLI-based system accessible to authorized law enforcement personnel addresses this by the proposed tool.

Unlike existing commercial forensic tools like Cellebrite and Magnet AXIOM which are powerful and expensive but lack the unique Sinhala-language support (Chang and Yen, 2020; Abu Hweidi et al., 2023), this tool is customized to detect cyberbullying and hate speech in both Sinhala Unicode and Sinhala-English transliterated formats, extracted from applications like Facebook and Instagram. Moreover, the tool offers a structured, auto-generated PDF report to support evidence presentation in legal contexts as well and this bridges a key usability gap identified in existing solutions (Chathurangi et al., 2024)

Furthermore, this research project elaborates the application of multilingual BERT in low-resource language settings specially such unique and complex, showing how fine-tuned transformer models can be used to improve cybercrime investigations in diverse language environments. It further contributes to the a special area of code-mixed language processing in NLP, where Sinhala-English social media content has previously received limited attention (Ruwandika and Weerasinghe, 2018).

Apart from the technical side of the solution, the project delivers practical value by offering a scalable, affordable solution suited for the unique socio-economic context of Sri Lanka, where forensic resources are frequently limited and professional capacity is under development. Therefore, the research not only advances academic understanding but also strengthens the operational capacity of digital forensic efforts in developing countries.

1.10. Research Challenge

1.10.1. The Challenges of Research

The main challenge of this research the development of this android forensic tool that can extract social media data from the mobile phone and accurately detect

cyberbullying through hate speech in Sinhala and Sinhala-English code-mixed content, while also fitting the unique context of Sri Lanka. At first glance, this might seem like a simple classification task but in reality, it's much more complex. Moreover, there are several practical barriers that make the system difficult to be designed and implemented effectively:

- Unique Language Complexity

Sinhala in the mother tongue of Sri Lanka and it is a native language limited only to Sri Lanka unlike English, Sinhala is a complex language with complex grammar, and letters. Furthermore, at present Sinhala in mixed with English for the convenience in communication through online platforms. Therefore, it makes the language even more complex, when Sinhala combined with English in transliterated and code-mixed formats, the structure becomes unpredictable because there is no standard alphabet or standard set of spellings. With that people often switch between mid-sentence languages, use informal grammar, and write in creative, unstructured ways. These variations make it hard for NLP models to consistently understand the true meaning behind a message especially when detecting hidden threats and sarcasm.

- Lack of Quality Datasets

One of the biggest challenges is the limited availability of annotated Sinhala and code-mixed hate speech datasets. Most available datasets are small, imbalanced, and do not reflect the real diversity of how people express themselves online. Without rich, high-quality dataset, even the most advanced models will struggle to make accurate predictions. As mentioned earlier all most all the existing data sets are small and limited to 5000 – 6000 comments. The accuracy of the tool totally depends on these datasets.

- Balancing Accuracy with Usability

Another key challenge is making the tool powerful yet simple to use. Law enforcement officers in Sri Lanka often lack technical training in digital forensics and machine learning. Therefore, the tool must deliver reliable results through a clean, command-line interface, without requiring users to understand the underlying ML

models and preprocessing steps. Therefore, maintaining this balance between usability and technical depth is not an easy task.

- Resource Constraints and Deployment Barriers

Developing countries like Sri Lanka often face limitations in terms of hardware, internet access, and funding for expensive digital forensics. therefore, cloud-based and high-cost solutions are not practical for the Sri Lankan context. With that the tool must be lightweight, run locally, and function well under these existing resources without compromising the performance.

1.10.2. Research Questions

In this research providing a solution for this issue in Sri Lanka, we specifically hope to answer the following research questions;

- Identification of techniques for effective data extraction and analysis of social media data.
- Analysis of the social media platforms chosen to be addressed in the research and justification for the selection.
- Investigation of the cybercrimes most commonly reported on Facebook and Instagram in Sri Lanka.
- Comparison of the accuracy of the proposed tool in detecting hate speech in Sinhala and English with the existing hate speech detection systems.
- Evaluation of the effectiveness of NLP (Natural Language Processing) models in detecting cybercrime evidence in Sinhala and English.
- Optimization of the structure of the generated PDF reports to ensure clarity, usability and easy interpretation by law enforcement and legal professionals.

1.11. Research Aim

The key aim of this research is to design and develop a multilingual android forensic tool that can assist in cyberbullying investigations by detecting hate speech and offensive content in Sinhala and code-mixed Sinhala-English formats. Moreover, the

tool is specialized to support Sri Lankan law enforcement by extracting relevant social media data from platforms like Facebook and Instagram and analyzing it using NLP techniques. It further aims to generate structured reports that are understandable and usable in legal contexts on crime investigations. With the focus on language adaptability, the user-friendly tool interface, and forensic relevance, this approach creates a practical, low-cost solution for facilitate cybercrime investigations in a resource-limited but language complex environment.

1.12. Research Objectives

2.1.1. Primary objectives

The main objective of this project is to develop an Android social media forensic tool to help in detecting cybercrime evidence in both Sinhala and English tailored to unique requirements in Sri Lanka.

2.1.2. Other specific objectives

Apart from the primary objective there are few more special other objectives of this project, they are;

- To develop the tool to extract data from popular social media platforms like Facebook and Instagram on Android devices.
- To integrate Natural Language Processing (NLP) capabilities to analyze and process content in both Sinhala and English languages.
- To further develop the forensic tool to detect cyberbullying, hate speech, and other cybercrimes from extracted social media data.
- To enable filtering of data based on user-defined keywords and categories to focus on specific investigations.
- To generate detailed and structured PDF reports summarizing:
    - Extracted data.
    - Detected hate speech and cyberbullying instances.

    - Keyword-related evidence.

- To ensure the tool addresses unique cybercrime patterns and trends in Sri Lanka.
- To develop a user-friendly and easy-going interface with much less complexity suitable for law enforcement officers.

1.13. Chapter Summary

Overall, this chapter provides an overview of why this research is relevant, especially within the Sri Lankan context, where cyberbullying is on the rise and existing forensic techniques fail short primarily due to language barriers and high cost. Moreover, it describes the way smartphones and social media usage has increased, the challenges that investigators confront, and the reason that this technology is essential. The chapter further explains the research goal, important objectives, and key challenges involved in developing a system that is both effective and simple for local law enforcement to use.

CHAPTER 02 : LITERATURE REVIEW

2.1. Chapter Overview

This chapter explores the background and existing research related to cyberbullying, digital forensics, and language challenges in Sri Lanka. It looks into current trends on platforms like Facebook and Instagram, the limitations of existing forensic tools, and the requirement for local language focused solutions. By reviewing related studies and systems, this chapter sets the stage for why a localized, multilingual forensic tool is essential for effectively addressing cyberbullying in the Sri Lankan context.

2.2. Concept Map

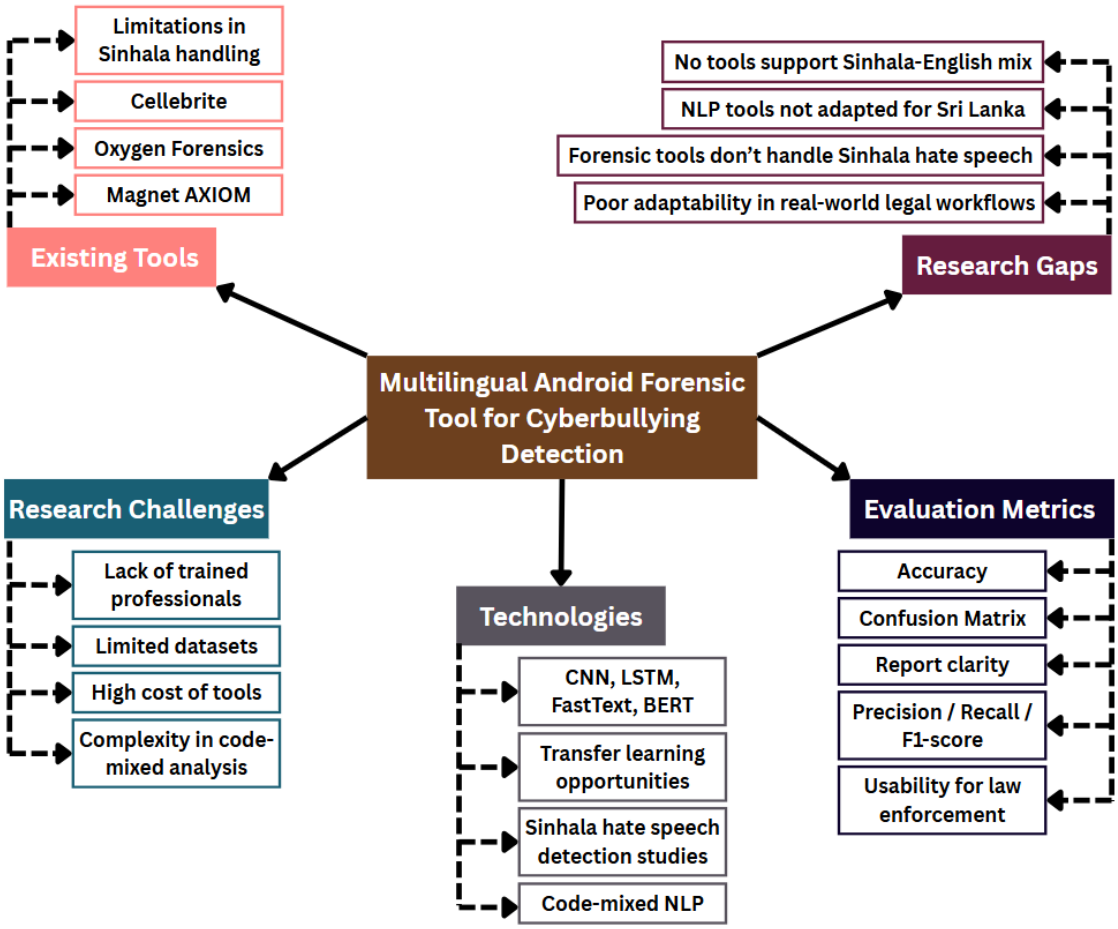


Figure 1 - Chapter 1: Concept Map

2.3. Problem Domain

2.3.1. Current Cybercrime Trends in Sri Lanka

Currently, with the worldwide evolution of technology, cybercrimes have become a common problem all over the world. Similarly, Sri Lanka witnessed a notable increase in cybercrimes with various emerging trends such as online financial fraud, phishing attacks, social media scams, fake websites, cyberbullying and online harassment. Among all these cybercrime trends, cyberbullying and online harassment have noticeably increased in the past few years with the wide spread of social media platforms and the civilian's usage patterns of social media platforms. Especially, the younger generations who are addicted to technology and social media are the main targets or the victims of this crucial cybercrime trend, cyberbullying. According to Daily Miror News, SLCERT has reported that over 9,000 cybercrime incidents were reported, with 80% linked to social media platforms between August and September 2024. Moreover, this includes 85 cases of cyberbullying involving children and 40 cases of online sexual abuse targeting minors (Barukanda, 2024).

2.3.2. Cyberbullying Cybercrime Trend in Sri Lanka

Cyberbullying is the usage of digital technology like social media platforms, emails, gaming platforms and messaging platforms to harass, threaten and embarrass an individual or a group of individuals. Unlike traditional bullying, cyberbullying has no specific time or place, it can occur anywhere at any time making it more critical and extremely hard to escape. It can be in any form such as spreading rumors, sharing private or false information, sending threatening messages, videos, photos, and impersonating an individual or a group of individuals online to cause harm.

However, it mostly leads to serious emotional distress, social isolation or even suicides and murders. Some types of Cyberbullying trends;

- Harassment – Harassment is sending hurtful or offensive messages, emails, or comments repeatedly to a person or a group of people.

- Flaming – Flaming is engaging in online fights, arguments in chat rooms or comment sections often using aggressive and offensive language.
- Outing – Outing is sharing private, personal confidential and sensitive information about a person or group of people without their permission to embarrass them and hurt them.
- Catfishing – Catfishing or impersonation is creating fake profiles or accounts in social media platforms pretending to be someone else to trick or harm a victim.
- Cyberstalking – Cyberstalking is intense online harassment which includes threats of harm or physical violence.
- Trolling – Trolling is provoking someone online by posting offensive and hurtful comments.
- Exclusion – Exclusion is kicking out or leaving someone out of online groups, chats and social activities to make them feel isolated and excluded.
- Meme Bullying – It is using edited images, GIFs, stickers or any sort of a meme to mock and humiliate a person or a group of people publicly.
- Sexting and Revenge Porn – Sexting is sharing deepfake or inappropriately edited images, messages and videos without permission to blackmail or take revenge from someone.
- False Accusations - this is spreading false information or rumors to damage the reputation of an individual or a group of individuals.

Despite the type of cyberbullying trends, each type has affects harmfully on its victims leading to motional, psychological, and legal consequences.

2.3.2.1. Rise of Cyberbullying Cases in Sri Lanka

The most significant reason for the rise of Cyberbullying cases in Sri Lanka is the rapid growth of social media usage in Sri Lanka. Despite the age group, all young, mid and older age groups use social media daily and it has gradually become a part of their day-to-day life. Hobbies, fun activities, family time and leisure time are consumed by social media. At present, even some middle school kids have mobile phones with access to all the social media platforms in Sri Lanka. SL Cert has reported that most of

the victims are university students, the young adults who are considered as the future of the country. Sri Lanka is a small island with 25 districts and a population of approximately 21.96 million so, comparing with other countries in the world it is a small country with an average amount of population. However, sources state that there were 8.20 million active social media user identities in Sri Lanka in January 2025 which is 35.4 percent of the total population (Kemp, 2025). These statistics clearly highlight that over the past years and with development of technology, social media has speeded out vastly all around the country within a very short time period.

With the recent rise in usage of social media, Sri Lankans are more updated, and they tend to judge other individuals and show hatred with comments, messages, posts etc. For an instance, Sri Lankan celebrities face this issue all the time, the public are connected to celebrities through social media closely and this had affected the privacy of celebrities, their personal details are published, they are criticized openly, they are trolled bullied for their creations, actions, speech and behavior. Not only celebrities this could happen to any civilian despite their age, gender, generation, career, status etc. because Sri Lankans now tend to use social media as a platform to speak up about the things that they are not comfortable speaking in person. Due to this reason cyberbullying, using hate speech and trolling in social media has become quite common in society. Moreover, sources state that, over 2,000 cases of cyberbullying have been reported in Sri Lanka so far in 2024, according to official records from the Sri Lanka Computer Emergency Readiness Team (SLCERT). Furthermore, Charuka Damunupola, an engineer of SLCERT revealed that approximately 9,500 cybercrimes had been reported during the first 10 months of the year 2024.(Weerasooriya, 2024)

2.3.2.2. Effects of Cyberbullying

Cyberbullying is an increasingly severe concern at present due to its lasting impacts on victims, especially the younger generations who are victimized right way. Various scholarly studies, research papers, and surveys provide evidence of the harmful effects of being cyberbullied such as;

- Psychological effects like depression, anxiety, low self-esteem, self-worth, isolation and loneliness are quite common among victims. Such psychological effects affect their daily lives and overall mental health. Ariyadasa (2019) highlighted that approximately 90% of university students subjected to cyberbullying in Sri Lanka reported symptoms of psychological distress, including anxiety and depression (Ariyadasa, 2019). Moreover, continuous harassment and humiliation online lead these victims to perceive themselves negatively, lowering their confidence, self-esteem, social withdrawal and isolating themselves from friends and family (Gohal et al., 2023).
- Cyberbullying results in the decline in academic performances reducing productivity and work performance. Continuous bullying negatively impacts the concentration and engagement of the victims, causing a decline in academic performance, future interests, disrupting the productivity, increasing dropout rates and reducing the overall work efficiency (Gohal et al., 2023).
- Cyberbullying also indirectly influences physical health like sleep disorders and physical Stress Symptoms. Victims frequently report disturbances in sleep patterns, such as insomnia, due to anxiety or distress caused by cyberbullying. Moreover, the chronic stress from bullying leads to physical symptoms like headaches, stomach issues, and fatigue (Gohal et al., 2023).
- Long-term exposure to cyberbullying can lead to noticeable behavioral changes like aggression, risk-taking behavior and violence. With this, victims in extreme scenarios expose aggressive behavior, sometimes leading to revenge and violence (Ariyadasa, 2019).
- A particularly severe consequence of cyberbullying is suicidal ideation and actions. The relentless nature of cyberbullying, combined with isolation and emotional distress, increases the risk of suicide. Moreover, Ariyadasa (2019) has elaboratively mentioned cases where cyberbullying incidents grown tragically worse, ending up suicide (Ariyadasa, 2019).
- Furthermore, cyberbullying negatively impacts social relationships by building up distrust between peers, friends, and family members, leading to damaged

relationships and disrupting social interactions, missing critical opportunities for personal growth and socialization.

- Apart from all the above cyberbullying indirectly affects financial implications by simply wasting money on psychological counseling and medical treatments, imposing financial burdens on families and individuals. Other than that, cases requiring legal action further enhance financial strains.

2.3.2.3. How to mitigate Cyberbullying in Sri Lanka

There are various options to mitigate cyberbullying in Sri Lanka. However, mitigating this type of cybercrime in Sri Lanka involves a varied set of approaches including;

- Awareness programs and education

  Holding educational programs and workshops by institutions like SLCERT and the National Child Protection Authority (NCPA), targeting schools, universities, and community centers to inform the young generation to identify, avoid, and respond against cyberbullying, are crucial to promote online safety awareness (Ariyadasa, 2019; Heshan Maduranga, 2024).

- Effective legislation and law enforcement

  Strengthening legal frameworks and cybercrime laws, with clear definitions, strict penalties, reporting channels, enabling victims to report cyberbullying without fear of revenge, is crucial for the mitigation of cyberbullying attempts (Sampath, 2023).

- Technical solutions

  Implementation of technological tools, such as multilingual hate-speech detection tools, software and social media monitoring systems specifically aligned for the unique context of Sri Lanka, to identify and manage cyberbullying is another effective mitigation strategy (Muthuthanthri and Smith, 2024).

- Support systems and counseling services

  Furthermore, providing accessible psychological counseling and support services to victims helps to mitigate the psychological harm caused and prevents long-term trauma suffered by the victims as a result of cyberbullying.

2.3.2.4. Importance in Mitigating Cyberbullying

Mitigating cyberbullying is crucial due to several reasons. Firstly, mitigation of this cybercrime directly helps in protecting the mental health of individuals, specifically the youth, helping to prevent conditions like anxiety and depression (Gohal et al., 2023). Additionally, it promotes the creation of safe digital environment where users have the ability to interact positively, minimizing the risks of being online bullied. trolled and harassed (Jayasinghe et al., 2024). Moreover, severe and dangerous outcomes such as loss of self-harm, suicide attempts, and social isolation can be reduced (Ariyadasa, 2019). Furthermore, reducing cyberbullying attempt leads to a variety of educational and social benefits, including better academic performance, enhanced productivity in professional settings, and stronger relationships within communities. With consideration to all these reasons mitigation of cyberbullying is crucial according to the current situation in Sri Lanka.

