

# **A DEEP LEARNING ENSEMBLE HATE SPEECH DETECTION APPROACH FOR SINHALA TWEETS**

Munasinghe Imiyage Sidath Asiri Munasinghe

(209358D)

Master of Science in Data Science

Department of Computer Science and Engineering  
Faculty of Engineering

University of Moratuwa  
Sri Lanka

March 2022

# **A DEEP LEARNING ENSEMBLE HATE SPEECH DETECTION APPROACH FOR SINHALA TWEETS**

Munasinghe Imiyage Sidath Asiri Munasinghe

(209358D)

Thesis/Dissertation submitted in partial fulfillment of the requirements for the degree  
Master of Science in Data Science

Department of Computer Science and Engineering  
Faculty of Engineering

University of Moratuwa  
Sri Lanka

March 2022

## DECLARATION

I declare that this is my work and this dissertation does not incorporate without acknowledgment any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief, it does not contain any material previously published or written by another person except where the acknowledgment is made in the text.

Also, I hereby grant to the University of Moratuwa the non-exclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic, or another medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature:

Date:

The above candidate has carried out research for the Master's dissertation under my supervision.

Signature of the supervisor:

Date:

## ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to all those who provided support to make my research on “A DEEP LEARNING ENSEMBLE HATE SPEECH DETECTION APPROACH FOR SINHALA TWEETS” successful.

First of all, I would like to express my gratitude towards my project supervisor Dr. Uthayashanker Thayasivam, Senior Lecturer, Department of Computer Science and Engineering. I am highly indebted to him for his guidance and constant supervision as well as for providing necessary information regarding the project and for his support in completing the project successfully.

I am sincerely thankful to the final year project coordinator Dr. Charith Chittaranjan, Senior Lecturer, Department of Computer Science and Engineering for the support given throughout the project time period. Further, I would like to extend my gratitude to Dr. Sapumal Ahangama for participating in evaluations and providing me very useful guidance to make research successful.

Especially I would like to thank Prof. Indika Perera, Head of Department, Department of Computer Science and Engineering for his assistance and coordination to conduct the research without any issues during the final year.

Finally, I wish to thank the academic and non-academic staff of Department of Computer Science and Engineering and everyone who supported me.

## Abstract

We live in an era where social media platforms play a key role in the society. With the advancement of technology, these platforms have become more closer to people and currently, they can interact with most of the native languages including the Sinhala language. This has enabled people to express their opinions more conveniently. At the same time, it is very common to observe that people express very hateful offensive opinions on social media platforms and in certain applications it is mandatory to block this kind of content.

Several studies have been carried out on this area for the Sinhala language with traditional machine learning models and as per the results, none of them have shown promising results. Further, current approaches are far behind the latest techniques carried out in high-resource languages like English. Hence this study presents a deep learning-based approach for hate speech detection which has shown outstanding results for other languages. Three deep learning models namely LSTM, CNN and BiGRU which have proven performance in Natural Language Processing domain have been considered here. Moreover, a deep learning ensemble was constructed from these three models to evaluate whether the ensemble technique can further improve the model performance. These models were trained and tested on a newly created dataset using the Twitter API. Moreover, the model generalizability was further tested by applying it to a completely new dataset.

As per the results, it can be clearly observed that the deep learning-based approach has outperformed the traditional machine learning models. Moreover, further tests on the model generalizability reveal that this approach is more generalized and produces better predictions than the prior approaches.

Finally, this study experiments with using extra features in addition to the Tweet content such as retweet count, favourited count, etc, to evaluate whether those can be utilized to improve the performance further. As per the results obtained in this study, it can be observed that there is an impact on the performance using extra features. It is recommended to experiment further on this area in future studies.

