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 a 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

LIST OF FIGURES

LIST OF TABLES