2.3.3. Forensic Investigation Challenges in Sri Lanka

2.3.3.1. Limited Use of Forensic Tools in Sri Lanka

The usage of digital forensic tools in Sri Lanka remains significantly low due to several factors like high costs, the complexity, lack of training and language limitations. The high costs of implementing and maintaining advanced forensic tools, such as Cellebrite and Oxygen Forensic Suite, limit their widespread adoption in Sri Lanka (Cellebrite, 2025; Oxygen, 2025). Moreover, the existing forensic tools are complex, require extensive training, hence only a limited number of trained personnel in Sri Lanka have the accessibility for such tools (Maduranga, 2024). Furthermore, popular forensic tools typically support global languages and are quite effective for Sinhala and Sinhala-English code-mixed contexts, causing difficulties when investigating in local cybercrime cases (Muthuthanthri and Smith, 2024).

2.3.3.2. Limited Security Professionals in Sri Lanka

Sri Lanka faces a shortage of skilled cybersecurity and digital forensic professionals at present due to various reasons. Limited availability of specialized

training programs in cybersecurity and digital forensics limits the growth of qualified professionals (Maduranga, 2024). Moreover, there is a significant difference between the demand for cybersecurity professionals and their availability, hindering effective response to rising cybercrimes (Sampath, 2023). Furthermore, at present talented professionals migrate abroad for better opportunities, expanding the local shortage (Sampath, 2023).

2.3.3.3. Lack of Knowledge of General Public About Forensic Investigations

However, despite the wide spread of technology, public awareness on digital forensic practices and procedures is significantly low. Many Sri Lankan citizens lack knowledge about what constitutes digital evidence, how forensic investigations are conducted, and how critical evidence can be preserved (Ariyadasa, 2019). Moreover, the limited understanding of forensic investigations leads to fear, mistrust, and doubt towards law enforcement, which results in the increase of underreported cases and creating gaps in addressing cybercrimes effectively (Maduranga, 2024).

2.3.4. Unique Sri Lankan Context – Multilingual Language Barriers

2.3.4.1. Complexity of Sinhala Language

Sri Lanka is a culture enriched country with a unique language called Sinhala, as its primary language, this unique and complex language challenges the existing digital forensic and NLP applications affecting in the effectiveness and accuracy. Sinhala has complex grammatical structures, with various conjugations, tenses, and significant expressions, in which an automated language processing model or tool particularly challenging (Samarasinghe, Meegama and Punchimudiyanse, 2020). Computational linguistic tools and resources like datasets, corpora, language models for Sinhala language are limited compared to global languages, hence NLP development, forensic tool effectiveness is limited, and it is challenging to train models with limited datasets (Samarasinghe, Meegama and Punchimudiyanse, 2020). Furthermore, the use of mixed up languages like using English letters to convey Sinhala messages further complicates

text extraction and analysis in forensic investigations (Samarasinghe, Meegama and Punchimudiyanse, 2020).

2.3.4.2. Complexity of Sinhala-English Code-Mixed Language

Code-mixing between Sinhala and English, using English letters to represent Sinhala words, significantly complicates cyber investigations and NLP analysis. This mixed up language system does not have any standardized spelling or representation when using English letters for Sinhala phonetic sounds hence it creates uncertainty and reduces the accuracy in automated text analysis (Muthuthanthri and Smith, 2024). Moreover, annotating datasets for training machine learning models are challenging due to inconsistencies and lack of clear language boundaries in code-mixed texts (Muthuthanthri and Smith, 2024).Furthermore the existing NLP models, designed for single-language analysis, produces results with reduced accuracy when handling complex Sinhala-English code-mixed scenarios (Muthuthanthri and Smith, 2024).

2.4. Existing Systems

2.4.1. Review of Existing Systems, Approaches and Tools.

Forensics Investigation on Social Media Apps (Abu Hweidi et al., 2023)

This study provides a clear and structured approach to forensic analysis on social media platforms like Facebook and Instagram, showing how digital evidences such as messages, images, timestamps, and metadata can be extracted for use in legal crime investigations. Moreover, it follows the NIST framework to ensure the process is thorough and legally sound, and it reviews different extraction tools used in real-life cases. However, the study does not cover language-related challenges, such as identifying hate speech and cyberbullying in Sinhala or Singlish, nor does it explore using AI and NLP for deeper content analysis. Despite this, it outlines the use of technologies like mobile forensic suites, SQLite analysis, the NIST methodology, and metadata tools. In conclusion, even though it offers a solid foundation for forensic work, it needs to include multilingual and AI-powered features to be fully effective in countries like Sri Lanka.

Evidence Gathering of Facebook Messenger on Android (Chang and Yen, 2020)

This research focuses on collecting digital evidence from Facebook Messenger on Android devices, detailing how to recover messages, timestamps, metadata, and files stored within the application. It identifies key data locations storing critical evidences like SQLite databases and cache directories and explains how to access them using ADB or physical methods. The paper is specifically relevant for Sri Lanka, as it offers clear, step-by-step guidance tailored to the widely used Android system and emphasizes maintaining legal standards through proper evidence handling. However, it only addresses data extraction and excludes post-extraction analysis that includes sentiment analysis, user behavior profiling, and cyberbullying identification. Furthermore, it mainly evaluates English content, omitting Sri Lanka's language diversity. Although the overall study offers strong technical guidance, its lack of content analysis reduces its usefulness in tackling modern digital crimes. To be more effective in Sri Lankan cybercrime investigations, especially those involving hate speech or abuse, it should be enhanced with NLP tools that can handle multilingual and context-rich content.

Forensic Analysis of Social Networking Applications (Menahil et al., 2021)

This study looks into the forensic analysis of social networking apps, focusing on user interactions and recovering critical data such as messages, timestamps, and file attachments. It covers both logical and physical data extracting methods, as well as deleted file recovery, which are frequently disregarded in ordinary investigations. This diverse research, which employs cross-platform analysis, includes the recovery of deleted information and forensic traces that are frequently overlooked in normal techniques. The systematic approach it takes is consistent with industry norms, making it appropriate for both law enforcement and academic application. However, the study focuses mostly on technical elements and avoids addressing local and language-specific problems. Additionally, it lacks tools for detecting cyberbullying and evaluating text using advanced methods such as NLP. Although this study provides a good framework for general forensic operations, it lacks the intelligence layer needed to analyze user activity and identify offending content. The study would have greater impact if its combined with natural language processing (NLP) and hate speech identification

techniques that are adapted to local language patterns in nations like Sri Lanka, in which cybercrimes frequently use local and code-mixed languages.

Identification of Hate Speech in Social Media (Ruwandika and Weerasinghe, 2018)

This initial research provides a foundation for Sinhala NLP research and focuses on identifying hate speech in Sinhala-language social media content. Using a manually gathered and annotated dataset from public posts, it employs fundamental supervised machine learning techniques, such as SVM, to categorize speech as hate and non-hate. Common emotions and language patterns associated with hate speech in Sri Lanka are also examined in the study. Although this lays a strong foundation to address this problem and offers insightful information, its limited dataset and usage of basic models limit its capacity to handle real-world situations and scale up. Additionally, it does not detect code-mixed Sinhala-English content, which can frequently be discovered online. SVM, emotion analysis, and Sinhala Unicode text tokenization are among the technologies used in this approach. Although the study provides a solid foundation overall, it does not include proof of concept. With support for transliterated Sinhala-English code-mixed material, deep learning, and larger datasets, this approach provides a great foundation for further research.

Hate Speech Detection for Code-Mixed Data (Muthuthanthri and Smith, 2024)

The goal of this study is to identify hate speech in transliterated, code-mixed Sinhala-English language, which is frequently found on Sri Lankan social media. It uses sophisticated natural language processing (NLP) techniques, particularly the BERT transformer model, to address the complexity of code-mixed language. Moreover, the algorithm is capable of understanding informal and natural online communication as it is trained on a specifically constructed dataset that comprises Sinhala words written in English letters. This study further addresses a critical gap in Sri Lankan hate speech identification by examining how people use language in online platforms. However, compared to previous approaches, BERT increases accuracy and facilitates the model's better understanding of context but still struggles with a few things like sarcasm, short texts, and figurative language, and it hasn’t been fully tested beyond its dataset. Overall, this study is one of the most advanced, and it directly addresses the specific Sri Lankan

context providing a strong foundation for developing multilingual forensic tools to identify cyberbullying as it uses of real-world language and latest NLP techniques.

Cyberbullying Detection using AI (Khairy, Mahmoud and Abd-El-Hafeez, 2021)

This study proposes an AI-based model for detecting abusive and bullying behavior on social media, with a focus on verbal abuse such as insults, threats, and caustic comments. The algorithm is trained on structured data identified by users and moderators, which allows it to efficiently identify abusive content. Moreover, it demonstrates how AI can enable real-time monitoring while reducing the requirement for constant human supervision. The system is flexible, so it is capable of adapting to different sorts of violent conduct, and its structured labeling method is compatible with existing filtering systems. However, it functions only with English-language data and does not support the Sinhala and Sinhala-English transliterated unique local languages found in countries like Sri Lanka. Furthermore, the technologies used in this approach include NLP preprocessing, classifiers such as logistic regression and SVM, and abuse categorization. Although the study provides an effective and scalable detection model, it requires retraining with local languages and usage patterns before it can be used in Sri Lankan forensics. Still, it provides a useful platform for developing AI-powered cybercrime detection systems.

LLMs for Cyberbullying Detection (Ogunleye and Dharmaraj, 2023)

This study explores the usage of Large Language Models (LLMs), namely GPT-3, to detect cyberbullying in social media conversations. This highlights how these models can recognize context, tone, and complex language that older systems sometimes overlook. One of their unique features is their use of zero-shot and few-shot learning, which enables them to perform effectively even with insufficient training data. These models reflect an advanced approach, capable of understanding complex language, slang, and a variety of text forms. However, the study focuses mostly on English and does not address multilingual or code-mixed content, limiting its usefulness in countries like Sri Lanka. Furthermore, the high computing demands of LLMs such as GPT-3 could be an obstacle for developing countries with less resources. GPT-3, NLP pipelines, and few-shot learning algorithms are among the technologies used. Though

the study reveals that LLMs have a high potential for detecting cyberbullying, their practical and effective implementation in Sri Lanka requires local language adaption, specialized training data, and cost-effective infrastructure.

Cyberbullying, Spam & Bot Detection (Chathurangi et al., 2024)

This study proposes a paradigm for detecting multiple cyber threats such as cyberbullying, spam, and bot activity on Sri Lankan social media platforms, understanding that these issues frequently overlap. It employs language analysis and rule-based methods to detect harmful content and conduct, with a heavy emphasis on local context. Furthermore, this study integrates cyberbullying detection with spam and bot analysis specifically for Sri Lanka, providing a more comprehensive picture of online threats. The approach, however, depends on keyword-based rules rather than deep learning or advanced NLP, making it less effective at dealing with complex or informal language, such as Sinhala-English code-mixed text. Classifiers powered by rules and simple pattern recognition were among the technologies deployed. Overall, the study provides a solid foundation for a localized detection system, but in order to be more effective, it must integrate updated AI methods like as transformer models and support for multilingual data, which is widespread in Sri Lankan online communication.

Threatening Language and Target Identification (Amjad et al., 2021)

This study focuses on recognizing threatening language in social media and determining who the threats are aimed at, individuals or organizations. It emphasizes the risks of both direct and indirect threats along with a text classification system capable of detecting violent and abusive content. Unlike simple analysis of emotions, it seeks to identify specific targets, making it useful in forensic and law enforcement investigations. The concept is intended to integrate into legal frameworks by identifying various forms of hazards. However, it is based on English-only data and does not account for local differences, multilingual and code-mixed communication, limiting its utility in countries like Sri Lanka. NLP, Named Entity Recognition (NER), and threat classification models are the main technologies used in this approach. In conclusion, although the study presents a valuable tool for identifying digital threats and their

targets, it would require extensive adaptation especially language support for Sinhala and code-mixed text to be genuinely effective in Sri Lankan cybercrime investigations.

Cyberbullying Prediction using ML (Al-Garadi et al., 2019)

This study addresses applying machine learning to anticipate cyberbullying before it occurs by evaluating user behavior, post content, and engagement trends. Rather than simply responding to threatening content, the tool identifies accounts that are more likely to engage in bullying based on historical trends. This proactive strategy is beneficial for moderation teams in detecting concerns early on. It combines activity tracking and textual analysis to detect recurrent abusive habits more efficiently. However, the model is based on English data and does not adjust for local language variances, hence its applicability in countries like Sri Lanka is limited. Additionally, it relies on huge, labeled datasets, which can be difficult to find in situations with limited resources. Behavioral analytics, supervised machine learning, sentiment analysis, and temporal feature analysis are the main technologies used for this approach. In conclusion, though the approach has great potential in preventing cyberbullying, it requires local language adaptation, in addition to locally relevant data, to be used effectively in Sri Lankan forensic investigations.

Sinhala Cyberbullying Classification (Amali and Jayalal, 2020)

This study proposes a machine learning approach for detecting cyberbullying in Sinhala, which is an unrepresented language in NLP. The model is trained on a customized dataset of Sinhala social media posts, and it intends to provide a useful, locally relevant tool for Sri Lankan law enforcement and social media platforms. Being research that investigates cyberbullying in Sinhala, it provides a significant contribution by dealing with both local language issues. However, the dataset is small and lacks a diverse range of bullying categories, therefore the model's accuracy is limited. It further does not allow code-mixed Sinhala-English text, which is widely used on networks like as Facebook. Supervised machine learning algorithms such as SVM and Naive Bayes were employed, in addition to a customized Sinhala dataset and text preprocessing. In conclusion, this study is a good start toward localized forensic tools in Sri

Lanka, however, it requires a larger dataset, support for code-mixed languages, and more advanced models such as deep learning.

Online Hate Speech in Sinhala (Shibly, Sharma and Naleer, 2021)

This study focuses on recognizing hate speech written in Sinhala using classic NLP and machine learning techniques. It emphasizes the difficulty of working with Sinhala's distinct syntax and vocabulary, hence proposes a methodology for detecting hate speech from normal conversation. The data set gathered locally helps to keep the social media context genuine. Additionally, the study is significant for the country as it specifically focuses on Sinhala. However, using simple models like Naive Bayes and logistic regression limits the system's capacity to detect deeper and complicated language, such as sarcasm and hidden hatred. Furthermore, the dataset is limited and unbalanced, hence it limits the generalizing ability of the model. In conclusion, this study is an ideal starting point for detecting Sinhala hate speech but requires using more advanced techniques like deep learning and expand its coverage of Sinhala-English code-mixed text to be more effective for real-world cybercrime detection in Sri Lanka.

Instagram Digital Forensics (Mubarik et al., 2021)

This study focuses on the digital forensic examination of the Instagram application on Android devices, demonstrating how to retrieve critical evidence such as chat logs, media files, and timestamps for criminal investigations. It emphasizes the importance of maintaining a chain of custody and describes a simple, practical methodology for data recovery and presentation, which is applicable to real-world legal concerns. However, while the forensic procedure is effective, the study does not address what to do with the data after extraction, such as recognizing cyberbullying or tracking user behavior. It further skips problems associated with language, which are critical in countries like Sri Lanka. Android forensic tools, SQLite viewers, and ADB are the technologies used in the research approach. Overall, the study provides a solid framework for gathering Instagram evidence, but it would be far more valuable for Sri Lankan cybercrime investigations if integrated with AI and NLP models capable of analyzing Sinhala and code-mixed information.

Integrated Detection of Cyberbullying (Jones, Winster and Valarmathie, 2022)

This study employs a comprehensive strategy to detect cyberbullying by merging several sorts of digital evidence, such as text messages, user activity, and platform metadata. It combines forensic tools and NLP approaches to create a more accurate and trustworthy investigation process, comparable to that which law enforcement conduct. Employing technologies like Oxygen Forensics in conjunction with NLP models enables both effective data extraction and deeper analysis. Although the procedure is accurate and forensically solid, it does not handle different languages, making it unsuitable for countries like Sri Lanka. It additionally is not suitable for low-resource and command-line-based systems. In conclusion, the study provides a solid foundation for cyberbullying identification and serves as a useful model. However, to be useful in Sri Lanka, it must offer multilingual content and remain more resource efficient.

Enhancing Sinhala Hate Speech Accuracy (Fernando and Deng, 2023)

This paper aims to improve the accuracy of hate speech detection in Sinhala by modifying how models process language and extract features. It combines techniques like emotion labeling and phrase structure analysis to better comprehend the overall mood and purpose of messages, with the goal of reducing false positives and negatives. Unlike many experiments, it goes beyond basic detection and focuses on improving model accuracy, focusing on Sinhala to maximize performance in this area of study. Its focus on Sinhala-specific elements of the language makes it especially useful for forensic purposes in Sri Lanka. However, it only handles Sinhala text and excludes the common usage of Sinhala-English code-mixed language, limiting its applicability in real-world situations. Overall, this study represents an important step forward in detecting Sinhala hate speech, but it must be expanded to include code-mixed and multilingual content in order to be truly relevant in Sri Lankan cybercrime investigations.

Bengali Hate Speech Detection (Senapati and Roy, 2023)

This study addresses hate speech detection in Bengali using deep learning models such as CNNs and RNNs. Bengali, like Sinhala, is a low-resource language, hence the research addresses issues such minimal annotated data and script complexity.

The models produce great results after carefully adjusting and preparing the data, proving that deep learning is still effective with small datasets. The study is interesting since it focuses on a closely connected language and illustrates how AI can perform well in such contexts. However, it does not address how to implement the model in real-world systems and is restricted to the Bengali context. CNN, RNN, and NLP pipelines are the main technologies deployed in this approach. Overall, the study provides useful insights for detecting Sinhala hate speech, especially on how to manage limited data, but it lacks support for code-mixing and multilingual usage, which are critical in Sri Lanka. It is a useful point of reference for adopting deep learning methodologies to Sinhala.

2.4.2. Comparative Analysis Review of Existing Systems

The review of current tools for cyberbullying detection and social media forensics reveals clear strengths and limitations across different approaches. Tools introduced by Abu Hweidi et al. (2023) and Chang & Yen (2020) are excellent at extracting data such as chat logs and metadata, but they stop at surface-level recovery and don does not help in understanding harmful or abusive content especially in multilingual environments like Sri Lanka (Chang and Yen, 2020; Abu Hweidi et al., 2023). On the other hand, studies by Muthuthanthri & Smith (2024) and Fernando & Deng (2023) makes solid progress in detecting hate speech in Sinhala-English code-mixed content using advanced models like BERT (Fernando and Deng, 2023; Muthuthanthri and Smith, 2024). Although it is effective, these models need high-quality datasets and still struggle with detecting subtle forms of abuse like sarcasm. Furthermore, earlier work by Ruwandika & Weerasinghe (2018) and Amali & Jayalal (2020) lay the foundation for Sinhala-language abuse detection using simpler models like SVM (Ruwandika and Weerasinghe, 2018; Amali and Jayalal, 2020). These are resource-efficient but not effective enough for complex and mixed-language online conversations at present.

