SENTIMENT ANALYSIS OF SINHALA NEWS COMMENTS

Isuru Udara Liyanage

(168243P)

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

Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

May 2018

DECLARATION

I declare that this is my own work and this dissertation does not incorporate without acknowledgement 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 acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my thesis/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).

---------- ----------

Isuru Udara Liyanage Date

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

---------- ----------

Dr. Surangika Ranathunga Date

ABSTRACT

Mining sentiment values from unstructured text uncovers interesting patterns that can be effectively used for many applications. One interesting yet poorly explored area is online news comment analysis, in particular for Sinhala language. Despite the uptrend in online Sinhala news articles and related comments, no efficient method exists for analyzing and identifying the public sentiment associated with them. In this research our effort is to classify online Sinhala news comments according to its sentiment orientation.

Most of the sentiment analysis research is done for English language. As for Sinhala, only one research can be found for classification of Sinhala news comments according to its sentiment values. Since it is an initial attempt it lacks the use of advanced text analysis methods and localization, and hence can be improved in many ways.

In this research we build a complete Sinhala sentiment analysis system, from data collection to sentiment classification. First we gather a dataset by crawling through a popular online news site. Complied dataset contains news items and related comments. Sufficient amount of comments are annotated according to its sentiment values. Finally sentiment analysis is carried out to identify sentiment values associated with each comment.

This research provides many valuable outputs to the research community, sentiment analysis for Sinhala text. Dataset, the labeled data set in particular, can be used for future Sinhala text analysis research. Finally direction and a baseline will be set for future research on sentiment analysis for Sinhala text.

ACKNOWLEDGEMENT

I would like to express my deepest appreciation to my thesis advisor Dr. Surangika Ranathunga for her continuous guidance throughout the research work. She steered me in the right direction from the beginning by providing valuable insights, resources and supervision. Without her guidance and motivation this wouldn't have been a success.

I am grateful to Dr. Charith Chitraranjan, Dr. Shehan Perera, Dr. Malaka Walpola and Dr. Indika Perera for their motivation and guidance for making this a success.

I would also like to thank my colleges at DirectFN for their support and the understanding during the MSc duration.

Finally I express my heartfelt gratitude for my parents for their love and support throughout my life.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

NLP	Natural Language Processing
POS	Part Of Speech
SVM	Support Vector Machines
SO	Sentiment Orientation
PMI	Pointwise Mutual Information
WSD	Word Sense Disambiguation
FNE	Focused Named Entities