AN OPTIMIZED MACHINE LEARNING PIPELINE FOR DETECTING RACIST COMMENTS WRITTEN IN SINHALA LANGUAGE

Chandrajith Priyadarshana Bandara Senadheera

189347N

Dissertation submitted in partial fulfillment of the requirements for the degree Master of Science in Computer Science

Department of Computer Science and Engineering

University of Moratuwa

Sri Lanka

February 2021

DECLARATION

I declare that this is my own work and this thesis does not incorporate without acknowledgement any material previously submitted for 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 acknowledgement is made in the text.

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

Signature: …… Date: …….

Name: C.P.B. Senadheera

The supervisor should certify the thesis with the following declaration.

The above candidate has carried out research for the Masters thesis under my supervision.

Name: Dr. Uthyasanker Thayasivam.

Signature of the supervisor: ……. Date: …….

ACKNOWLEDGEMENT

I warmly convey my gratitude to my supervisor Dr. Uthyasanker Thayasivam, for his continuous guidance and support. His patience and knowledge inspired me and helped me to improve this study. He always steered me in the right direction whenever he thought I needed it and motivated me to complete my research.

Additionally, I appreciate the support provided by all the other university lecturers of the Department of Computer Science and Engineering, who supported me from the beginning and during the Masters degree.

Finally, I express my gratefulness to my family members for supporting me and believing in me. They kept me going on and this work would not have been possible without their support. Finally I would like to thank my closest friends who helped me in many ways and encouraged me to complete this research successfully.

ABSTRACT

Text classification is one of the core area of machine learning applications and it continuously improving with time. Hate speech classification in e-content became very popular over last few years and search engines, social media were the main interested parties of this area. “Racism content” is a sub part of the “Hate speech” area and many studies were carried out in this area. But surprisingly “Racism text classification in Sinhala language” is not very much popular. This study focus on building a machine learning pipeline for detecting racism comments written in Sinhala language. Sinhala racism text classification techniques can be used to remove unnecessary texts from social media, pages, blogs and many more text sources. This study was done to provide a solution to Sinhala racist text classification problems in social media, but also this study is valid for any text source that contain Sinhala text as Unicode text.

As the initial step, previous similar studies were reviewed and documented. Used techniques and results of similar studies were documented and reviewed. One similar study was selected as the baseline and set the baseline performance measures as the lower margin of the performance target. An architecture for the pipeline was designed and a methodology was selected as the next step. In this methodology, as the initial step, dataset was extracted and preprocessed. Stemming, stop word removal and conversion of text to basic character word were the main preprocessing steps. Features were extracted next and due to less number of racism data, an oversampling method was used to increase the training data. Six machine learning classifiers were selected and those were Random forest, Naïve bayes, SVM, Logistic regression, Ada boost and XGBoost. All the classifiers were trained with oversampled data. This was a major improvement point of the results. In order to increase the performance of these classifiers further, hyperparameter tuning was performed. Ensemble techniques were also used to increase the performance. As the ensemble techniques bagging, boosting and voting was used. After selecting the best classifiers from bagging and boosting, best three classifiers were set as the input to voting classifier to get the final results.

The study shows that each preprocessing steps improves the performance of classifiers and results. Each classifier behave differently in each step. This study highlights these differences and sensitivity of each classifier to various changes. In oversampling step and hyperparameter tuning step, all the classifiers reached to a stable level. Hyperparameter tuning identified as a critical step in Sinhala text classification. Finally the best results were shown by the voting classifier and selected as the best model in the study. The proposed pipeline performed better and highlighted the each classifiers performance differences. The pipeline was able to outperform the baseline with a greater accuracy. The study proves that the racism text classification is possible with selected classifiers and accurate data. Study also proves that a better Sinhala racism text classification pipeline can be built using other classifiers with the help of ensemble techniques and enhanced preprocessing techniques.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Description
BoW	Bag of Words
CNN	Convolutional Neural Network
IDF	Inverse Document Frequency
PoS	Part of Speech
ROC	Receiver Operating Characteristic
SATC	Semi-Automated Text Classifier
SMOTE	Synthetic Minority Oversampling Technique
TF	Term Frequency
UTF	Unicode Transformation Format
WWW	World Wide Web

LIST OF APPENDICES