Moreover, some studies shift their focus to behavior, like Al-Garadi et al. (2019), who try to predict bullying based on user activity patterns (Al-Garadi et al., 2019).

Although the work is exceptional, these models are not tailored for Sri Lankan users. Similarly, Amjad et al. (2021) propose systems to identify targets of online threats, but they do not align with unique local language and context (Amjad et al., 2021). However, Ogunleye & Dharmaraj (2023) use powerful LLMs like GPT-3 to interpret abuse with impressive accuracy but, such models are expensive and poorly adapted to local context, hence, is hard to implement in countries like Sri Lanka (Ogunleye and Dharmaraj, 2023). Chathurangi et al. (2024) introduces a more practical, locally aware approach combining detection of spam, bots, and cyberbullying using rule-based systems which are easier to run but not flexible enough for detecting latest and more complex threats (Chathurangi et al., 2024). Finally, specific studies like Asim Mubarik et al. (2021) focus on Instagram forensics, helping investigators recover important evidence but they lack tools for understanding the meaning behind the content (Mubarik et al., 2021). Overall, the tools are specified for a specific task, some are great at retrieving data, others at analyzing language so, a system that brings both together which is capable of handling multilingual, real-world cybercrimes fits the digital landscape Sri Lanka.

2.5. Technological Review

The development of the multilingual Android forensic tool for cyberbullying detection required a careful integration of datasets, preprocessing techniques, and algorithmic models all selected to suit the complex linguistic and forensic environment of Sri Lanka. In developing this system, which is focused on using advanced machine learning models ensuring that the tool is practical, culturally adaptive, capable of handling the cyberbullying attempts and its language usage in Sinhala, Sinhala - English code-mixed formats.

Dataset

The model was developed using two key annotated datasets, a Sinhala Hate Speech Dataset and a Sinhala-English Code-Mixed Hate Speech Dataset. Both datasets are compiled from real-world social media content, gathered from platforms like Facebook, where cyberbullying and hate speech are prominent. Each dataset includes

user-generated comments manually labeled with categories such as hate speech, cyberbullying, and neutral content.

The Sinhala Hate Speech Dataset includes Unicode Sinhala text, covering a range of offensive expressions commonly used in native Sinhala. It provides structured insight into abusive language in a fully Sinhala context, hence it is crucial for detecting hate speech written in the local language, the mother tongue of Sri Lanka.

On the other hand, the Code-Mixed Hate Speech Dataset which contains comments where Sinhala words are typed using the English alphabet commonly known as transliterated Sinhala often mixed with English phrases and most of the Sri Lankan use this method in social media platforms for communication especially the younger generations. Such comments are typically informal, context-dependent, and highly variable in spelling and grammar; therefore, its uniqueness challenges automated detection. However, both datasets were relatively small and suffered from class imbalance, with fewer samples labeled as severe hate speech and cyberbullying compared to neutral comments. To address this:

- Cleaned the data by removing duplicated, mislabeled, and incomplete records.
- Balanced the class distribution through oversampling of minority categories.
- Performed light data augmentation by introducing spelling variants, synonym substitutions, and transliterated duplicates to reflect the noisy, real-world nature of user content.

Moreover, both the datasets were combined into a unified training set, allowing the model to learn from the full language spectrum formal Sinhala, informal Sinhala, and Sinhala-English code-mixed text. This decision was critical for generalization, ensuring that the model could perform reliably on diverse user inputs during real forensic casework. In conclusion, the dataset strategy not only enhanced the model's language coverage but also reflected the social and cultural communication patterns unique to Sri Lanka. This dual-dataset approach forms the backbone of the system's ability to detect doubtful, context-driven cyberbullying content across local language forms.

Preprocessing

The dual-dataset nature of the model in Sinhala Unicode and Sinhala-English code-mixed transliteration the preprocessing pipeline accommodated two distinct language forms. This stage is not just about cleaning the text it is about creating a bridge between informal social media language and structured machine learning input. With that custom tokenization, using separate configurations for Sinhala script and English-Sinhala character transliterations. Sinhala Unicode characters are tokenized based on native syllabic structure, while transliterated Sinhala followed word-boundary rules which are optimized for code-mixed texts. Moreover, to standardize inputs, normalization was applied which is converting all text to lowercase, removing punctuation, emojis, special characters, and standardizing spellings where possible.

Given that both datasets contained a mix of offensive and neutral content, therefore stop word removal was applied cautiously ensuring we did not remove emotionally charged or context-bearing words. Furthermore, a language tagging module is applied which identifies the segments as Sinhala or English. This allowed text to feed into multilingual embeddings with better emotional alignment. Additionally, class labels were then encoded into hate speech, cyberbullying, and neutral categories, enabling structured training and meaningful evaluation metrics such as accuracy, recall, and F1-score.

With that to address the spelling variations and informal grammar common in user-generated content, we introduced minor augmentation techniques like synthetic misspellings and variant transliterations to help the model generalize better. This stage is especially important for improving performance on informal and youth based online language usage, where grammar and spellings are often intentionally distorted. However, together, these preprocessing steps enabled the system to understand and adapt to the complex digital languages of modern Sri Lankan social media, allowing the model to make meaningful predictions on highly variable, multilingual user input.

Algorithm Selection

Choosing the right model architecture is the core to develop a tool capable of understanding not just language, but multilingual online expression. After exploring

several candidate algorithms, BERT (Bidirectional Encoder Representations from Transformers) is adapted as the primary model. Its ability to capture bidirectional context made it ideal for interpreting mixed-language posts, detecting implied and indirect hate speech, and understanding fragmented syntax which is a common trait in cyberbullying. Moreover, a multilingual BERT fine-tuned using the Hugging Face Transformers library on PyTorch as this version of BERT had already been pretrained on multiple languages including Sinhala to some extent, it is specifically effective in handling transliterated Sinhala-English content. Even when sentences contained in middle switches in language and tone, BERT is able to contextualize them more reliably than classical models.

However, before finalizing BERT, the models below were compared as well:

- CNN with LSTM Hybrid: Combined convolutional pattern recognition with sequential modeling. This is effective for structured sentences, but fails to capture contextual content in short, slang-heavy and emotionally charged social media posts.
- Naive Bayes (Baseline): Used for benchmarking, this model offered fast predictions but lacked semantic depth. It often misclassifies the content due to its inability to detect sarcasm and hidden emotions.
- Transfer Learning: Although this is not implemented in the final tool, this is evaluated the potential of cross-lingual transfer learning. Embeddings pretrained in morphologically similar languages (e.g., Hindi) showed the capacity for enhancing Sinhala model performance with minimal training data.

Training was conducted locally using PyTorch, where available hardware resources are used to manually tune hyperparameters such as learning rate, batch size, and number of epochs. Cloud-based platforms like Google Colab limit the scalability of experimentation, but using hardware resources allowed to maintain full control over the training process and environment.

In summary, the tool combines localized datasets, a structured preprocessing pipeline, and a fine-tuned BERT model to effectively detect cyberbullying in the Sri

Lankan context. It understands the unique mix of Sinhala, English, and code-mixed language and aligns modern ML with real-world local usage for digital forensic investigations.

2.6. Evaluation and Benchmarking

2.6.1. Evaluation Criteria for Multilingual Cyberbullying Detection

In order to evaluate the effectiveness of the Android forensic cyberbullying detection tool, various evaluation standards were designed based on best practices in NLP, machine learning, and forensic usability;

- Accuracy and F1 Score: These are primary indicators for model performance. Accuracy helps to assess overall accuracy, while the F1 Score offers a balance between accuracy and recall especially important in imbalanced datasets.
- Precision and Recall: Precision highlights how many of the flagged instances are cyberbullying, while recall measures how many of the cyberbullying cases were correctly identified. High recall is important in forensic applications as a missing harmful post can compromise an investigation.
- Language Robustness: Given the code-mixed nature of the input, evaluates how well the model handles Sinhala Unicode, transliterated Sinhala and English. This is done by measuring individual class performance on samples containing each script type, ensuring the tool works across language forms.
- Confusion Matrix Analysis: Confusion matrices are used to identify specific weaknesses, such as false positives which are flagging non-abusive content as cyberbullying and false negatives which are failing to catch actual hate speech. This helps to fine-tune the model's thresholding and improve real-world trustworthiness.
- Practical Relevance: Beyond machine learning metrics, the usability of the tool is evaluated in real-life forensic contexts by measuring the clarity of the generated reports and the usefulness of flagged content. Informal pilot feedback

from testers highlights that the system was able to present evidence in a format that's understandable even by non-technical law enforcement personnel.

2.6.2. Benchmarking Methods and Best Practices

- Standardized Datasets: Although custom and real-world datasets are used, the benchmarked model performance is using a fixed testing subset, and it avoids mixing it during training. This helps to ensure that the results are replicable and trustworthy.
- Cross-Validation: Stratified k-fold cross-validation is used to validate the model consistency. This allows to confirm that the model is not overfitting and maintains stability across different segments of the data.
- Baseline Comparisons: The fine-tuned BERT model is benchmarked against simpler models such as Naive Bayes and CNN with LSTM hybrids. However, the baseline models are computationally lighter, hence their performance drops significantly especially on code-mixed and informal comments. This confirms the superiority of BERT in application.
- Performance Metrics Documentation: All training and testing results, including precision, recall, and F1 scores for each class, are logged and documented for transparency. This not only supports evaluation but also ensures that future enhancements can be compared against these results.
- Real-World Relevance: The tool's design and testing were framed around practical law enforcement needs. Report clarity, CLI usability, and the ability to detect threats in mixed language are continuously benchmarked against real-life communication styles found on Facebook and Instagram.

In conclusion, the tool is evaluated not only for raw machine learning performance but also for its practical impact in real-world forensic settings. By applying both academic and real-life standards, the system is proven to be reliable, scalable, and highly applicable to multilingual cybercrime investigations in Sri Lanka.

2.7. Chapter Summary

In summary, this chapter provided a detailed analysis of the cyberbullying landscape in Sri Lanka, the challenges faced in digital investigations, and the gaps in existing forensic tools especially concerning language limitations. It highlights the importance of developing a cost-effective, multilingual solution capable of analyzing both Sinhala and code-mixed content. By critically evaluating existing systems, datasets, and machine learning approaches, this review justifies the need for a localized Android forensic tool that not only extracts social media data but also detects cyberbullying with accuracy and local relevance to unique Sri Lankan context.

CHAPTER 03 : METHODOLOGY

3.1. Chapter Overview

This chapter outlines the methodology used to design, develop, and evaluate the proposed multilingual Android forensic tool, focusing on detection of cyberbullying through the identification of hate speech on social media platforms. This tool is developed to analyze Sinhala and Sinhala-English code-mixed content which is commonly known as Singlish, from platforms like Facebook and Instagram. Furthermore, the tool addresses a critical requirement in Sri Lanka where the language diversity and limited digital forensic resources challenging the effective crime investigations.

Even though the primary objective of the tool is cyberbullying detection, its practical applications can be extended further. In real-world crime investigations, law enforcement often examines mobile devices and forensics of the suspects in order to uncover harmful behavior and patterns of violence. Therefore, in such cases, this tool can play a valuable role in identifying hate-filled and aggressive inappropriate communication, not only to cyberbullying but also to other forms of cybercrimes and even physical offenses. With that, implementing this automated detection tool and clear reporting of such content, the tool supports evidence gathering and strengthens the investigative process.

Moreover, this chapter continues to explain the research methodology used, followed by the system architecture, dataset collection, preprocessing strategies, model training, and tool implementation. It further outlines the libraries and frameworks used, the reasoning behind model choices, and the performance metrics used for the evaluation of the tool. Additionally, the overall methodology aims to ensure that the system is both technically accurate and practically usable in real-world scenarios and crime investigations in Sri Lanka.

3.2. Research Methodology

This research follows the structure of the *Research Onion* model proposed by Saunders et al., which helps to organize and justify the different methodological layers of the project. A pragmatic and deductive approach was adopted as this study aims to develop a practical forensic tool that uses NLP to detect cyberbullying in Sri Lankan social media. Furthermore, the following table outlines each layer of the methodology with an explanation on what was used and the reason for using this specific research context.

Layer	Used Approach	Justifications
Philosophy	Pragmatism	This study targets the practical problem of detecting cyberbullying on Sri Lankan social media using a flexible, outcome-oriented approach. Pragmatism allows the combination of quantitative analysis like model performance and qualitative needs like local language context, supporting a real-world solution that balances theory and practical implementation.
Approach	Deductive	The project starts with existing theories and tools in natural language processing like BERT and applies them to a specific, underexplored context like Sinhala and Sinhala-English code-mixed hate speech detection. Moreover, it tests how well these existing models work in this new scenario and builds a tool around the observed outcomes, making deduction a suitable approach.

Methodological Choice	Quantitative with elements of applied experimentation	The research primarily relies on measurable results like accuracy, precision, recall and F1-score to evaluate model and tool performance. Although it is not research with a standard combination of methods, real-world applications like CLI usability and language handling are evaluated using iterative development and observed functionality, which adds qualitative value.
Strategy	Experimentation	BERT-based models are trained, tested, and fine-tuned on labeled hate speech datasets in order to conduct experimental evaluations. Additionally, the strategy further includes integrating model output into a command-line forensic tool and testing its results in a simulated forensic environment. Furthermore, this approach helps demonstrate real-world readiness.
Time Horizon	Cross-sectional	The project evaluates the model performance and system functionality based on the data collected and tested during a fixed time duration. Furthermore, it concentrates on developing and assessing the tool in its present form, without any continuous retraining or longitudinal tracking.

Data Collection and Analysis	Two annotated datasets: Sinhala Hate Speech Dataset & Code-Mixed Hate Speech Dataset Analysis metrics: Accuracy, Precision, Recall, F1-Score	These datasets were chosen based on their relevance to real-world Sri Lankan social media usage patterns. These present a good foundation for developing a hate speech detection algorithm that adapts to the local problem. Model evaluation is carried out using standardized categorization measures to provide objective and comparable findings.

Table 2 - Chapter 3: Table of Research Methodologies

3.3. Development Methodology

This development methodology section outlines how the proposed multilingual Android forensic tool is developed, from initial requirements gathering to system design and execution. Moreover, this methodology aims to meet the unique requirements of cybercrime investigations in Sri Lanka, with a focus on detecting cyberbullying through hate speech analysis in Sinhala and code-mixed Sinhala-English media. The development method is directed by a combination of survey feedback, tool analysis, and iterative testing, ensuring that the tool was both technically accurate and feasible for non-technical users like law enforcement agents.

3.3.1. Requirement Elicitation Methodology

For this project literature review, tool review and general survey was used for requirement identification to ensure that the final solution was grounded in real-world needs and challenges.

- Literature and Tool Review – The technical evaluation of the existing tools such as Cellebrite, Oxygen Forensics, Magnet AXIOM and insights drawn from academic literature review of government cybercrime reports, existing researches, approaches and frameworks were used as a primary sources of data collection to

identify the unique requirements in Sri Lanka. Even though the above-mentioned set of existing tools are highly effective at extracting raw data from mobile devices, yet they have limited capabilities to support for analyzing content written in Sinhala and Sinhala-English code-mixed content in Social Media. Furthermore, their high cost and technical complexity makes them inappropriate for widespread adoption in Sri Lanka, especially within public law enforcement divisions.

- A Survey - A general survey was conducted targeting Sri Lankan active social media users. The goal was to understand their experiences with cyberbullying, the languages they commonly use online, and their awareness of existing digital safety tools. However, the majority of respondents indicated they frequently encounter cyberbullying and hate speech from the two languages of Sinhala Unicode and Sinhala-English code-mixed language, especially on Facebook and Instagram platforms. Furthermore, many participants reported witnessing online harassment and hate speech but were unaware of how to report such incidents and if any tools existed to support law enforcement. These responses validated the need for a tool that could detect harmful content in both Sinhala and Sinhala-English code-mixed formats and present the results in a simple, legally usable format.

In conclusion, by combining user feedback with an evaluation of current limitations, the following core requirements were identified:

- The tool must support text extraction from Android devices, specifically from Facebook and Instagram.
- It must detect hate speech in Sinhala Unicode and Sinhala-English code-mixed formats.
- It should present results in a PDF report format that is structured which can be legally presented in investigations as crime evidence.
- The tool must be simple to use, ideally through a Command-Line Interface (CLI).
- It should be lightweight and operable offline, especially in resource-limited environments.

These requirements formed the foundation for system design, model training, and interface development described in the following sections.

3.3.2. Design Methodology

The design methodology of the proposed forensic tool was guided by the core requirements identified through research survey and literature, technical review. Moreover, the system is designed to be modular, lightweight, and compatible with the practical requirements of the digital forensic investigators in Sri Lanka. Moreover, the design process follows a layered structure, to ensure that each module of the tool is independently developed, tested, and improved without affecting the rest of the tool.

Major Functional Layers	Description
Data Extraction Layer	This layer connects to an Android device through ADB (Android Debug Bridge) to extract text-based data from Facebook and Instagram applications. The extraction process targets key directories and databases where chat logs and message content are stored.
Preprocessing Layer	Extracted text is often unstructured and jumbled, especially in informal and code-mixed messages. Therefore, the preprocessing module is responsible for cleaning, tokenizing, and normalizing the content. Special attention is given to handle Sinhala Unicode text, transliterated code-mixed Sinhala-English content.
Classification Layer	This is the core intelligence of the tool, built around a fine-tuned BERT model trained to identify hate speech. Moreover, the model was adapted specifically for low-resource languages using real-world Sri Lankan datasets. However, each message is passed through this layer and classified as Hate or Non-Hate.

Report Generation Layer	The results from the classifier are compiled into a structured PDF report. This report includes the extracted messages, prediction labels, timestamps, and metadata required for investigative use. The format was designed to be simple, readable, and legally usable by non-technical personnel.
CLI Interface Layer	Instead of a graphical user interface, the tool uses a Command-Line Interface (CLI) to allow investigators to operate the system easily from any terminal. This keeps the tool lightweight and ensures compatibility with systems that have limited resources and lack GUI support.

Table 3 - Chapter 3: Table of Research Methodologies

This layered design ensures that each component remains independent and scalable, making it easier to update and expand the tool in the future for example, to support more languages, platforms, and additional forensic features.

3.3.3. Programming Paradigm

Python programming language is used to develop the tool, following a simple and organized structure where each part of the system is built separately. This means that tasks like extracting data, cleaning text, detecting hate speech, and creating reports were all handled by different sections of the code which makes it easier to test, update, and fix specific parts without affecting the whole tool. With that instead of using a complex interface, the tool was designed to work through the command line interface (CLI), where users are capable of running the tool using simple commands. This approach is ideal as it is lightweight, does not require high-end computers, and is easier for use of the investigators who may not have much of a technical background as well.

Additionally for the machine learning model, the hate speech detection model was trained using labeled data and standard Python libraries like PyTorch and Hugging Face Transformers. Furthermore, the code is written in a clean, straightforward style to keep everything simple and maintainable. Overall, the development approach is focused on keeping things clear, practical, and easy to improve in the future.

