

Identify Hateful Comments in Sinhala Language on Social Media

W.W.E.N. Fernando
189461H

Faculty of Information Technology
University of Moratuwa

July 2021

Identify Hateful Comments in Sinhala Language on Social Media

W.W.E.N. Fernando
189461H

Dissertation submitted to the Faculty of Information Technology, University of Moratuwa,
Sri Lanka for the partial fulfillment of Degree of Master of Science in Information
Technology.

July 2021

Declaration

We declare that this thesis is our own work and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education. Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.

Name of Student

W.W.E.N. Fernando

Signature of Student

Date

Supervised by

Name of Supervisor

Mr. S. C. Premaratne

Signature of Supervisor

Date

Acknowledgements

I am so grateful to my supervisor Mr. Saminda Premaratne, Senior Lecturer, Faculty of Information Technology, University of Moratuwa for academic guidance, advice, time, encouragement and patience that has seen me through, making this work success.

Also I would like to give a special thanks to the Prof. Rathnasiri Arangala, Senior Professor in Sinhala, University of Sri Jayewardenepura for his guidance, advice, time and encouragement in carrying out this research. As well as special thanks to Mr. K.K. Premarathne, Former Principle, Bandaragama National College for his precious support as an annotator and his encouragement.

Furthermore, my thanks also go out to my beloved family members for their support, love, advice and encouragement throughout the period of my study. I lastly wish to extend my sincere gratitude toward my colleagues who have tirelessly guided and encouraged me at all times through my course.

Abstract

In present, the spread of hate speech through social media has become a very serious problem, both globally and locally. The route cause for this is the increasing use of social media with the rapid expansion of computer science and information technology. Therefore, it is very important to use same to control this kind of situations. Although there is a mechanism in place on social media to automatically control such hate speech in English language, but it is still not seen in Sinhala Language. The reason for this is the lack of knowledge about the native languages such as Sinhala in the social media service providers. Therefore, the identification of hateful contents in Sinhala language is an urgent and vital task that needs to be addressed.

This research propose lexicon based and machine learning based approaches for the automatic identification of hateful speech in Sinhala on social media. With different pre-processing techniques and machine learning algorithms, machine learning algorithm based approach was conducted with four different approaches. These approaches were begun with 3000 comments which is equally divided into hateful and non-hateful. Using these comments, it was able to identify the most appropriate featured groups and model to identify the hateful speech in Sinhala language on social media.

Table of Contents

Declaration.....	i
Acknowledgements.....	ii
Abstract.....	iii
Table of Contents.....	iv
List of Figures.....	vi
List of Tables.....	vii
Chapter 1 - Introduction.....	1
1.1 Prolegomena.....	1
1.2 Background and Motivation.....	2
1.3 Problem Statement.....	3
1.4 Aim and Objectives.....	3
1.4.1 Aim.....	3
1.4.2 Objectives.....	3
1.5 Proposed Solution.....	4
1.6 Structure of the theses.....	4
Chapter 2 - Literature Review.....	5
2.1 Introduction.....	5
2.2 Related work for Social Media Text Mining for Identify Objectionable Content.....	6
2.3 Objectionable Content Identification.....	16
2.4 Tools Available for Sinhala Language.....	16
2.5 Summary.....	17
Chapter 3 - Adopted Technologies.....	18
3.1 Introduction.....	18
3.2 Text Mining Techniques.....	18
3.3 Multinomial Naïve Bayes.....	18
3.4 Support Vector Machine.....	19
3.5 Rapid Miner Studio.....	19
3.5.1 Text Processing.....	20
3.5.2 Weka.....	20
3.5.3 Operator Toolbox.....	20
3.6 Summary.....	21
Chapter 4 - Research Methodology.....	22
4.1 Introduction.....	22
4.2 Hypothesis.....	22

4.3	Input	22
4.4	Output	23
4.5	Process	23
4.6	Summary	23
Chapter 5 - Analysis and Design		24
5.1	Introduction.....	24
5.2	High level Architecture of System.....	24
5.3	Summary	25
Chapter 6 - Implementation		26
6.1	Introduction.....	26
6.2	Data Corpus Construction.....	26
6.3	Text Pre-processing	27
6.4	Construct Negative Word List	28
6.5	Dictionary based Classification	29
6.6	Feature Extraction.....	30
6.7	Feature Vectorization.....	33
6.8	Machine Learning based Classification	34
6.9	Performance Measurements.....	34
6.10	Summary	36
Chapter 7 - Evaluation		37
7.1	Introduction.....	37
7.2	Evaluation of Classification Techniques.....	37
7.3	Summary	40
Chapter 8 - Conclusion and Future Work		41
8.1	Introduction.....	41
8.2	Conclusion	41
8.3	Limitations	42
8.4	Future Developments.....	42
8.5	Summary	42
References.....		43

List of Figures

Figure 5.1: High Level Architecture of the Design	24
Figure 6.1: Procedure to obtain a detailed list of words from training data set	28
Figure 6.2: Word List of Training Data Set	29
Figure 6.3: Negative word List	29
Figure 6.4: Dictionary based Classification Model	30
Figure 6.5: Example for Word N-gram.....	31
Figure 6.6: Example for Character N-gram	32
Figure 6.7: Machine Learning based Classification Model	34
Figure 6.8: Performance Measurements	35

List of Tables

Table 2.1: Summary of Feature Extraction Techniques	9
Table 2.2: Summary of Feature Vectorization Techniques	10
Table 2.3: Summary of Machine Learning Techniques used in Hate Speech Detection.....	14
Table 2.4: Performance Evaluation Summary of surveyed Machine Learning Techniques....	15
Table 6.1: Annotated Data Corpus.....	26
Table 6.2: Word N-gram Feature Groups	31
Table 6.3: Character N-gram Feature Groups.....	32
Table 7.1: Performed Experiments	37
Table 7.2: Results of Experiment A - Dictionary Based Hate Speech Detection	38
Table 7.3: Results of Experiment B1.1 - with MNB Classification Technique.....	38
Table 7.4: Results of Experiment B1.2 - with SVM Classification Technique	38
Table 7.5: Results of Experiment B2.1 - with MNB Classification Technique.....	38
Table 7.6: Results of Experiment B2.2 - with SVM Classification Technique	39
Table 7.7: Details of methods used and the results of best fit model.....	39