

**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. Uthayasanker 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 immensely 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

Declaration.....	i
Abstract.....	ii
Acknowledgement.....	iii
1 Introduction.....	1
1.1 Background.....	1
1.2 Research Problem.....	1
1.3 Research Objective.....	2
2 Literature Survey.....	3
2.1 Sentiment Analysis.....	7
2.2 Related Work.....	9
2.2.1 Sentiment Analysis Using Transfer Learning.....	9
2.2.2 Sentiment Analysis Using Transfer Learning for Other Languages.....	15
2.2.3 Sentiment Analysis Using Transfer Learning for Sinhala Language.....	16
2.2 Sentiment Analysis Using Transformer Models.....	17
2.2.1 Transfer Learning and Transformer Models in NLP.....	17
2.2.1 Sentiment Analysis Using BERT.....	20
2.2.2 Sentiment Analysis Using XLNet.....	26
2.2.3 Sentiment Analysis Using ELMO.....	27
2.2.3 Sentiment Analysis Using ULMFit.....	27
2.3 Multilingual Transformer Models.....	28
3 Research Methodology.....	29
3.1 Data Collection.....	29
3.2 Data Preprocessing.....	31
3.3 Sentiment Analysis with BERT.....	33
3.4 Fine Tuning with BERT.....	34
3.5 SinBERT -small and SinBERT-large.....	37
3.6 Architecture of Proposed Model with BERT with Feature Based Approach.....	38
4 Experiments.....	39
4.1 Evaluation - BERT.....	40
4.1 Experiments with SinBERT-small and SinBERT-large without Stop Words.....	44

4.2 Experiments with SinBERT-small and SinBERT-large with Stop Words.....	44
4.3 Experiments on Feature Based Approach Using BERT.....	45
4.4 Evaluation – XLnet.....	46
5.1 Future Work.....	48
References.....	49

List of Figures

Figure 1 Different Learning Approaches among Transfer and Learning Traditional Machine Learning.....	4
Figure 2 An Overview of Different Settings of Transfer.....	6
Figure 3 Graphical illustration of attention.....	18
Figure 4: Achitecture of transformers.....	20
Figure 5: BERT Achitecture compared with other models.....	21
Figure 6: Overview of BERT Architecture.....	22
Figure 7: Count of Positive and negative comments.....	30
Figure 8: Count of Positive and negative comments.....	30
Figure 9: Pre-training and fine-tuning models in BERT.....	33
Figure 10: The layers of BERT architecture.....	34
Figure 11: Architecture of the proposed model of BERT.....	35
Figure 12: Architecture of Proposed Model with SinBERT.....	37
Figure 13: Achitecture of BERT Model with Feature Based.....	38
Figure 14: Analysis for Data Preparation for db1.....	39
Figure 15: Training validation loss and accuracy for dataset.....	40
Figure 16: Training validation loss and accuracy for dataset 2.....	40
Figure 17: Predictions for dataset 1.....	41
Figure 18: Predictions for dataset 2.....	42
Figure 19: Accuracy for dataset 1.....	42
Figure 20: Accuracy for dataset 1.....	42
Figure 21: Benchmark.....	43
Figure 21: Benchmark.....	Error! Bookmark not defined.

List of Tables

Table 1 Relationship among Various Transfer Learning Settings and Traditional Machine Learning	5
Table 2 Different Settings of Transfer Learning	5
Table 3: Experiments with SinBERT without StopWords for db 1.....	44
Table 4: Experiments with SinBERT without StopWords for db2.....	44
Table 5: Experiments with SinBERT with StopWords for db1.....	44
Table 6: Experiments with SinBERT wit StopWords for db2.....	44
Table 7: Experiments with Feature Based for db1.....	45
Table 8: Experiments with Feature Based for db2.....	46

List of Equations

Equation 1: Formula for y at time t.....	18
Equation 2: Formula for st.....	18
Equation 3: Formula for ct.....	18