3.3.4. Evaluation Methodology

The evaluation of the proposed tool's effectiveness is the key part of this project as the main goal of this tool is to detect cyberbullying through hate speech analysis, the evaluation focused on how accurately and reliably the system capable of identifying harmful content in social media platforms. Moreover, the performance of the model is measured using common classification metrics like accuracy, precision, recall, and F1-score which helps to determine accuracy of the tool in differentiating hate speech from normal messages.

However, after training the BERT-based model using two datasets of Sinhala hate speech and Sinhala-English code-mixed text, the tool is tested on unseen data to check the accuracy of its generalization. Moreover, a confusion matrix is used to visualize correct and incorrect predictions by giving a more clear picture on the strengths and weaknesses of the model. In addition to testing the model, the usability of the tool itself is evaluated by running it through practical test cases using Facebook and Instagram data extracted from test Android devices. However, here the focus is to check the efficiency, usability and the user friendliness of the CLI interface, the accuracy and the quality of the reports, and the reliability of the tool working in an offline or low-resource environment.

In conclusion the combination of these evaluation steps helps to ensure that the tool is not just a technical approach but also practical and effective approach for real-world crime and cybercrime investigations in Sri Lanka.

3.3.5. Solution Methodology

As mentioned earlier, the core solution implemented through this research project is a command-line-based android forensic tool that detects cyberbullying by analyzing media on social media platforms for hate speech like comments, messages, captions etc. Moreover, this tool focuses on content written in Sinhala and Sinhala-English code-mixed formats which are two common languages among Sri Lankan social media users.

Therefore, in order to fulfil this requirement, a BERT-based machine learning model is fine-tuned using two datasets of Sinhala hate speech and a code-mixed Sinhala-English hate speech. Moreover, these datasets are preprocessed to clean the text, and format it in a way that the model cab learn from. However, the trained model is considered as the heart of the tool's detection system.

Furthermore, the tool connects to the Android device through ADB (Android Debug Bridge) to extract data from Facebook and Instagram applications and once the data is collected, it is passed through the BERT classifier, which identifies whether each message contains hate speech. Finally, the results are presented in a PDF report, which includes flagged messages, their timestamps, and classifications, making it easy for law enforcement officers to use during investigations and use as legal evidence in courts.

In conclusion, by combining natural language processing (NLP) with basic digital forensic techniques, the tool provides a lightweight, localized solution for this growing problem making it a practical choice for investigators working with limited resources but facing serious cybercrime threats.

3.4. Project Management Methodology

Managing this project requires a structured and flexible approach to ensure that each phase from research and design to implementation and evaluation progressed smoothly and stayed aligned with academic milestones. However, a simplified, Agile-inspired development model was followed, allowing for adaptability and iterative

progress, especially important in a project that combines both machine learning and forensic tool development.

The entire project was divided into five major stages:

- Requirement gathering and analysis
- Dataset collection and preprocessing
- Model selection, fine-tuning, and testing
- Tool design and planned integration
- Evaluation and future reporting integration

Each phase is guided by clearly defined deliverables and soft deadlines, helping to maintain a steady progress while allowing room for modifications. Frequent feedback from the project supervisor helped to refine the direction of the model development and future tool design.

Although the full working tool has not yet been implemented, sprint-style planning was used during model training and evaluation, especially for iterative tasks like dataset balancing, parameter tuning, and performance testing. Moreover, this approach allowed core components to be tested and validated in isolation, ensuring they were reliable before future integration. Furthermore, the following subsections detail the project's scope, schedule, resource requirements, and risk management strategies that are considered during this development phase.

3.4.1. Project Scope

As mentioned before, this project aims to develop a multilingual forensic tool to assist Sri Lankan law enforcement in detecting cyberbullying through hate speech analysis on social media and the primary focus is on analyzing text content in Sinhala, and Sinhala-English code-mixed formats which are the languages commonly used by Sri Lankan youth on platforms like Facebook and Instagram.

The scope of the project is divided into two main phases:

- Phase 1

  At this point Phase 1 is completed which is training and evaluating a BERT-based model capable of identifying hate speech in Sinhala and code-mixed text using two publicly available datasets.

- Phase 2

  Mostly phase 2 is ongoing and planned. This phase includes the development of a command-line interface (CLI) tool that integrates the model, allows input of social media data extracted through ADB, and generates structured PDF reports for forensic use.

In-Scope Features

- Fine-tuning and evaluation of a BERT model for multilingual hate speech detection
- Use of Sinhala Unicode and code-mixed Sinhala-English datasets
- CLI-based tool structure for forensic use.
- Basic Android data extraction from Facebook/Instagram using ADB.
- Text classification and flagging of hate speech content
- Generation of structured PDF reports.

Out of Scope Features (Future Expansions)

- Advanced forensic capabilities (e.g., password bypass, app decryption).
- Support for iOS or non-Android platforms.
- Graphical user interface (GUI).
- Expand for other Social Media Applications.
- Detection of other forms of cyberbullying beyond hate speech.

3.4.2. Schedule

The Gantt chart below outlines the planned timeline for each major task in the project, divided into two key phases, the midpoint phase which is highlighted in blue are already completed and the final development phase which highlighted in green are yet to be completed. The midpoint phase is focused on research, model development, and structuring the initial chapters of the report. These tasks, including the training and evaluation of the BERT-based hate speech detection model, were completed as scheduled.

The final phase includes the development of the CLI-based forensic tool, PDF report generation, Android data extraction, and the remaining report chapters. The structured scheduling of individual tool components and report chapters ensures that the project remains on track for successful completion and submission.

This timeline reflects a realistic, well-phased workflow that balances technical development with academic documentation, helping ensure consistent progress toward the final deliverables.

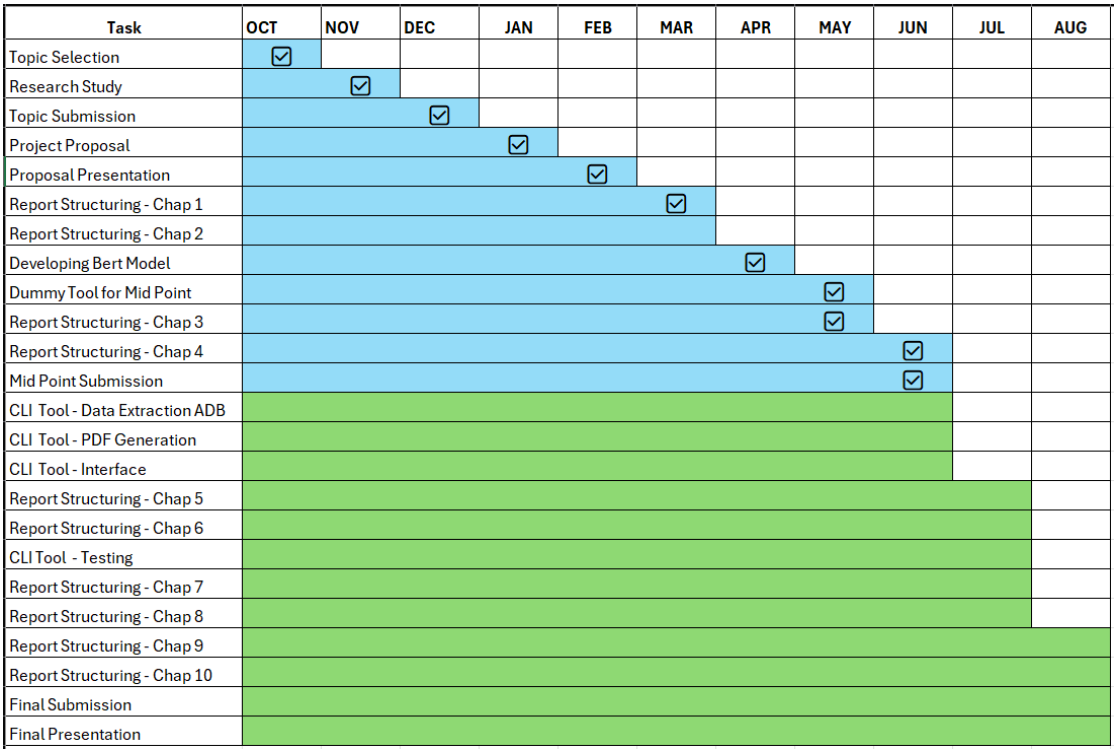


Figure 2 - Chapter 3: Schedule and Project Planning

3.4.3. Resource Requirements

To successfully develop and evaluate the proposed multilingual forensic tool, several resources including both hardware and software are required. Moreover, the project further relies on academic support and publicly available datasets to meet technical and domain-specific requirements.

Human Resources	
Researcher/Developer	Responsible for literature review, dataset collection, model development, evaluation, and documentation basically the whole project.
Supervisor/Mentor	Provides regular guidance, feedback, and helps to refine the scope and methodology.
Software Resources	
Python	Main programming language is used for model development and tool integration.
VS Code	Development environments used for model training, testing, and experimentation.
PyTorch and Transformers (Hugging Face)	Libraries are used for fine-tuning and evaluating the BERT model.
Pandas, NumPy	Data handling and preprocessing.
FPDF	For planned PDF report generation.
ADB (Android Debug Bridge)	Planned for extracting data from Android devices.
GitHub	Version control and project backup.
VM Ware	For testing ADB-based data extraction from Facebook and Instagram.
Hardware Resources	

Personal Laptop	Intel Core i7 13th Gen 16GB RAM NVIDIA RTX 4060 GPU Used for model training and development
Dataset Resources	
Sinhala Hate Speech Dataset (sinhala-hate-speech-dataset new.csv)	Used to train and evaluate the BERT model for accurate hate speech classification.
Code-Mixed Sinhala-English Dataset (singlish_dataset1.csv + singlish_dataset2.csv)	

Table 4 - Chapter 3: Table of Resource Requirements

These resources were carefully selected to balance effectiveness, accessibility, and cost. Open-source tools and publicly available datasets allowed for efficient development without incurring licensing fees, making the tool practical and scalable for use in Sri Lanka.

3.4.4. Risks and Mitigation

Like any software and research-based project, the development of this forensic tool came with several risks, especially given the integration of machine learning, multilingual NLP, and forensic investigation workflows. The table below outlines the key risks identified, their potential impact, and the mitigation strategies applied or planned.

Risk	Impact	Likelihood	Mitigation Strategy
Limited availability of Sinhala and code-mixed datasets	High	Medium	Two datasets were merged and preprocessed to improve data quality and balance.
Model overfitting or low generalization	Medium	Medium	Used dropout, validation split, and performance metrics like F1-score for tuning.

Delay in CLI/tool implementation	High	High	Focused on completing model first; tool development scheduled as Phase 2 deliverable.
Difficulty in accessing real Android device data	Medium	Medium	Planned to use sample exports or emulator-based test data if device access is limited.
Complexity of handling code-mixed text patterns	High	Medium	Applied robust preprocessing and used a transformer model suited for mixed languages.
Limited hardware resources for large-scale model training	Medium	Low	Optimized batch size and model layers; used GPU where possible.
Unfamiliarity with ADB or Android extraction process	Medium	High	Planned to learn period and testing with dummy device data before full integration.
Time constraints with academic deadlines	High	High	Prioritized core features (model development); tool components planned in phases.

Table 5 - Chapter 3: Table of Risk Analysis and Mitigation

3.5. Chapter Summary

This chapter outlines the research and development methodology use to develop a multilingual forensic tool to detecting cyberbullying in Sri Lanka. Firstly, the research philosophy and approach based on the Research Onion model is elaborated, emphasizing a pragmatic and deductive strategy supported by quantitative evaluation. Following that the chapter further details how requirements were collected through both user survey and analysis of existing tools, building up the foundation for system design and planning.

Furthermore, the development methodology of the tool which is focused on developing a hate speech detection model using a fine-tuned BERT classifier, trained on Sinhala and Sinhala-English code-mixed datasets. Although the full CLI-based forensic tool has not been implemented at this stage, its design and planned integration were clearly defined. Tool components such as the command-line interface, Android

data extraction, and report generation are scheduled for development in the upcoming phase.

Finally, the chapter further covers the project management strategies, including scope definition, timeline planning, resource allocation, and risk mitigation. In conclusion, with the core model completed and a well-structured roadmap in place, the project is on track to achieve its final goals in the next development phase.

CHAPTER 04 : SOFTWARE REQUIREMENTS SPECIFICATION

4.1. Chapter Overview

This chapter presents the Software Requirements Specification (SRS) for the proposed multilingual Android forensic tool aimed at detecting cyberbullying through hate speech analysis in Sinhala and Sinhala-English code-mixed content. The chapter outlines the purpose of the system, target users, functional expectations, and non-functional requirements. Furthermore, it includes visual representations such as a rich picture, stakeholder models, use case diagrams, and prioritization tables to elaborate both the scope and the intended functionality of the system. Additionally, the content is based on insights gathered from literature, user surveys, and stakeholder evaluation to ensure the tool meets the real-world forensic requirements in the Sri Lankan context.

4.2. Rich Picture

The rich picture below illustrates the real-world context in which the proposed multilingual Android forensic tool will operate. Moreover, it visually represents the relationships between key stakeholders involved in a cyberbullying investigation, the core problem of online abuse, and how the tool integrates into the existing forensic process.

Firstly, the diagram begins with the victim, who experiences emotional distress due to bullying and hates speech on social media platforms like Facebook and Instagram. Upon reporting the incident to the police, the prime suspect, the criminal is identified and taken into custody. The suspect's mobile device is seized and handed over to the cybercrime investigation unit.

At this stage, the proposed forensic tool comes into play. It is designed to extract mobile social media data manually through ADB, analyze text-based content, and detect hate speech, particularly in Sinhala and Sinhala-English code-mixed

formats. Once the tool processes the data, it generates a structured report highlighting detected hate speech instances.

Finally, this report is then passed back to the police, supporting them with clear, contextual evidence that can be used in a legal investigation and a court case. The rich picture highlights how the tool fills a critical gap in existing cybercrime workflows, enabling more effective evidence gathering and decision-making in Sri Lankan forensic environments.

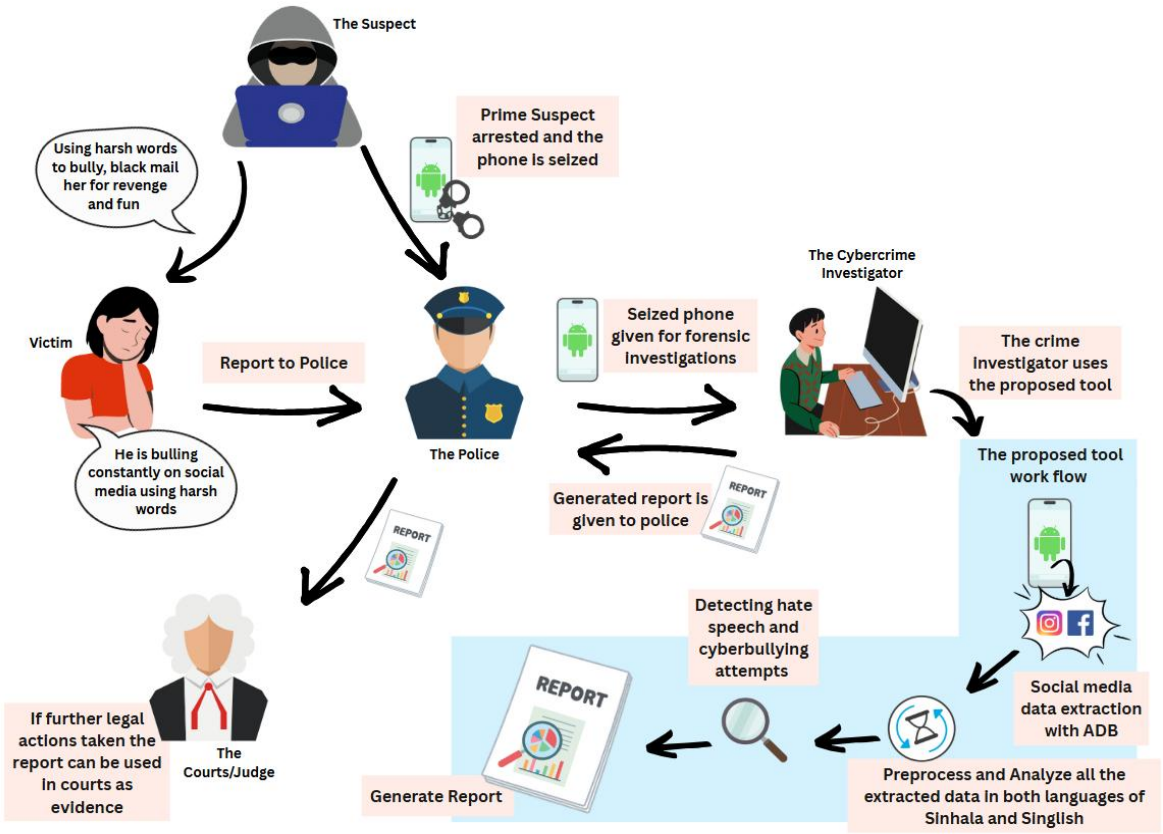


Figure 3 - Chapter 4: Rich Picture

4.3. Stakeholder Analysis

4.3.1. Stakeholder Description

Stakeholder	Stakeholder Type	Description
Cybercrime Investigator	Primary Operator	Direct user of the tool who is responsible for running the analysis, interpreting results, and preparing evidence for legal proceedings.
Police Officer	Support Operator	Acts as the first responder in collecting digital devices from suspects and coordinates with cybercrime units for forensic analysis.
Victim	Affected Individual	The target of online harassment or bullying and indirectly benefits from the investigation and evidence collected using the tool.
Criminal/Suspect	Subject of Investigation	The individual is under investigation for cyberbullying. Their device is analyzed using the tool.
Court/Judge	Decision-Maker	Use the forensic report produced by the tool as part of legal proceedings and evidence evaluation.
General Public	External Stakeholder	Indirectly impacted by the success of such tools in reducing cyberbullying and improving online safety.
Social Media Platforms (e.g., Facebook, Instagram)	External Data Source	Platforms from which communication data is extracted and analyzed by the tool.

