



Enhancing Sinhala Hate Speech Detection in Online Platforms

W. M. R. D. Silva
(Reg. No.: MS23006966)

A THESIS
SUBMITTED TO
SRI LANKA INSTITUTE OF INFORMATION TECHNOLOGY
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE IN INFORMATION TECHNOLOGY
(ENTERPRISE APPLICATIONS DEVELOPMENT)

December 2024

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Dr. Dilshan De. Silva

Approved for MSc. Research Project:

MSc. Programme Co-ordinator, SLIIT

Approved for MSc:

Head of Graduate Studies, FoC, SLIIT

DECLARATION

This is to certify that the work is entirely my own and not of any other person, unless explicitly acknowledged (including citation of published and unpublished sources). The work has not previously been submitted in any form to the Sri Lanka Institute of Information Technology or to any other institution for assessment for any other purpose.

Sign: 

W. M. R. D. Silva

Date: 15/01/2025

ABSTRACT

Enhancing Sinhala Hate Speech Detection in Online Platforms

W. M. R. D. Silva

MSc. in Information Technology (EAD)

Supervisor: Dr. Dilshan De Silva

December 2024

The rise of deep learning methodologies has indeed revolutionized text analysis, enabling more sophisticated and nuanced understanding of language dynamics. With the proliferation of social media platforms, these advancements have been particularly crucial in navigating the vast amounts of data generated by online interactions. However, amidst the benefits of this digital age, the prevalence of hate speech has emerged as a pressing concern, transcending linguistic and cultural boundaries. In the context of Sinhala, a language rich in nuances and deeply intertwined with cultural complexities, the challenges in detecting and mitigating hate speech are further compounded. Language is not merely a tool for communication but also a reflection of societal norms, values, and power structures. In the Sinhala-speaking context, historical legacies, religious beliefs, and political tensions intertwine to shape discourse in multifaceted ways. Consequently, any hate speech detection mechanism must navigate these intricate layers of meaning, accounting for cultural sensitivities and contextual nuances to ensure accurate identification of harmful content. The integration of deep learning techniques and advanced semantic analysis holds promise in enhancing hate speech detection in Sinhala. By leveraging the power of neural networks to discern patterns and contexts within textual data, such mechanisms can offer a more nuanced understanding of language dynamics. Moreover, the evaluation of these tools on real-world social media data not only validates their effectiveness but also provides insights into the evolving nature of online discourse. Ultimately, addressing hate speech in Sinhala and similar low-resource languages requires a multifaceted approach that combines technological innovation with cultural sensitivity and community engagement to foster safer and more inclusive online spaces.

ACKNOWLEDGEMENT

I would like to express my deepest gratitude to Dr. Dilshan De Silva and Mr. Lasitha Petthawadu for their invaluable support and guidance throughout the course of this research. Dr. De Silva's expert advice, insightful feedback, and encouragement were instrumental in shaping the direction and success of this study. His deep understanding of the subject matter and willingness to share his knowledge greatly enhanced the quality of the research.

Without their dedicated support and mentorship, this research would not have been possible. Their contributions have been invaluable, and I am deeply appreciative of their efforts to help me achieve my research goals.

TABLE OF CONTENTS

DECLARATION	ii
ABSTRACT	iii
ACKNOWLEDGEMENT	iv
TABLE OF CONTENTS	v
List of Figures	viii
List of Tables	ix
Chapter 1 Introduction	1
1.1 Background	1
1.2 Problem Statement	2
1.3 Objectives of the Research	3
1.3.1 Investigate Deep Learning Architectures for Hate Speech Detection in Sinhala	3
1.3.2 Integrate Advanced Semantic Analysis Techniques into the Detection Framework	3
1.3.3 Evaluate the Proposed Mechanism on Real-World Social Media Data	4
1.3.4 Compare the Proposed Mechanism with Existing Approaches	4
1.4 Research Questions	4
1.4.1 How to develop a mechanism for hate speech detection in Sinhala that integrates deep learning techniques with advanced semantic analysis to achieve more consistent and reliable results?	5
1.5 Significance of the Study	6
1.6 Scope and Limitations	7
Chapter 2 Literature Review	9
2.1 Related Work	9
2.1.1 A Deep Learning Ensemble Hate Speech Detection Approach for Sinhala Tweets [2]	9
2.1.2 Hate Speech Detection in Sinhala-English Code-Mixed Language [6]	10
2.1.3 Identifying False Content and Hate Speech in Sinhala YouTube Videos by Analyzing the Audio [7]	11
2.1.4 Machine Learning Approach for the Detection of Hate Speech in Sinhala Unicode Text [8]	12
2.1.5 Sinhala Hate Speech Detection in social media Using Machine Learning and Deep Learning [9]	13
2.1.6 Fine-tuning XLM-R for the Detection of Sinhala Hate Speech Content on Twitter and YouTube [10]	15
2.1.7 Identifying Racist Social Media Comments in Sinhala Language Using Text Analytics Models with Machine Learning [11]	16
2.2 Research Gap	17
Chapter 3 Methodology	19
3.1 Research Design	19
3.2 Data Collection	20
3.2.1 Sources	20

3.2.2 Data Description	20
3.2.3 Ethical Considerations	21
3.3 Data Preprocessing.....	22
3.3.1 Text Cleaning and Label Encoding.....	22
3.3.2 Data Splitting	23
3.4 LaBSE Model.....	23
3.4.1 LaBSE Model Architecture.....	23
3.4.2 BertTokenizer	26
3.4.3 Classifier Architecture	27
3.4.4 Classifier Architecture Diagram	30
3.5 Model Training and Evaluation	30
3.5.1 Training Process.....	30
3.6 REST API and Google Chrome Extension	31
3.7 Real-World Applications	33
3.7.1 Social Media Platforms.....	33
3.7.2 Online Communities and Forums	34
3.7.3 Government and Regulatory Bodies	34
3.7.4 Educational Platforms.....	34
3.7.5 Corporate and Brand Protection.....	35
3.7.6 Crisis Management and Incident Response	35
3.7.7 Content Filtering and Recommendation Systems	35
Chapter 4 Results	36
4.1 Model Performance.....	36
4.1.1 Accuracy and Loss Curves.....	36
4.1.2 Comparative Analysis	36
4.2 Case Studies	37
4.2.1 Sample Tweets Analysis	39
4.2.2 Difficult Cases	40
Chapter 5 Discussion	43
5.1 Interpretation of Results.....	43
5.2 Challenges and Considerations in Model Deployment.....	45
5.3 Challenges in Hate Speech Detection	46
5.4 Comparison with Previous Work.....	49
5.5 Implications for Online Platforms.....	49
5.6 Legal Implications	50
5.7 Impact on Free Speech and Moderation Policies.....	51
5.8 Impact on Social Media Policy and Regulation.....	52

5.9 Bias in Hate Speech Detection.....	53
5.9.1 Bias in Data Collection	53
5.9.2 Annotation Bias	54
5.9.3 Bias in Model Prediction.....	54
5.9.4 Mitigating Bias.....	55
5.9.5 Responsible Practices and Implications	55
5.10 Future Work	56
References.....	58
Appendix.....	60
Appendix 1: Web Extension	60
Appendix 2: Sample Request and Response JSON	61
Appendix 3: Tweets collected from X and evaluated by the classifier model.....	62

List of Figures

Figure 3.1 Classifier Architecture Diagram	30
Figure 3.2 Flow of the main components	33
Figure 3.3 Work flow of the proposed study	33

List of Tables

Table 4.1 Evaluated Tweets summary	38
Table 5.1 Comparison with previous studies	44