SENTIMENT ANALYSIS OF SINHALA TWEETS

Warna Ieshaka Karunarathne

(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 Karunarathne

(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.Karunarathne

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. Uthyasanker Thayasivam

ACKNOWLEDGEMENT

I would like to convey my genuine appreciation to my supervisor Dr. Uthyasanker 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

Table of tables

Table of figures