Table 6 - Chapter 4; Table of Stakeholder Descriptions

4.3.2. Stakeholder Onion Model

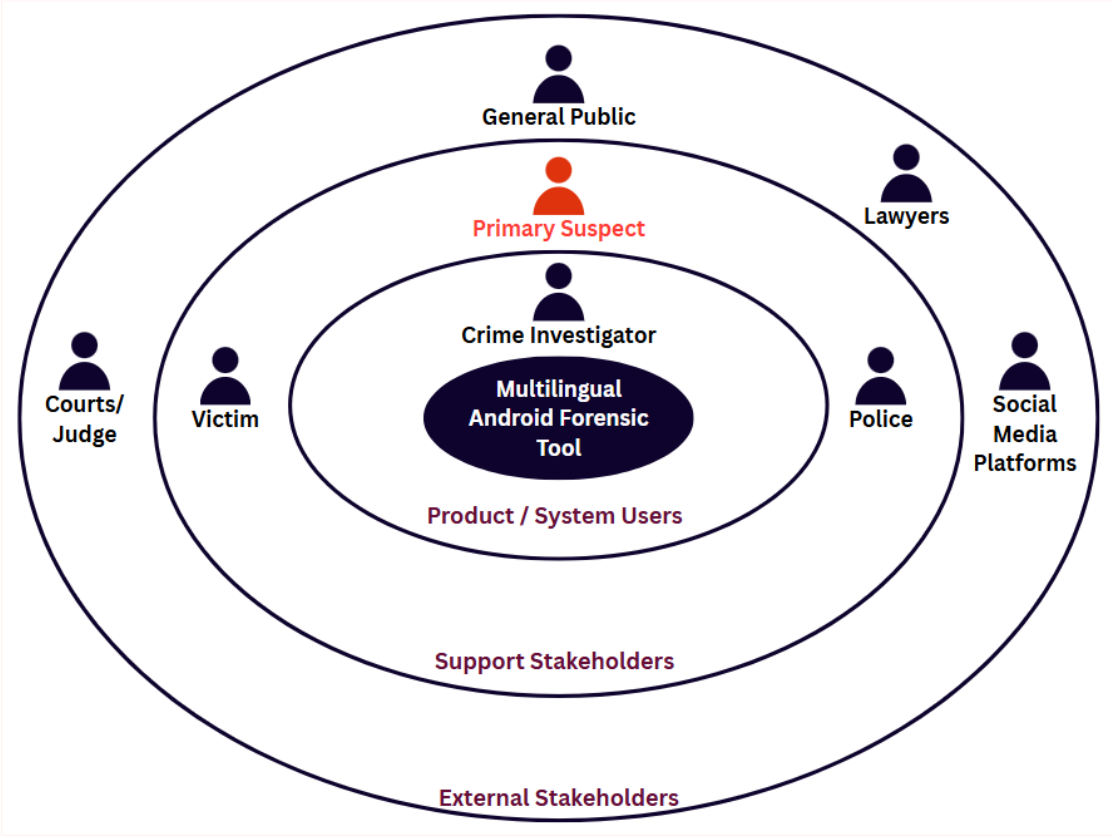


Figure 4 - Chapter 4: Stakeholder Onion Model

4.4. Requirement Elicitation Methods

Selected Method	Justification
Literature Review	Conducting a literature review helped to understand the current landscape of existing tools, technologies, and research gaps. Moreover, this method supported the technical direction of the project by justifying the need for a localized, multilingual forensic tool in Sri Lanka.

Public Survey	The survey allowed for gathering real-world input from a diverse set of potential social media users of all sorts and it helped to identify user needs, expectations, and crucial points related to cyberbullying and digital forensics in the Sri Lankan context.

Table 7 - Chapter 4: Table of Requirement Elicitation Methods

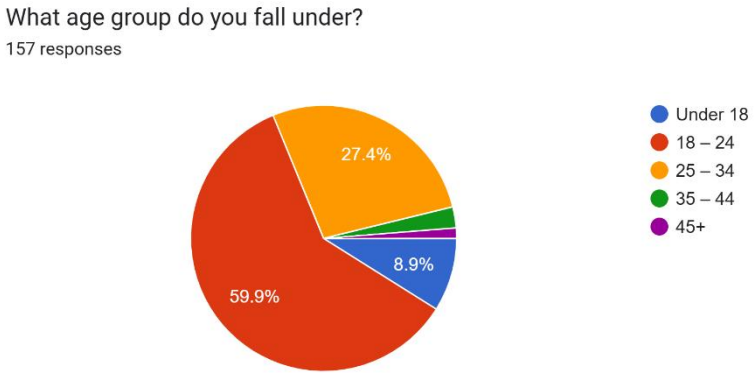
4.5. Discussion of Results

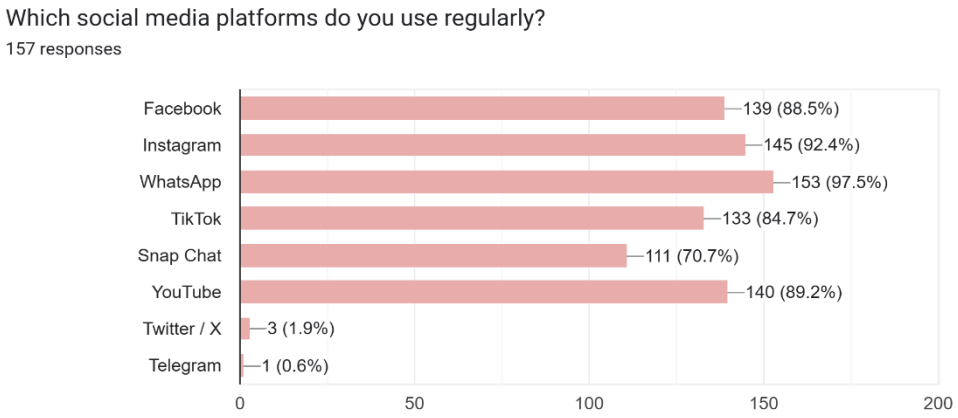
4.5.1. Literature Review

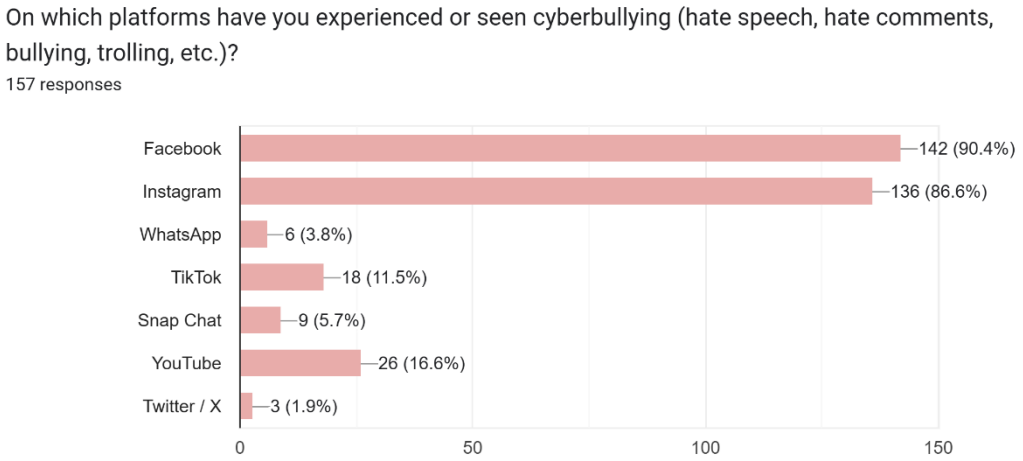
Citation(s)	Findings
Abu Hweidi et al. (2023), Chang & Yen (2020)	Most forensic tools used in Sri Lanka are expensive, foreign-developed, and do not support Sinhala or code-mixed content, creating a gap in local investigation capabilities.
Ruwandika & Weerasinghe (2018), Ishara & Jayalal (2020)	Traditional ML models like SVM and logistic regression underperform with code-mixed Sinhala text and struggle to generalize.
Muthuthanthri & Smith (2024), Fernando & Deng (2023)	BERT and transformer-based models significantly improve hate speech detection accuracy, especially in multilingual or code-mixed contexts.
Jayasinghe et al. (2024), Kemp (2024)	Sri Lankan youth are the most vulnerable to cyberbullying on platforms like Facebook and Instagram, often in Sinhala-English hybrid language.
Asim Mubarik et al. (2021), Ogunleye & Dharmaraj (2023)	There is a lack of integrated tools that combine data extraction, NLP-based analysis, and structured evidence reporting in a forensic workflow.

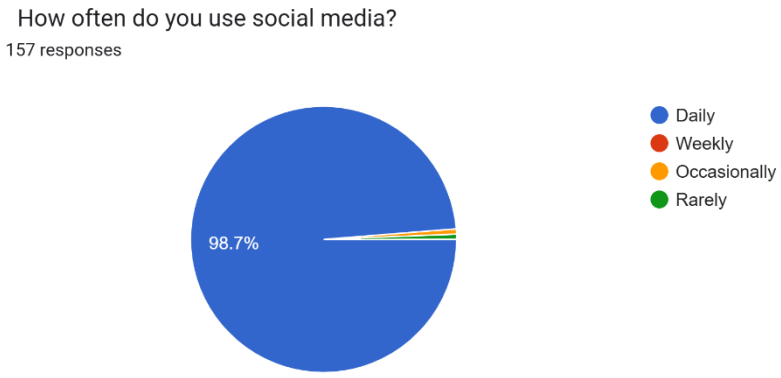
Table 8 - Chapter 4: Table of Literature Review Findings

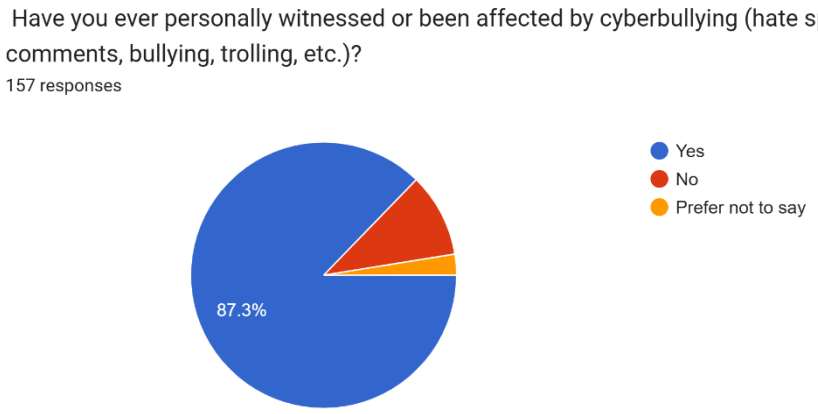
4.5.2. Survey Findings

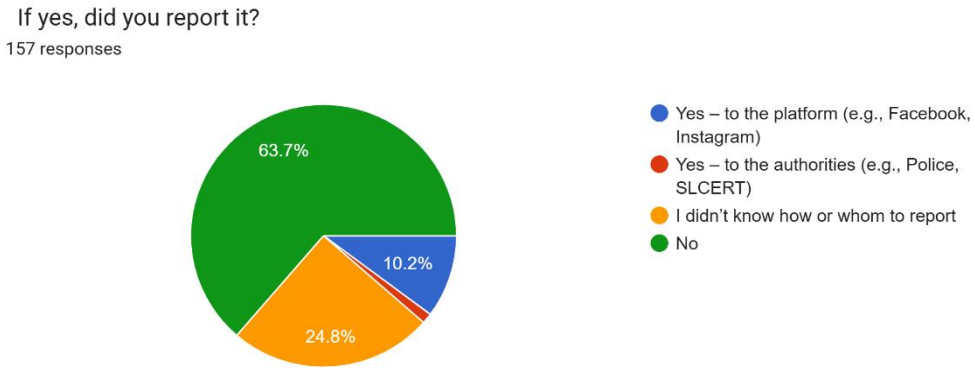
Question	What age group do you fall under?
Purpose	The purpose of this question was to identify the dominant age demographic of social media users in Sri Lanka who may be exposed to, witnessed or affected by cyberbullying. Understanding the age group distribution helps ensure that the tool is designed to address the requirements of the most vulnerable and most active users online.
	
Discussion	Based on the 157 responses collected, it is highlighted that the majority (59.9%) of the participants fall within the 18–24 age range, followed by 27.4% in the 25–34 age range. This clearly shows that young adults are the most active on social media and are therefore most likely to encounter and be affected by cyberbullying. Moreover, the respondents under 18 (8.9%) and above 34 are very less compared to the other two age groups, confirming that the primary users remain as the youth and young adults. Furthermore, these findings directly support the decision to focus the tool on platforms like Facebook and Instagram, where these age groups are highly active. It also pops out the fact that it is importance addressing this issue as these social media platforms are commonly used by younger users in Sri Lanka.

Question	Which social media platforms do you use regularly?
Purpose	The aim of this question was to identify which social media platforms are most commonly used in Sri Lanka. Moreover, this helps to determine which platforms should be prioritized during data extraction and tool compatibility design, ensuring the tool is developed around platforms where cyberbullying is most likely to occur.
	
Discussion	The results show that WhatsApp (97.5%), Instagram (92.4%), and YouTube (89.2%) are the most commonly used platforms among respondents, followed closely by Facebook (88.5%) and TikTok (84.7%). Platforms like Snapchat (70.7%) also show notable usage, while Twitter/X (1.9%) and Telegram (0.6%) have very low engagement among this group of people. However, these findings strongly validate the decision to focus the forensic tool on platforms like Facebook and Instagram, where both message-based and public interaction-based bullying are frequent. Even though WhatsApp is ranked the highest, it involves more private encrypted messaging, which requires different extraction techniques and legal considerations which are outside the scope of this project. Twitter and Telegram are not a development priority due to their low usage in this

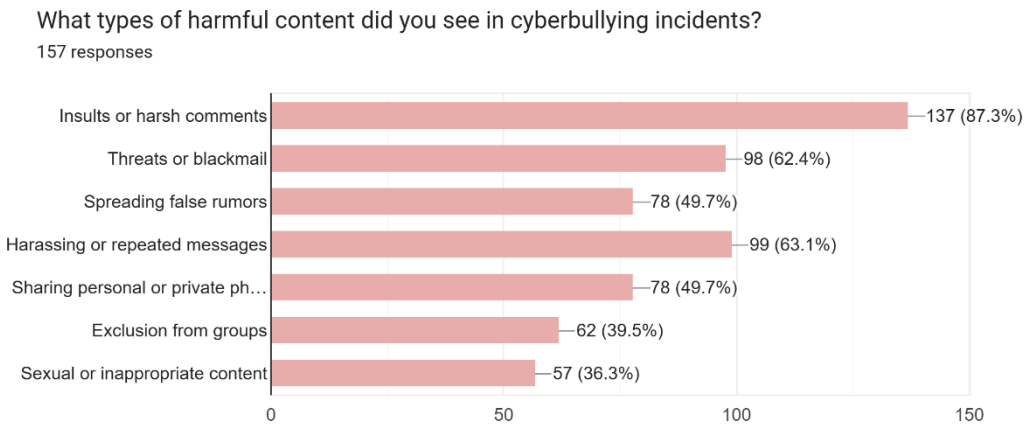
	demographic. Finally, the results help to strengthen relevance of the chosen platforms in the project scope and ensure that the tool focuses on real-world, highly used environments.
Question	**On which platforms have you experienced or seen cyberbullying (hate speech, hate comments, bullying, trolling, etc.)?**
Purpose	This question was designed to identify which social media platforms are most frequently associated with cyberbullying incidents. Understanding this is crucial and it helps the project focus its data extraction and hate speech detection efforts on the most relevant platforms where harmful content is most commonly encountered.
	
Discussion	Out of 157 respondents, a striking 90.4% reported experiencing or witnessing cyberbullying on Facebook, followed closely by Instagram (86.6%). This clearly confirms that these two platforms are the primary digital spaces where online abuse is most visible among the surveyed group. Moreover, other platforms, like YouTube (16.6%) and TikTok (11.5%), show some reports of cyberbullying but with comparatively lower rates. Furthermore, WhatsApp (3.8%), Snapchat (5.7%), and Twitter/X (1.9%) were reported as much less

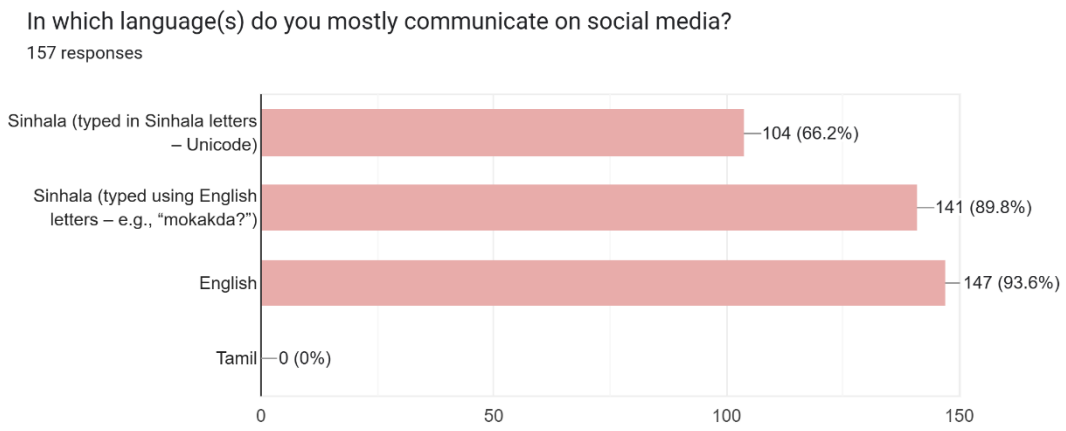
	common spaces for visible bullying in this survey. However, these findings further support the decision to prioritize Facebook and Instagram in the development of the tool, especially when designing modules for data extraction, classification, and report generation. The results further support the inclusion of features that detect various forms of hate speech, trolling, and harassment, which appear to be widespread on these platforms.
Question	**How often do you use social media?**
Purpose	This question aims to determine the frequency of social media usage among respondents and understanding the usage patterns helps to justify the requirement for real-time forensic investigation tools and supports the decision to prioritize social media platforms where users are highly active.
	
Discussion	Out of the 157 participants, a solid 98.7% reported using social media daily. Only a very small percentage indicated weekly, occasional, and rare usage. This result clearly highlights that social media is an integral part of daily life for the majority of the surveyed population. The high level of daily engagement highlights that the risk of encountering cyberbullying is also frequent and ongoing.

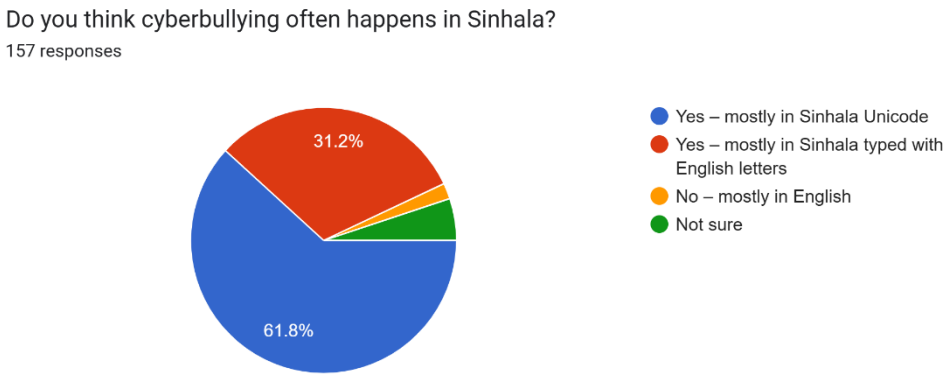
	Additionally, this finding validates the choice to focus on platforms like Facebook and Instagram, which are checked and used daily by most users. It further elaborates the urgency of implementing systems that helps the investigators to quickly identify harmful interactions in a timely manner.
Question	**Have you ever personally witnessed or been affected by cyberbullying (hate speech, hate comments, bullying, trolling, etc.)?**
Purpose	The purpose of this question is to measure how many individuals have direct or indirect experience with cyberbullying and helps to validate the real-world relevance of the problem which the proposed forensic tool is trying to solve.
	
