

SENTIMENT ANALYSIS OF SINHALA TWEETS

Warna Ieshaka Karunaratne

(189328H)

Degree of Master of Science

Department of Computer Science and Engineering

University of Moratuwa

Sri Lanka

July 2020

SENTIMENT ANALYSIS OF SINHALA TWEETS

Warna Ieshaka Karunaratne

(189328H)

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

Department of Computer Science and Engineering

University of Moratuwa

Sri Lanka

July 2020

DECLARATION

I declare that this is my own work and this dissertation 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 dissertation, 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: W.I.Karunaratne

The supervisor/s should certify the thesis/dissertation with the following declaration.

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

Signature of the supervisor:

Date:

Name: Dr. Uthayasanker Thayasivam

ACKNOWLEDGEMENT

I would like to convey my genuine appreciation to my supervisor Dr. Uthyaanker Thayasivam for his determined efforts, continuous supervision and assistance. I have been extremely fortunate to have him as my supervisor, whose guidance and enthusiasm helped me to improve. His motivation, persuasion and patience helped me to overcome many crisis situations and complete this research successfully.

My heartfelt appreciation is rendered to all my friends for their continuous assistance and encouragement given to me during this challenging and hectic endeavour.

Most significantly, this wouldn't have been possible without the love and support of my parents. They have been a continuous source of love, concern, support and strength all these years. I am grateful to my parents for believing in me, cheering for me and being patient and supportive during this critical period in my academic life, which motivated me to successfully complete the research.

ABSTRACT

Sentiment analysis has become a popular topic since the last decade. The increase in the use of internet has led to the increase of user-generated content. This has played an important role in making sentiment analysis more popular among researchers. The user-generated content can provide some valuable insight about the public opinion to the government and various industries.

This research has mainly focused on sentiment analysis of Sinhala language. Sinhala is the most spoken language in Sri Lanka. With the increased use of the internet and social media, there is a considerable amount of information communicated via Sinhala. This has presented a good opportunity to mine the information presented in Sinhala language. Performing Sinhala language sentiment analysis has some difficulties, as Sinhala is morphologically rich and is a language of free order compared to English. Lack of Sinhala language resources has brought challenges from gathering and generating data sets to stemming / lemmatizing algorithms. This research has tried to address the above challenges by developing a Sinhala dataset suitable for sentiment analysis and by developing a stemming algorithm for Sinhala. The dataset is developed by collecting Tweets from Twitter and it has been manually annotated.

In addition to the resource creation, sentiment analysis of Sinhala language is also performed using word embedding as features. Several sentiment analysis experiments are performed by using several machine learning techniques. The accuracy as well as precision and recall are used to identify the best performing model. The problems faced when conducting sentiment analysis for Sinhala language are discussed in the research. The research has discussed the difference between the user-generated content in English and Sinhala.

Table of Content

DECLARATION	i
ACKNOWLEDGEMENT	ii
ABSTRACT.....	iii
Table of Content	iv
Table of tables.....	vii
Table of figures	vii
Chapter 1 Introduction.....	1
1.1 User-Generated Content.....	2
1.2 Web Content in Sinhala Language.....	3
1.3 Sentiment Analysis for Sinhala Language	4
1.4 General Approaches to Sentiment Analysis.....	4
1.5 Main Challenges in Sentiment Analysis	5
1.5.1 Challenges in Sentiment Analysis overall	5
1.5.2 Challenges in Sentiment Analysis for Sinhala.....	6
1.5.3 Challenges in collecting Sinhala Tweets	6
1.6 Motivation	6
1.7 Applications	7
1.8 Objectives of the Research.....	7
1.9 Contribution of the Research.....	7
Chapter 2 Literature Review.....	9
2.1 History and Growth of Sentiment Analysis	10
2.2 General Methods Used in Sentiment Analysis.....	11

