AN ANALYTICAL STUDY OF PRE-TRAINED MODELS FOR SENTIMENT ANALYSIS OF SINHALA NEWS COMMENTS

M. Lishani Sadna Dissanayake

189314L

M.Sc. in Computer 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).

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

M. Lishani Sadna Dissanayake Date

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

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

Dr. Uthyasanker Thayasivam Date

ABSTRACT

In the area of natural language processing, due to the large-scale text data availability sentiment analysis has become a prevalence topic. Sentiment analysis is a text classification which is mainly focusing on classifying recommendations and reviews as positive or negative. Earlier for this classification task, most of methods require product reviews and label them. Using these reviews then a classifier is trained with their relevant labels. For this training procedure a huge number of labeled data is needed to train these classification models for each of the product, considering the facts that the distribution of the reviews can be different between different domains and to enhance the performance of these classification models. Nevertheless, the procedure of labeling the data is very expensive and time consuming. For low resource languages like Sinhala language, the existence of annotated Sinhala data is limited compared to the languages like English language. The need of applying classification algorithms in order to perform sentiment classification for Sinhala language is challenging. Apart from applying traditional algorithms to analyze sentiments, here using pre-trained models(PTM)s, experimenting on whether the outcome of these experiments outperform the traditional methods. In natural language processing, PTM is performing an important role, since it paves the way for applying PTMs for downstream tasks. Therefore, this research takes the step to applying PTMs such as BERT and XLnet to classify sentiments. Experiments have been done using two approaches on BERT model as fine tuning the BERT model and feature based approach. Also using the existing Roberta-based Sinhala models, named as SinBERT-small and SinBERT-large which are available in Huggingface official site which have trained using a large Sinhala language corpus.

ACKNOWLEDGEMENT

First and foremost, I would like to convey my deepest appreciation to my supervisor Dr. Uthyasanker Thayasivam for his continuous support, invaluable advice, determined efforts and assistance. I have been immmensly fortunate to have him as my supervisor, whose plentiful experience and immense knowledge have guided me to making this research a success.

Dr. Surangika Ranthunga has also guided me in some things and therefore I would like to express my gratitude towards her.

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

Finally, none of this would have been achievable without the love and patience of my parents. They have been a constant source of love, strength, concern and support all these years. I am grateful to my parents for their tremendous understanding, motivating and being patient and supportive during this critical period in my academic life, which motivated me to complete the research successfully.

Table of Content

List of Figures

List of Tables

List of Equations