Discussion	According to the responses, 87.3% of participants reported that they have either personally experienced or witnessed some form of cyberbullying, while only a small percentage responded "No" and chose "Prefer not to say". This clearly points out that cyberbullying is widespread and affects a large majority of social media users in Sri Lanka. Moreover, the result strongly justifies the need for a tool that is capable of detecting hate speech and abusive language, especially in Sinhala and code-mixed formats. Furthermore, it supports the idea

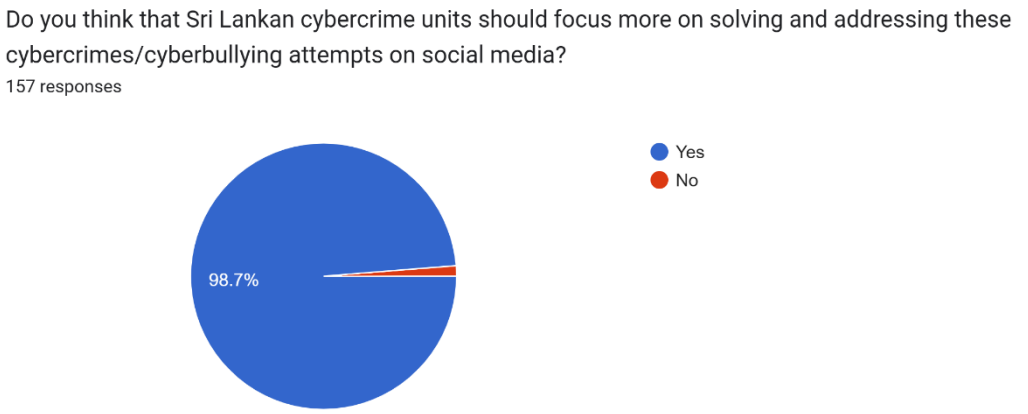
	that such a tool could have a real social impact by helping victims seek justice and enabling investigators to gather evidence more effectively. However, the finding strengthens the urgency and importance of developing localized solutions to address this issue.
Question	**If yes, did you report it?**
Purpose	This question aims to understand how the victims or witnesses of cyberbullying respond to incidents and whether they take any action. Additionally, it further reveals the awareness and accessibility of the existing reporting mechanisms, helping to determine whether users trust the current systems in place.
	
Discussion	Among the 157 respondents, a significant 63.7% admitted that they did not report the cyberbullying incident at all. Meanwhile, 24.8% indicated that they did not know how or whom to report to, highlighting a major gap in awareness. With all that only a small portion reported the incident to the platform itself (10.2%), and an even smaller percentage went to the authorities (1.3%). Moreover, this result is highly crucial, as it reflects both a lack of reporting and a disconnect between social media users and available support systems. It point outs that despite the high rate of cyberbullying exposure, most victims or witnesses feel unsure, helpless, and

	unwilling to report. However, for this project, this finding supports the need to develop a forensic tool empowers authorities by giving them clearer, locally understandable evidence of abuse, especially since users themselves are often not taking direct action. It further illustrates the possible use of such tools to fill in the reporting gap by conducting the back-end investigations when front-end reporting fails.
Question	**If you didn't report it, why not?**
Purpose	This question is to further clarify the reasons behind hesitation or refusal of the users to report cyberbullying incidents. It helps to identify the emotional, social, and practical barriers that prevent users from seeking help, which in turn supports the development of backend tools that enable authorities to intervene even when cases go unreported.
	
Discussion	Out of the few responses collected, common themes included uncertainty, fear of involvement, feeling it was not serious enough, and believing there was no point. Specific reasons included: "I didn't know what to do." "Didn't want to get in unnecessary trouble." "No point."

	"Didn't think of it much because I was not bullied, although I witnessed it." These answers highlight a pattern of low confidence in reporting and fear of retaliation and social conflict. It further reflects a lack of awareness about support channels and legal protections.
Question	**What types of harmful content did you see in cyberbullying incidents?**
Purpose	This question is designed to identify the most common forms of harmful and abusive content experienced or witnessed by active social media users. Moreover, these findings are critical for understanding what kinds of hate speech and bullying behaviors the tool needs to detect and categorize effectively.
	
Discussion	These responses reveal that the most frequently seen form of harmful content was "insults or harsh comments" (87.3%), followed by "harassing or repeated messages" (63.1%), and "threats or blackmail" (62.4%). A specific portion of users also reported exposure to spreading false rumors (49.7%), sharing of personal/private photos (49.7%), and exclusion from groups (39.5%). Lastly, 36.3% had seen sexual or inappropriate content. With that, these critical findings

	show that text-based abuse is by far the most common and damaging form of cyberbullying by validating the focus of the project on natural language processing (NLP) for hate speech detection. Additionally, it further suggests that any future expansions of the tool may need to consider other aspects of bullying, such as media-sharing and group manipulation, though these are currently out of scope. However, the results directly support the need for a system capable of recognizing toxic language, verbal violence across Sinhala and code-mixed texts.
Question	**In which language(s) do you mostly communicate on social media?**
Purpose	This question was asked to determine the most commonly used languages and input formats in online communication among Sri Lankan social media users. This helps define the language scope of the model and supports the need for multilingual and code-mixed language processing in the tool.
	
Discussion	Out of the157 responses recorded 93.6% communicates in English, 89.8% use Sinhala typed using English letters (commonly known as Sinhala-English code-mixed or "Singlish"), 66.2% use Sinhala Unicode (typed in Sinhala letters) and 0% selected Tamil. Additionally, these findings confirm the widespread use of code-

	mixed Sinhala-English interaction highlighting the importance of the tool supporting informal transliterated Sinhala content in addition to Unicode Sinhala and standard English. However, it further supports the decision to focus on Sinhala and code-mixed Sinhala-English as the primary language inputs for the BERT-based hate speech detection model. The lack of Tamil responses suggests that including Tamil language support is not immediately necessary for this version of the tool, though it could be considered for future expansion. Overall, the finding strongly justifies the multilingual focus and language preprocessing strategy defined in the system design.
Question 11	**Do you think cyberbullying often happens in Sinhala?**
Purpose	This question was asked to assess perceptions of the participants about the language usage nature on cyberbullying specially whether abusive behavior online frequently occurs in Sinhala, including Unicode or code-mixed formats. The aim was to validate the need for local language support in hate speech detection.
	
Discussion	According to the results, 61.8% of participants believe that cyberbullying mostly happens in Sinhala Unicode, while 31.2% believe that it occurs mainly in Sinhala typed with English letters (code-mixed format) meanwhile a very small percentage selected "mostly in English" (1.3%) and "not sure" (5.7%). With that the

	majority confirms that Sinhala is the most dominant language used for cyberbullying. Furthermore, the insight directly supports the core requirement of this project to detect hate speech in Sinhala and Sinhala-English code-mixed messages, which are often not supported by existing forensic tools. This finding strengthens the language scope of the project justifying valid reasons for choosing a multilingual NLP model and customized preprocessing steps tailored to handle both Unicode Sinhala and transliterated Sinhala.
Question 12	**Do you think that Sri Lankan cybercrime units should focus more on solving and addressing these cybercrimes/cyberbullying attempts on social media?**
Purpose	The purpose of this question was to assess public opinion on the role of Sri Lankan cybercrime units in tackling social media-based cyberbullying. This helps to validate the importance of developing a tool that directly assists law enforcement agencies in identifying and handling such cases more effectively.
	
Discussion	Out of 157 responses, a solid 98.7% of participants said "Yes", highlighting a public demand for stronger enforcement and investigation into cybercrimes, particularly on social media platforms. Meanwhile, only a very small minority (1.3%) disagreed.

	However, this high level of agreement strongly supports the relevance and urgency of the proposed project. Moreover, it highlights that users not only recognize the seriousness of online bullying but also expects proactive action from local authorities. Furthermore, the findings support the development of this tool as a direct enabler for cybercrime units, helping them to bridge the gaps in language support, data analysis, and digital reporting when dealing with Sinhala and code-mixed cyberbullying content.

Table 9 - Chapter 4: Table of Survey Findings

4.6. Summary Findings

Findings	Literature Review	Survey Observation
Most cyberbullying cases occur on Facebook and Instagram.	✅	✅
Sinhala and Sinhala-English code-mixed content are the most common languages in online communication	✅	✅
Existing forensic tools do not support Sinhala or Sinhala-English code-mixed text.	✅	
Majority of victims do not report cyberbullying due to fear, confusion, and lack of awareness		✅
Public expects stronger involvement from Sri Lankan cybercrime units in tackling cyberbullying.		✅
Hate speech is the most frequent form of bullying, followed by harassment and threats.	✅	✅
Transformer models like BERT show improved performance for multilingual hate speech detection.	✅	

Table 10 - Chapter 4: Table of Summary Findings

4.7. Context Diagram

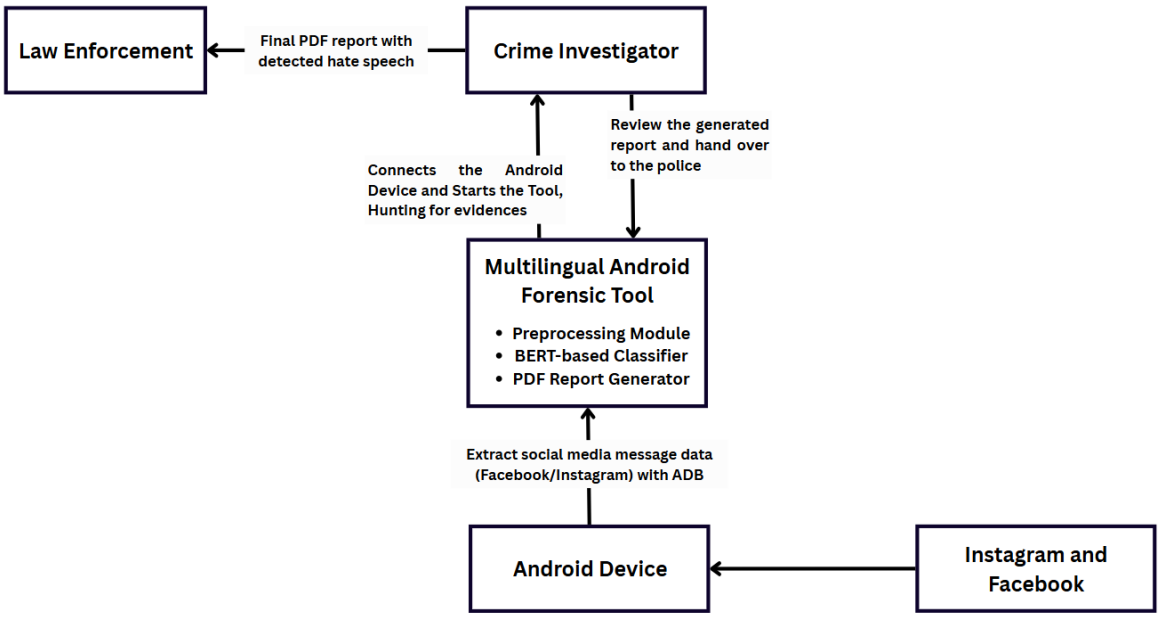


Figure 5 - Chapter 4: Context Diagram

This context diagram illustrates the interaction between the proposed system and its external entities. The crime investigator operates the tool by extracting social media data from an Android device using ADB. The tool internally preprocesses the data, classifies hate speech using a BERT-based model, and generates a structured PDF report. The investigator reviews this report and forwards it to law enforcement for further action. Instagram and Facebook are indirectly involved as the original data sources stored on the device.

4.8. Use Case Diagram and Description

No	Use Case ID	Description
1	Use Case Name	Extract Social Media Data (ADB)
	Actor	Crime Investigator, Android Device

	Description	Extracts chat/message data from Facebook and Instagram stored on a suspect's Android device using ADB.
	Precondition	Android device is connected, and ADB is functional.
	Postcondition	Data is successfully extracted and saved for preprocessing.
	Main Flow	• Investigator connects Android device. • System initiates ADB session. • Chat database is located and pulled. • Data is stored locally for analysis.
2	Use Case Name	Preprocess Extracted Text
	Actor	System
	Description	Cleans and formats the raw message data (Sinhala Unicode and code-mixed text) for NLP processing.
	Precondition	Raw data extracted and accessible.
	Postcondition	Text is structured and ready for classification.
	Main Flow	• Read raw text files. • Clean symbols. • Normalize Sinhala and Singlish. • Save preprocessed data.
3	Use Case Name	Run Hate Speech Classification
	Actor	Crime Investigator
	Description	Runs the BERT model to detect hate speech in preprocessed messages.
	Precondition	Preprocessed data must be available.
	Postcondition	Messages are classified with hate or non-hate labels.

	Main Flow	• Investigator initiates classification. • Data passed to BERT model. • Predictions returned and saved.
4	Use Case Name	Generate Report (PDF)
	Actor	Crime Investigator
	Description	Generates a structured PDF report from classified data.
	Precondition	Classified messages are available.
	Postcondition	A printable, shareable report is created.
	Main Flow	• Investigator selects generate report. • System compiles flagged content. • PDF is formatted and saved.
5	Use Case Name	View or Review Results
	Actor	Crime Investigator
	Description	Investigator views and verifies flagged content before finalizing the report.
	Precondition	Classification must be completed.
	Postcondition	Verified dataset is saved for report generation.
	Main Flow	• System displays flagged messages. • Investigator reviews each message. • Adjust selections if needed. • Confirms for reporting.

Table 11 - Chapter 4: Table of Usecase Diagram Description

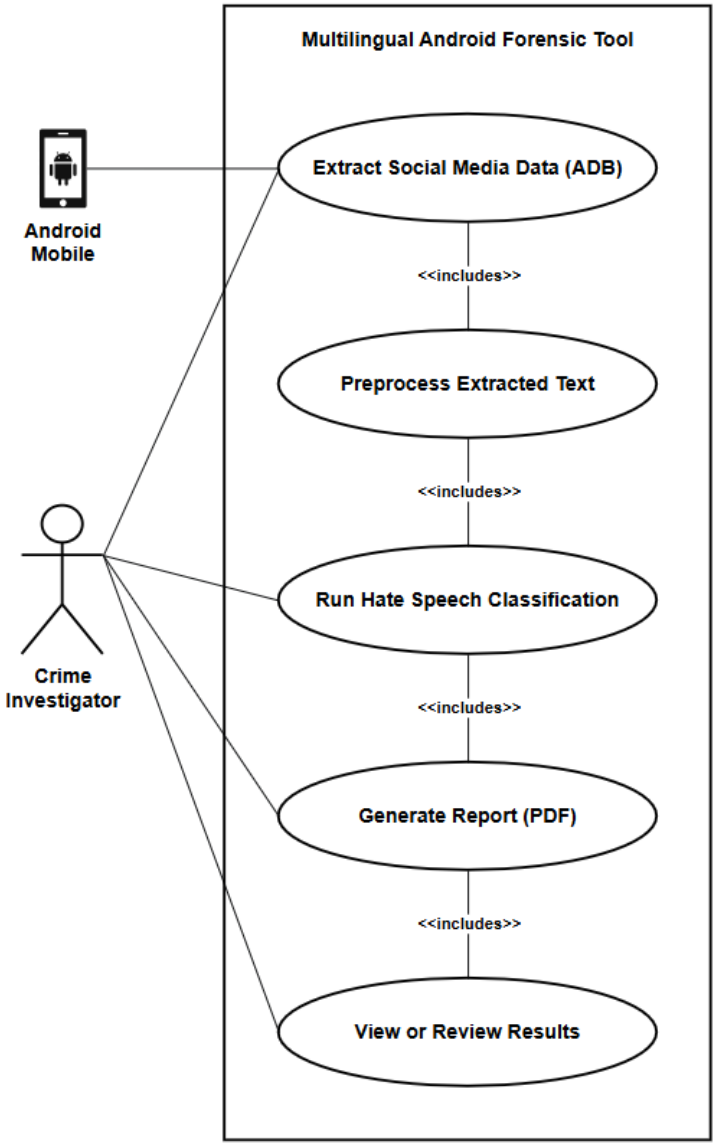


Figure 6 - Chapter 4: Usecase Diagram

4.9. Requirements

4.9.1. Prioritization

Must have	- Extract data from Android device (ADB) - Preprocess Unicode & code-mixed text - Run hate speech classification (BERT) - Generate PDF report with results

Should have	- Basic error messages/logs - Command-line progress feedback - Support for multi-file input
Could have	- Optional GUI layer - Export results in CSV format - Offline user manual/help flag
Will not Have	- iOS device support - Live social media monitoring - Multimedia (image/video) content analysis

Table 12 - Chapter 4: Table of Requirements Prioritization

4.9.2. Functional Requirements

S. No	Requirement	Prioritization	Use case mapping
1	The system shall allow investigators to extract social media data from Android devices using ADB.	Must Have	UC1
2	The system shall preprocess Unicode Sinhala and code-mixed Sinhala-English content.	Must Have	UC2
3	The system shall classify messages as hate or non-hate using the trained BERT model.	Must Have	UC3
4	The system shall generate a structured PDF report containing flagged messages.	Must Have	UC4
5	The system shall accept manually entered text input (for testing).	Should Have	UC2 (extends)
6	The system shall display basic feedback in the CLI interface (e.g., status).	Should Have	UC3, UC4 (optional)

7	The system shall export results in both PDF and CSV formats.	Could Have	UC4

Table 13 - Chapter 4: Table of Functional Requirements

4.9.3. Non-functional Requirements

S. No	Non-functional requirements	Prioritization
1	The system should be usable via CLI with minimal technical knowledge.	Must Have
2	The tool should run efficiently on mid-range laptops with 8–16 GB RAM.	Must Have
3	The output should be legally usable and readable (PDF format).	Must Have
4	The system should support offline operation (no internet dependency).	Should Have
5	The system should be modular for future extension (e.g., more languages).	Could Have
6	The system shall include brief CLI documentation/help flags.	Could Have

Table 14 - Chapter 4: Table of Non-functional Requirements

4.10. Chapter Summary

This chapter documents the detailed software requirement specifications for the proposed multilingual Android forensic tool. Starting off with a rich picture and stakeholder analysis to visually represent the ecosystem surrounding the system and to identify all relevant actors. Then a combination of literature review and survey findings which help to guide the functional and non-functional requirements, while the stakeholder onion model and survey data highlighted the practical need for localized

cyberbullying detection. The context diagram and usecase diagram illustrates how the system will interact with external entities and what actions it will perform. Moreover, the chapter also outlines the prioritization of system features and maps each requirement to its respective use case. Finally, the combination of these sections provides a clear blueprint for building and evaluating the system throughout its development lifecycle.

CHAPTER 05 : CONCLUSION

5.1. Chapter Overview

This conclusion chapter summarizes the outcomes and progress achieved during the initial development phase of the multilingual Android forensic tool. Moreover, it outlines the challenges faced, how they were solved, any modifications from the original plan, and presents proof of concept for the implemented components. The initial results are briefly discussed in order to reflect the model's functionality, followed by a link to a demo video showcasing the active workflow of the system. With that this chapter wraps up the midpoint deliverables and provides a foundation for the next phase of tool development and testing.