2.2.1 Sentiment analysis using subjective lexicon	11
2.2.2 Sentiment analysis using n-gram modelling.....	13
2.2.3 Sentiment analysis using machine learning techniques.....	15
2.3 Sentiment Analysis for Other Languages.....	16
2.4 Sinhala Language Sentiment Analysis	19
2.5 Summary	21
Chapter 3 Methodology	27
3.1 Sinhala Dataset	28
3.2 Sinhala Lexicon.....	28
3.3 Stemming	29
3.3.1 Stemming Method	29
3.4 Sinhala Word Embedding	30
3.5 Sentiment Analysis.....	30
Chapter 4 Implementation	31
4.1 Sinhala Dataset	32
4.2 Sinhala Sentiment Lexicons	32
4.3 Stemming Algorithm.....	33
4.4 Sinhala Word Embedding	35
4.5 Sinhala Sentiment Analysis.....	35
4.5.1 Two-way Sinhala Sentiment Analysis (Experiment 1)	35
4.5.2 Three-way Sinhala Sentiment Analysis (Experiment 2)	36
4.6 Problems Faced in Implementation.....	36
4.6.1 Problems Faced When Collecting Sinhala Dataset	36
4.6.2 Problems Faced When Creating Sinhala Sentiment Lexicons	37

Chapter 5	Results.....	38
5.1	Two-way Sentiment Analysis (Experiment 1).....	39
5.1.1	Two-way Sentiment Analysis using Naïve Bayes.....	39
5.1.2	Two-way Sentiment Analysis using SVM (Linear).....	40
5.1.3	Two-way Sentiment Analysis using SVM (rbf).....	40
5.1.4	Two-way Sentiment Analysis using LightGBM.....	41
5.1.5	Two-way Sentiment Analysis using XgBoost.....	42
5.1.6	Two-way Sentiment Analysis using AdaBoost.....	43
5.2	Three-way Sentiment Analysis (Experiment 2).....	43
5.2.1	Three-way Sentiment Analysis using LightGBM.....	44
5.2.2	Three-way Sentiment Analysis using XgBoost.....	44
5.2.3	Three-way Sentiment Analysis using AdaBoost.....	45
5.4	Summary.....	45
Chapter 6	Discussion, Conclusion and Possible Future Work.....	47
6.1	Discussion.....	48
6.2	Conclusions.....	49
6.3	Possible Future Work.....	50
References	51

Table of tables

Table 2-1 Review of two Hindi Sentiment analysis papers	18
Table 2-2 Review of two Russian Sentiment analysis papers	19
Table 2-3 Summary of Sentiment Analysis papers.....	21
Table 5-1 Model performance summary table	45

Table of figures

Figure 3.1: Flow chart of the stemming algorithm	30
Figure 4.1 Annotated Sinhala Tweets	32
Figure 4.2 A part of “Ingiya” Dictionary	33
Figure 4.3 Sinhala lexicons with word combinations	33
Figure 4.4 Sinhala word and its stem.....	34
Figure 4.5 100 frequent words in the dataset	34
Figure 4.6 Errors in the stemming output	35
Figure 4.7 Part of "Ingiya" translator.....	37
Figure 5.1 Naive Bayes Model Performance in experiment 1	39
Figure 5.2 AUC Curve for Naive Bayes	39
Figure 5.3 SVM (linear) Model Performance in experiment 1	40
Figure 5.4 AUC Curve for SVM-Linear	40
Figure 5.5 SVM (rbf) Model Performance in experiment 1	40
Figure 5.6 AUC Curve for SVM-rbf.....	41
Figure 5.7 LightGBM Model Performance in experiment 1	41
Figure 5.8 AUC Curve for LightGBM	41
Figure 5.9 XgBoost Model Performance in experiment 1	42
Figure 5.10 AUC Curve for XGBoost	42

Figure 5.11 Adaboost Model Performance in experiment 143
Figure 5.12 AUC Curve for AdaBoost43
Figure 5.13 LightGBM Model Performance in Experiment 2.....44
Figure 5.14 XgBoost Model Performance in Experiment 244
Figure 5.15 AdaBoost Model Performance in Experiment 2.....45