# TABLE OF CONTENTS

DECLARATION .....	i
ACKNOWLEDGEMENTS .....	ii
Abstract .....	iii
LIST OF FIGURES .....	vi
LIST OF TABLES .....	vii
1 INTRODUCTION .....	1
1.1 Hate Speech Detection Approaches .....	1
1.2 Challenges in Hate Speech Detection.....	2
1.3 Research Objectives .....	2
1.4 Contributions of Research .....	2
2 LITERATURE REVIEW.....	4
2.1 Approaches in Detecting Abusive Text.....	5
2.2 Studies Carried Out for the Sinhala Language .....	7
2.3 Studies carried out for the English language.....	12
2.4 Studies Carried Out for Other Languages .....	23
3 METHODOLOGY.....	31
3.1 Data Collection.....	32
3.2 Dataset Description .....	33
3.3 Data Pre-processing and Preparation .....	34
3.3.1 Tokenization.....	34
3.3.2 Removal of stop words .....	34
3.3.3 Stemming .....	35
3.3.4 Data Shuffling .....	35
3.4 Feature Engineering .....	35
3.4.1 Emoji Count .....	36
3.4.2 Tweet Length .....	36
3.5 Exploratory Data Analysis .....	37
3.5.1 Reply Count .....	37
3.5.2 Retweet Count.....	37
3.5.3 Possibly Sensitive Editable .....	38
3.5.4 Is Quote Status .....	39
3.5.5 Favourite Count.....	39
3.6 Model Construction.....	40
3.6.1 Convolution Neural Network (CNN).....	40
3.6.2 Long Short-Term Memory (LSTM).....	41

3.6.3	Bidirectional Gated Recurrent Unit (BiGRU).....	43
3.6.4	Ensemble of Deep Learning Models.....	44
4	EVALUATION.....	45
4.1	K-Fold Cross-Validation .....	45
4.2	Accuracy.....	45
4.3	Precision .....	45
4.4	Recall.....	46
4.5	F-Score .....	46
4.6	Receiver Operating Characteristic (ROC).....	47
4.7	Area Under the Curve (AUC).....	47
5	RESULTS .....	49
5.1	Performance by Deep Learning Models .....	49
5.2	Performance by Traditional Machine Learning Models .....	49
5.3	Performance on the Separate Dataset .....	50
5.4	Performance with Extra Features .....	51
6	DISCUSSION .....	53
7	CONCLUSION.....	54
8	REFERENCES.....	55

## LIST OF FIGURES

Figure 1: The flow of how the research area for Sinhala language evolved from creating the NLP related tools to building machine learning classification models ....	7
Figure 2: The flow of how the research area for English language evolved from lexicon-based approaches to deep learning approaches.....	12
Figure 3: The overall data collection process by fetching Tweets via Tweet API ....	31
Figure 4: The ensemble model construction process with the majority vote.....	32
Figure 5: Emoji count vs class label revealing a higher emoji count in non-offensive Tweets .....	36
Figure 6: Tweet length showing that the offensive Tweets have a shorter review length.....	36
Figure 7: Reply count depicting that non-offensive Tweets have considerably large reply counts .....	37
Figure 8: Retweet count showing a higher count for non-offensive Tweets .....	38
Figure 9: possibly_sensitive_editable revealing that a large proportion of offensive Tweets when the Tweet doesn't contain any links.....	38
Figure 10: is_quote_status depicting that it doesn't show a significant relation to the class label .....	39
Figure 11: favourite_count showing that the majority of non-offensive Tweets have a higher favourite count .....	39
Figure 12: Architecture of a CNN having embedding, convolution, pooling, flatten and output layers .....	40
Figure 13: Structure of an LSTM unit having input, input modulation, output, and forget gates .....	42
Figure 14: Structure of a set of LSTM units connected in a serial manner to handle sequential data.....	42
Figure 15: General structure of a RNN having an update gate and a reset gate .....	43
Figure 16: ROC Curve showing the behaviour of curves for different scenarios.....	47
Figure 17: AUC showing the TP rate vs FP rate quantifying the performance of the model by the area. ....	48
Figure 18: Classification report for Logistic Regression model by train-test split....	50
Figure 19: Classification report per fold in Ensemble of deep learning models.....	51
Figure 20: Performance vs feature set with showing highest performance with set 2	52



## LIST OF TABLES

Table 1: Dataset Description.....	33
Table 2: Performance metrics for deep learning models.....	49
Table 3: Performance metrics for traditional machine learning models .....	50
Table 4: Performance metrics for deep learning models on a new dataset .....	50