5.2. Problem Encountered and Solution

Problem	Description	Solution
Lack of Balanced Datasets	Difficulty in finding complete and class-balanced Sinhala and Sinhala-English code-mixed hate speech datasets.	Combined two separate datasets and applied custom preprocessing to create a unified, usable dataset.
Limited Computational Resources	Fine-tuning the BERT model was slow and resource-heavy on personal hardware.	Used GPU acceleration, optimized batch size, and limited training epochs to maintain performance.
Complexity in Android Data Extraction (ADB)	Initial struggles with setting up ADB, understanding where social media data is stored on Android devices and also behavior of Android Forensics.	Focused on testing with mock data while studying ADB behavior, commands and file paths therefore full integration planned in final phase.
Challenges in Handling Code-	Inconsistencies in spelling, language mixing, and informal	Applied regex based normalization and language

Mixed Sinhala-English Text	structures made preprocessing difficult.	specific text cleaning techniques.

Table 15 - Chapter 5: Table of Problem Encountered and Solution

5.3. Deviations

When working on the project, similar to most real-world projects, it didn't go exactly as planned therefore, a few changes had to be made along the way to manage time, technical complexity, and unexpected learning curves. Among the changes one of the biggest adjustments was postponing the Android data extraction feature. Originally, the tool was supposed to connect to a device and extract data using ADB by this phase. However, by digging deep into forensics I realized that modern Android file structures and app permissions are a bit harder to figure and consume more time and research. Due to this the development of this extraction feature was postponed to the next phase.

Another area that shifted slightly was the CLI interface which was originally supposed to be partially built and functional by now, more time was spent on improving the dataset and model accuracy. So, due to the time limitations the CLI part is still in the planning stage and will be handled too in the next phase.

Despite these two changes, the main goal was achieved which is developing a working, reliable hate speech detection model for Sinhala and Sinhala-English code-mixed content. However, these small deviations were necessary adjustments to maintain the quality and foundation of the final tool.

5.4. Proof of Concept

In order to validate the feasibility of the proposed multilingual Android forensic tool, a functional proof of concept was developed and tested. The main goal at this stage was to build a hate speech detection model capable of analyzing Sinhala and Sinhala-English code-mixed social media text, which serves as the foundation of the entire tool.

However, the BERT-based classification model was successfully trained using two publicly available datasets which was a Sinhala Unicode hate speech dataset and a

code-mixed Sinhala-English dataset. The model demonstrated its ability to identify hate speech patterns in both languages after appropriate preprocessing. While the full integration of the tool is still ongoing, the classification module is already producing good results in terms of detection accuracy and adaptability to local language use.

This proof of concept proves that localized hate speech detection using NLP is technically viable, and it validates the core approach of the project. With that the next phase will involve wrapping this model within a CLI tool, integrating data extraction from Android devices, and generating structured PDF reports.

Codes

The appendix includes screenshots of three code files

- Code for Merging 3 Datasets : This includes the code used to merge three datasets into one data set for model training. Due to the lack proper data set in Sinhala Unicode hate speech and English-Sinhala transliterated hate speech, two Sinhala datasets and one English-Sinhala transliterated dataset was combined. View Code
- Code for Training the Model : This includes the code used to train the BERT model to detect Sinhala Unicode and English-Sinhala transliterated hate speech on social media. View Code
- Code for the Dummy Cli Tool : This includes the code used to develop the dummy cli tool to check the functionality of the model trained. View Code

Output

Moreover the appendix includes the screenshots of the outcome of the dummy tool and the model. View Output.

5.5. Initial Results

At this stage of the project, the core model has been successfully trained using two local datasets one in Sinhala Unicode and the other in Sinhala-English code-mixed format. The fine-tuned BERT model is able to classify hate speech with good accuracy, showing strong potential in identifying harmful content written in local languages. Early

testing on sample data confirms that the system can detect a variety of cyberbullying behaviors including explicit insults, harassment, and hate-driven speech patterns. While a detailed accuracy evaluation will be performed in the final phase, the current results are encouraging and support the continuation of development into the full CLI tool.

5.6. Demo Video Link

Unlisted YouTube video link of the Hawkeye Multilingual Android Forensic Tool to Detect Cyberbullying Attempts through Hate Speech Detection

https://youtu.be/bRi5PfW41S8

5.7. Chapter Summary

This chapter presents the conclusion of the project's first development phase. Moreover, it outlines the key problems encountered, justified the small deviations from the original plan, and confirmed the feasibility of the proposed tool through a working proof of concept. Furthermore, the initial model testing shows strong potential, especially in identifying hate speech in Sinhala and Sinhala-English code-mixed content. Although certain parts of the tool are still in development such as the Android extraction and CLI interface the foundation is stable, and the direction is clear for the next stage of implementation.

REFERENCES

Abu Hweidi, R.F., Jazzar, M., Eleyan, A. and Bejaoui, T., 2023. Forensics Investigation on Social Media Apps and Web Apps Messaging in Android Smartphone. In: *2023 International Conference on Smart Applications, Communications and Networking (SmartNets)*. [online] 2023 International Conference on Smart Applications, Communications and Networking (SmartNets). Istanbul, Turkiye: IEEE. pp.1–7. https://doi.org/10.1109/SmartNets58706.2023.10216267.

Al-Garadi, M.A., Hussain, M.R., Khan, N., Murtaza, G., Nweke, H.F., Ali, I., Mujtaba, G., Chiroma, H., Khattak, H.A. and Gani, A., 2019. Predicting Cyberbullying on Social Media in the Big Data Era Using Machine Learning Algorithms: Review of Literature and Open Challenges. *IEEE Access*, 7, pp.70701–70718. https://doi.org/10.1109/ACCESS.2019.2918354.

Amali, H.M.A.I. and Jayalal, S., 2020. Classification of Cyberbullying Sinhala Language Comments on Social Media. In: *2020 Moratuwa Engineering Research Conference (MERCon)*. [online] 2020 Moratuwa Engineering Research Conference (MERCon). Moratuwa, Sri Lanka: IEEE. pp.266–271. https://doi.org/10.1109/MERCon50084.2020.9185209.

Amjad, M., Ashraf, N., Zhila, A., Sidorov, G., Zubiaga, A. and Gelbukh, A., 2021. Threatening Language Detection and Target Identification in Urdu Tweets. *IEEE Access*, 9, pp.128302–128313. https://doi.org/10.1109/ACCESS.2021.3112500.

Ariyadasa, A., 2019. Harassment Beyond Borders: Sexting, Cyber Bullying and Cyber Stalking in Social Media. Can Sri Lanka Protect Victims? *SSRN Electronic Journal*. [online] https://doi.org/10.2139/ssrn.3382683.

Barukanda, B., 2024. Over 9,000 cybercrime complaints within two months – SLCERT – Read Sri Lanka. Available at: <https://readsrilanka.com/2024/10/07/over-9000-cybercrime-complaints-within-two-months-slcert/> [Accessed 2 April 2025].

Cellebrite, 2025. *Cellebrite Forensics*. [online] Cellebrite Forensics. Available at: <https://cellebrite.com/en/home/> [Accessed 23 January 2025].

Chang, M.S. and Yen, C.P., 2020. Evidence Gathering of Facebook Messenger on Android.

Chathurangi, M.D.D., Nayanathara, M.G.K., Gunapala, K.M.H.M.M., Dayananda, G.M.R.G., Abeywardena, K.Y. and Siriwardana, D., 2024. Detecting Cyberbullying, Spam & Bot Behavior and Fake News in Social Media Accounts Using Machine Learning. In: *Geometric and Algebraic Properties of the Eigenvalues of Monotone Matricesg*. Proceedings Nonnegative Matrices and Finite Markov Chains 2024. Italy: International Research Conference Proceedings. pp.17–23.

Fernando, E.N. and Deng, J.D., 2023. Enhancing Hate Speech Detection in Sinhala Language on Social Media using Machine Learning.

Gohal, G., Alqassim, A., Eltyeb, E., Rayyani, A., Hakami, B., Al Faqih, A., Hakami, A., Qadri, A. and Mahfouz, M., 2023. Prevalence and related risks of cyberbullying and its effects on adolescent. *BMC Psychiatry*, 23(1), p.39. https://doi.org/10.1186/s12888-023-04542-0.

Heshan Maduranga, 2024. *Cybercrime Analysis - Sri Lanka*. https://doi.org/10.13140/RG.2.2.18522.93123.

Jayasinghe, Y.A., Kanmodi, K.K., Jayasinghe, R.M. and Jayasinghe, R.D., 2024. Assessment of patterns and related factors in using social media platforms to access health and oral health information among Sri Lankan adults, with special emphasis on promoting oral health awareness. *BMC Public Health*, 24(1), p.1472. https://doi.org/10.1186/s12889-024-19008-5.

Jones, G.M., Winster, S.G. and Valarmathie, P., 2022. Integrated Approach to Detect Cyberbullying Text: Mobile Device Forensics Data. *Computer Systems Science and Engineering*, 40(3), pp.963–978. https://doi.org/10.32604/csse.2022.019483.

Kemp, S., 2025. *Digital 2025: Sri Lanka*. [online] DataReportal – Global Digital Insights. Available at: <https://datareportal.com/reports/digital-2025-sri-lanka> [Accessed 2 June 2025].

Khairy, M., Mahmoud, T.M. and Abd-El-Hafeez, T., 2021. Automatic Detection of Cyberbullying and Abusive Language in Arabic Content on Social Networks: A Survey. *Procedia Computer Science*, 189, pp.156–166. https://doi.org/10.1016/j.procs.2021.05.080.

Maduranga, H., 2024. *Cybercrime Analysis - Sri Lanka*. https://doi.org/10.13140/RG.2.2.18522.93123.

Menahil, A., Iqbal, W., Iftikhar, M., Shahid, W.B., Mansoor, K. and Rubab, S., 2021. Forensic Analysis of Social Networking Applications on an Android Smartphone. *Wireless Communications and Mobile Computing*, 2021(1), p.5567592. https://doi.org/10.1155/2021/5567592.

Mubarik, M.A., Wang, Z., Nam, Y., Kadry, S. and Azam Waqar, M., 2021. Instagram Mobile Application Digital Forensics. *Computer Systems Science and Engineering*, 37(2), pp.169–186. https://doi.org/10.32604/csse.2021.014472.

Muthuthanthri, M. and Smith, R.I., 2024. Hate Speech Detection for Transliterated English and Sinhala Code-Mixed Data. In: *2024 4th International Conference on Advanced Research in Computing (ICARC)*. [online] 2024 4th International Conference on Advanced Research in Computing (ICARC). Belihuloya, Sri Lanka: IEEE. pp.155–160. https://doi.org/10.1109/ICARC61713.2024.10499768.

Ogunleye, B. and Dharmaraj, B., 2023. The Use of a Large Language Model for Cyberbullying Detection. *Analytics*, 2(3), pp.694–707. https://doi.org/10.3390/analytics2030038.

Oxygen, 2025. *Oxygen Forensics*. [online] Oxygen Forensics. Available at: <https://www.oxygenforensics.com/en/> [Accessed 23 January 2025].

Ruwandika, N.D.T. and Weerasinghe, A.R., 2018. Identification of Hate Speech in Social Media. In: *2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer)*. [online] 2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer). Colombo, Sri Lanka: IEEE. pp.273–278. https://doi.org/10.1109/ICTER.2018.8615517.

Samarasinghe, S.W.A.M.D., Meegama, R.G.N. and Punchimudiyanse, M., 2020. Machine Learning Approach for the Detection of Hate Speech in Sinhala Unicode Text. In: *2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer)*. [online] 2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer). Colombo, Sri Lanka: IEEE. pp.65–70. https://doi.org/10.1109/ICTer51097.2020.9325493.

Sampath, K.K., 2023. Cyber Crimes and Cyber Laws in Sri Lanka: Safeguarding Digital Spaces. *Medium*. Available at: <https://kalhara-sampath.medium.com/cyber-crimes-and-cyber-laws-in-sri-lanka-safeguarding-digital-spaces-56e203627651> [Accessed 23 January 2025].

Senapati, C. and Roy, U., 2023. Bengali Hate Speech Detection Using Deep Learning Technique.

Shibly, F.H.A., Sharma, U. and Naleer, H.M.M., 2021. Detection of online hate speech in Sinhala text using machine and deep learning algorithms: A comparative study.

Weerasooriya, S., 2024. More than 2,000 cyberbullying cases reported in the country so far this year. Available at: <http://island.lk/more-than-2000-cyberbullying-cases-reported-in-the-country-so-far-this-year/> [Accessed 4 June 2025].

APPENDICES

A. Codes

i. Screenshots of the Code for Merging 3 Datasets

train_model.py | cli_tool.py | prepare_dataset.py

prepare_dataset.py > ...

```python
import pandas as pd

# === Singlish Dataset ===
singlish_df = pd.read_csv("data/singlish_dataset1.csv", encoding="utf-8")
singlish_df = singlish_df[['Sentence', 'Hate_speech']]
label_map = {"Offensive": 1, "Not offensive": 0}
singlish_df = singlish_df[singlish_df['Hate_speech'].isin(label_map)]
singlish_df['label'] = singlish_df['Hate_speech'].map(label_map)
singlish_df.rename(columns={'Sentence': 'text'}, inplace=True)
singlish_df = singlish_df[['text', 'label']]

# === Sinhala Unicode Dataset 1 ===
sinhala1_df = pd.read_csv("data/sinhalauni_dataset1.csv", encoding="utf-8")
sinhala1_df.rename(columns={'comment': 'text', 'label': 'label'}, inplace=True)
sinhala1_df['label'] = pd.to_numeric(sinhala1_df['label'], errors='coerce')
sinhala1_df.dropna(subset=['label'], inplace=True)
sinhala1_df['label'] = sinhala1_df['label'].astype(int)
sinhala1_df = sinhala1_df[['text', 'label']]

# === Sinhala Unicode Dataset 2 ===
sinhala2_df = pd.read_csv("data/sinhalauni_dataset2.csv", encoding="utf-8")
print("✅ sinhalauni_dataset2 columns:", sinhala2_df.columns)

# Rename and clean
sinhala2_df.rename(columns={'comment': 'text', 'label': 'label'}, inplace=True)
sinhala2_df['label'] = pd.to_numeric(sinhala2_df['label'], errors='coerce')
sinhala2_df.dropna(subset=['label'], inplace=True)
sinhala2_df['label'] = sinhala2_df['label'].astype(int)
sinhala2_df = sinhala2_df[['text', 'label']]

# === Merge All ===
combined_df = pd.concat([singlish_df, sinhala1_df, sinhala2_df], ignore_index=True)
combined_df.dropna(inplace=True)
combined_df.to_csv("merged_dataset.csv", index=False)

# Summary
print("Final dataset created: merged_dataset.csv")
print(f"Total rows: {len(combined_df)}")
print("Class balance:\n", combined_df['label'].value_counts())

```

ii. Screenshots of the Code for Training the Model

train_model.py | cli_tool.py | prepare_dataset.py

train_model.py > ...

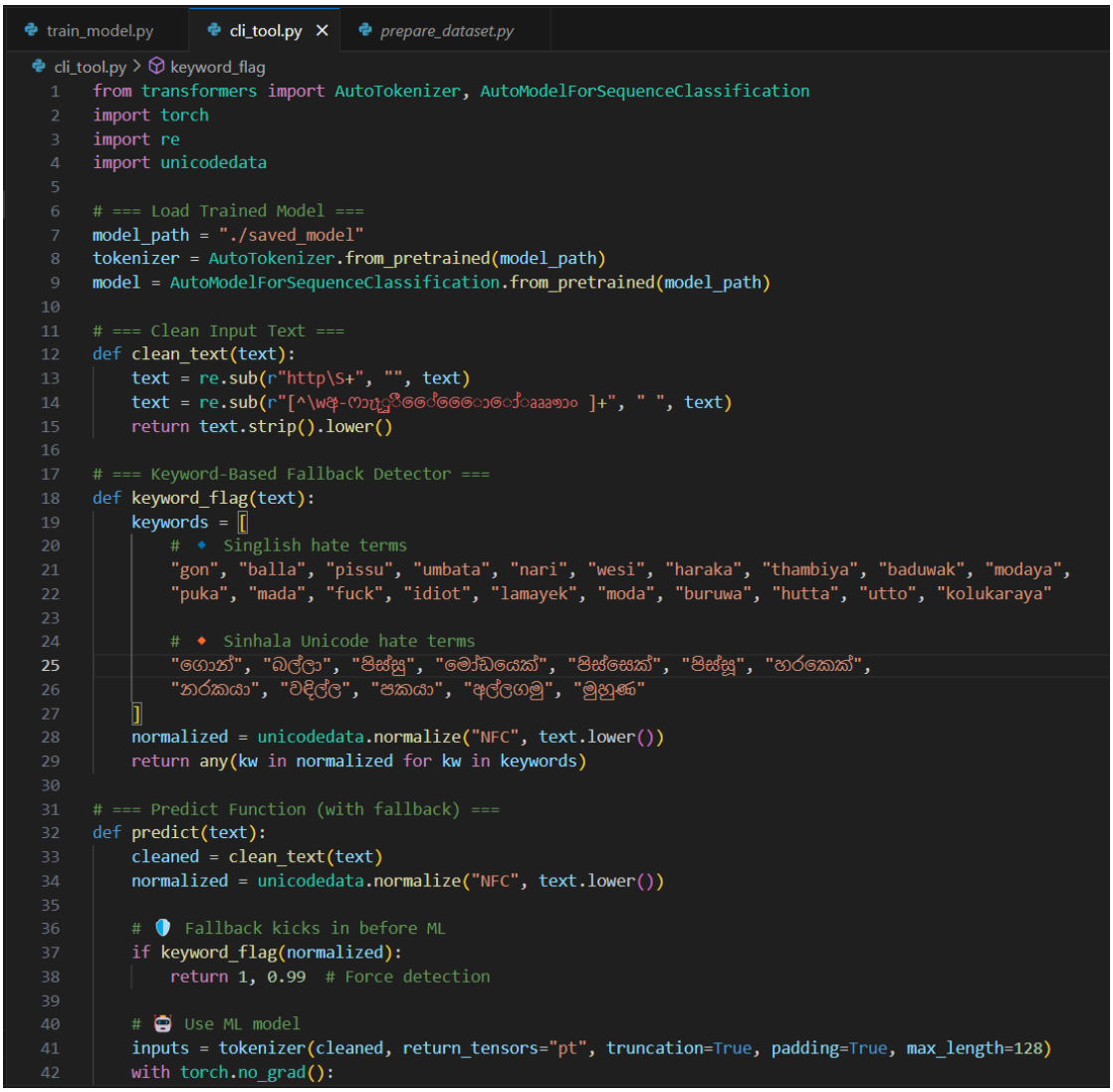
```python
from datasets import Dataset
from transformers import AutoTokenizer, AutoModelForSequenceClassification, Trainer, TrainingArguments
from sklearn.model_selection import train_test_split
import pandas as pd
import torch

# === Load merged dataset ===
df = pd.read_csv("merged_dataset.csv")
train_texts, val_texts, train_labels, val_labels = train_test_split(df["text"], df["label"], test_size=0.2, random_state=42)

# === Tokenization ===
model_name = "distilbert-base-multilingual-cased"
tokenizer = AutoTokenizer.from_pretrained(model_name)
train_encodings = tokenizer(train_texts.tolist(), truncation=True, padding=True)
val_encodings = tokenizer(val_texts.tolist(), truncation=True, padding=True)

# === Convert to Hugging Face Dataset format ===
train_dataset = Dataset.from_dict({**train_encodings, "label": train_labels.tolist()})
val_dataset = Dataset.from_dict({**val_encodings, "label": val_labels.tolist()})

# === Load pre-trained multilingual BERT model ===
model = AutoModelForSequenceClassification.from_pretrained(model_name, num_labels=2)

# === Training Configuration ===
training_args = TrainingArguments(
    output_dir="saved_model",
    per_device_train_batch_size=8,
    per_device_eval_batch_size=8,
    num_train_epochs=5,
    evaluation_strategy="epoch",
    save_strategy="epoch",
    logging_dir="./logs",
    logging_strategy="epoch",
    load_best_model_at_end=True,
    save_total_limit=1
)

# === Evaluation Metric ===
def compute_metrics(eval_pred):
    logits, labels = eval_pred
    preds = torch.argmax(torch.tensor(logits), axis=1)
    accuracy = (preds == torch.tensor(labels)).float().mean().item()
```

train_model.py | cli_tool.py | prepare_dataset.py

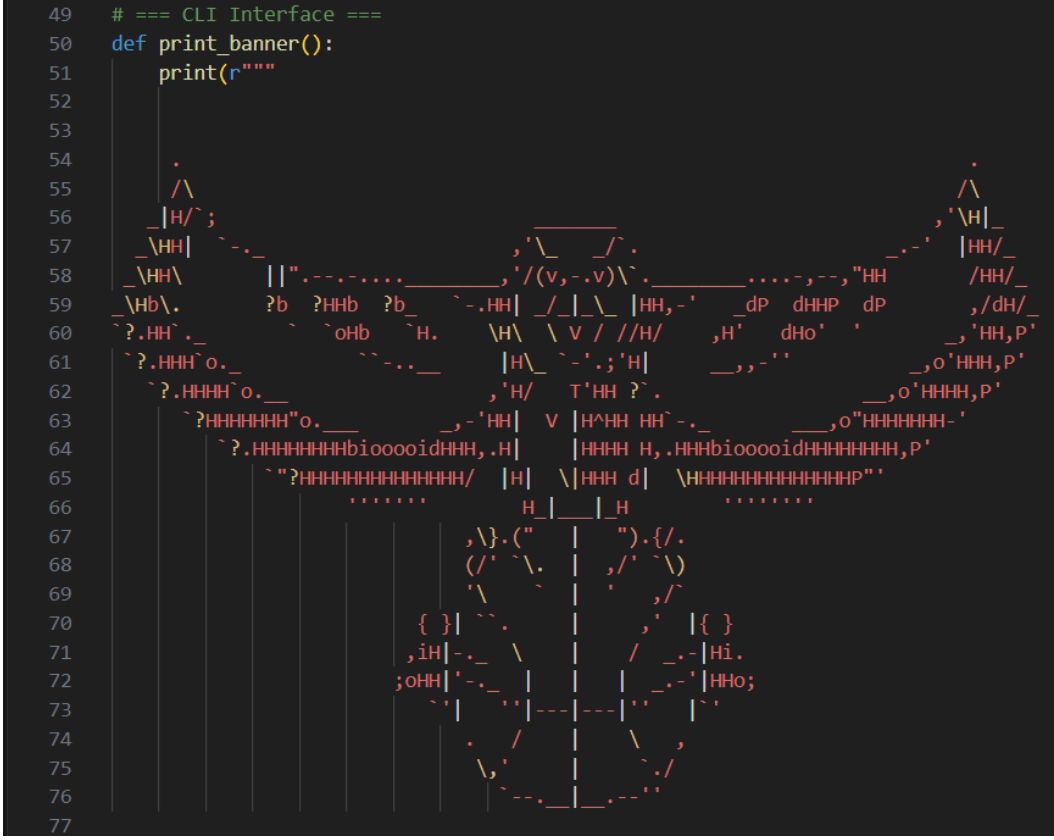
train_model.py > ...

```python
# === Evaluation Metric ===
def compute_metrics(eval_pred):
    logits, labels = eval_pred
    preds = torch.argmax(torch.tensor(logits), axis=1)
    accuracy = (preds == torch.tensor(labels)).float().mean().item()
    return {"accuracy": accuracy}

# === Trainer Setup ===
trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=train_dataset,
    eval_dataset=val_dataset,
    compute_metrics=compute_metrics
)

# === Train and Save ===
trainer.train()
trainer.save_model("saved_model")
tokenizer.save_pretrained("saved_model")
print("✅ Training complete. Model saved in 'saved_model/'")
```

iii. Screenshots of the Code for the Dummy Cli Tool



```
        outputs = model(**inputs)
        probs = torch.softmax(outputs.logits, dim=1)
        predicted_class = torch.argmax(probs).item()
        confidence = probs[0][predicted_class].item()
        return predicted_class, confidence

# === CLI Interface ===
def print_banner():
    print(r"""
```



```
88                                  88
88                                  88
88                                  88
88,dPPYba,  ,adPPYYba, 8b      db      d8 88   ,d8    8888888  8b    8b  8888888
88P'    "8a ""     `Y8 `8b    d88b    d8' 88 ,a8"     8b     8b  8b  8b  8b     8b
88       88 ,adPPPPP88  `8b  d8'`8b  d8'  8888[      8b    8888b  8b8b  8b    8888b
88       88 88,    ,88   `8bd8'  `8bd8'   88`"Yba,   8b           8b    8b
88       88 `"8bbdP"Y8     YP      YP     88   `Y8a  88888888b    8b     88888888b

    """)
    print("Type 'exit' to quit.\n")

def main():
    print_banner()
    while True:
        user_input = input("Enter a comment: ")
        if user_input.lower() == "exit":
            break

        label, confidence = predict(user_input)
        if label == 1:
            print(f"⚠️  Detected: It's Hate Speach (Confidence: {confidence:.2f})\n")
        else:
            print(f"✅ All clear! Not Hate Speach :) (Confidence: {confidence:.2f})\n")

if __name__ == "__main__":
    main()
```

B. Results

i. Screenshots of the Outcome

C. Screenshots of the Survey – Questionnaire

Survey on Cyberbullying Attempts, Language, and Usage Patterns on Social Media Platforms in Sri Lanka

This anonymous survey is part of an undergraduate research project focused on developing a multilingual digital forensic tool to detect cyberbullying on Facebook and Instagram.

The aim is to understand how social media is used in Sri Lanka, how cyberbullying occurs, and the types of language (Unicode Sinhala, Sinhala typed with English letters, or English) commonly used in such incidents.

Your responses will help to create a tool that supports local languages and assists law enforcement with online abuse investigations.

🔒 No personal information will be collected.
⏱ The survey will take less than 5 minutes to complete.

Thank you for your valuable input in making Sri Lanka's digital space safer.

SECTION 1: Demographic Information

Description (optional)

What age group do you fall under? *

- ○ Under 18
- ○ 18 – 24
- ○ 25 – 34
- ○ 35 – 44
- ○ 45+

Your gender? *

- ○ Female
- ○ Male
- ○ Prefer not to say
- ○ Other...

SECTION 2 : Social Media Usage

Description (optional)

Which social media platforms do you use regularly? *

- ☐ Facebook
- ☐ Instagram
- ☐ WhatsApp
- ☐ TikTok
- ☐ Snap Chat
- ☐ YouTube
- ☐ Twitter / X
- ☐ Other...

On which platforms have you experienced or seen cyberbullying (hate speech, hate comments, bullying, trolling, etc.)? *

- ☐ Facebook
- ☐ Instagram
- ☐ WhatsApp
- ☐ TikTok
- ☐ Snap Chat
- ☐ YouTube
- ☐ Twitter / X
- ☐ Other...

How often do you use social media? *

- ○ Daily
- ○ Weekly
- ○ Occasionally
- ○ Rarely

SECTION 3: Cyberbullying Experience & Awareness

Description (optional)

Have you ever personally witnessed or been affected by cyberbullying (hate speech, hate comments, bullying, trolling, etc.)? *

- ◯ Yes
- ◯ No
- ◯ Prefer not to say

If yes, did you report it? *

- ◯ Yes – to the platform (e.g., Facebook, Instagram)
- ◯ Yes – to the authorities (e.g., Police, SLCERT)
- ◯ I didn't know how or whom to report
- ◯ No

If you didn't report it, why not?

Short answer text

What types of harmful content did you see in cyberbullying incidents? *

- ☐ Insults or harsh comments
- ☐ Threats or blackmail
- ☐ Spreading false rumors
- ☐ Harassing or repeated messages
- ☐ Sharing personal or private photos
- ☐ Exclusion from groups
- ☐ Sexual or inappropriate content
- ☐ Other...

SECTION 4: Language Usage on Social Media

Description (optional)

In which language(s) do you mostly communicate on social media? *

- [ ] Sinhala (typed in Sinhala letters – Unicode)
- [ ] Sinhala (typed using English letters – e.g., "mokakda?")
- [ ] English
- [ ] Tamil
- [ ] Other...

Do you think cyberbullying often happens in Sinhala? *

- ( ) Yes – mostly in Sinhala Unicode
- ( ) Yes – mostly in Sinhala typed with English letters
- ( ) No – mostly in English
- ( ) Not sure

SECTION 5: Suggestions

Description (optional)

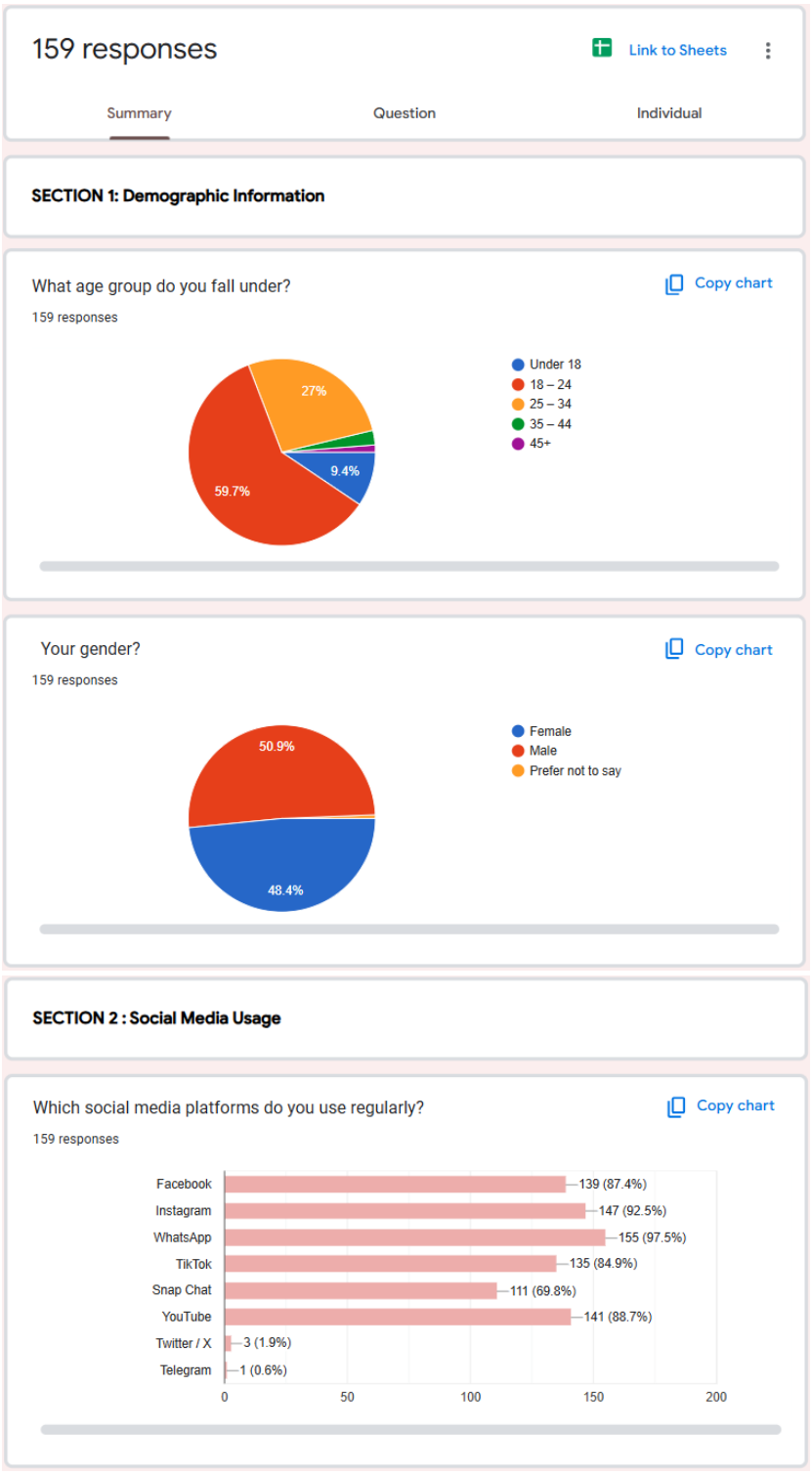
Do you think that Sri Lankan cybercrime units should focus more on solving and addressing these cybercrimes/cyberbullying attempts on social media? *

- ( ) Yes
- ( ) No

Do you have any suggestions for tools or actions that could help reduce cyberbullying in Sri Lanka?

Long answer text

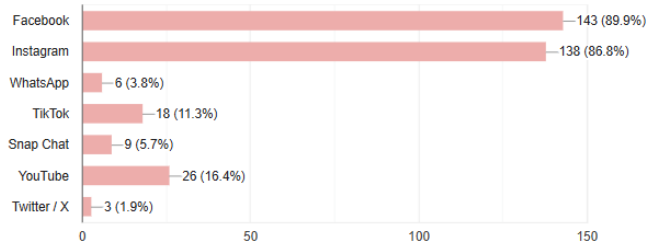
D. Screenshots of the Survey Results



On which platforms have you experienced or seen cyberbullying (hate speech, hate comments, bullying, trolling, etc.)?

Copy chart

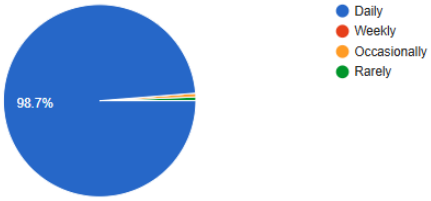
159 responses



How often do you use social media?

Copy chart

159 responses

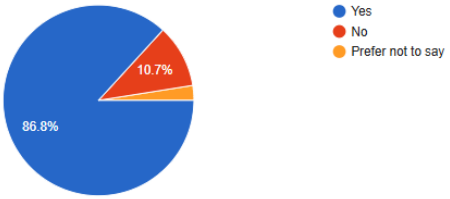


SECTION 3: Cyberbullying Experience & Awareness

Have you ever personally witnessed or been affected by cyberbullying (hate speech, hate comments, bullying, trolling, etc.)?

Copy chart

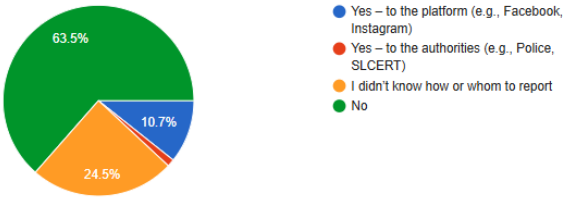
159 responses



If yes, did you report it?

Copy chart

159 responses



If you didn't report it, why not?

6 responses

I didn't knew what to do

I don't know

I reported it

Didn't think of it that much beacuse i was not bullied although i witnessed

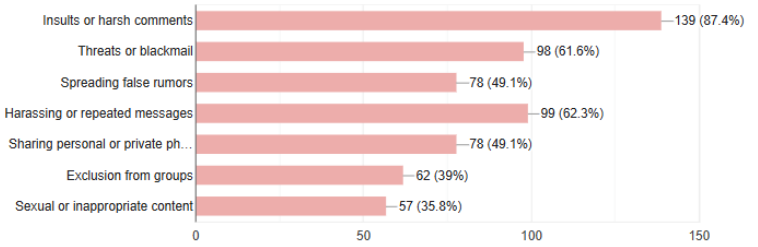
No point.

Didn't want to get in Unecessary Trouble

What types of harmful content did you see in cyberbullying incidents?

Copy chart

159 responses

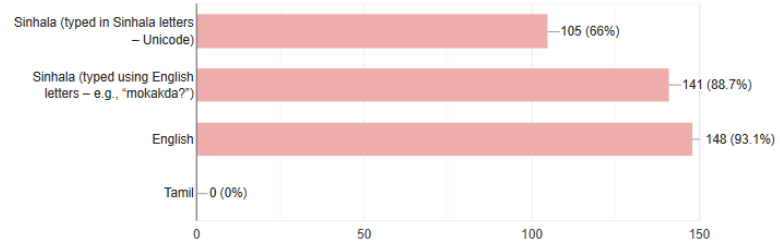


SECTION 4: Language Usage on Social Media

In which language(s) do you mostly communicate on social media?

Copy chart

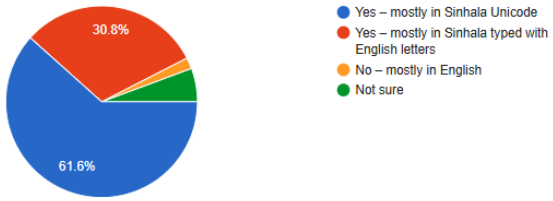
159 responses



Do you think cyberbullying often happens in Sinhala?

Copy chart

159 responses

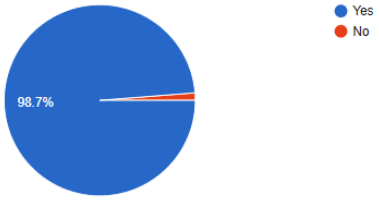


SECTION 5: Suggestions

Do you think that Sri Lankan cybercrime units should focus more on solving and addressing these cybercrimes/cyberbullying attempts on social media?

Copy chart

159 responses



Do you have any suggestions for tools or actions that could help reduce cyberbullying in Sri Lanka?

10 responses

Create a centralized national reporting platform for cyberbullying (like Report Harmful Content UK). Enhance capabilities of existing platforms like CERT|CC Sri Lanka (www.cert.gov.lk) to specifically handle cyberbullying complaints.

They can create a anonymous reporting systems

Make other ones aware of it...

Can we have a mechanism to identify which people do cyberbullying and have a personnel record for everyone in the country which done any bullying.

Now

Yes,

Implement nationwide digital citizenship education in schools, universities, and community centers, teaching youth about responsible online behavior.

Run multilingual public awareness campaigns (Sinhala, Tamil, and English) about the impact of