EXPLOITING MULTILINGUAL CONTEXTUAL EMBEDDINGS FOR SINHALA TEXT CLASSIFICATION

G.V. Dhananjaya

218039D

Master of Science (Major Component Research)

Department of Computer Science and Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

May 2022

EXPLOITING MULTILINGUAL CONTEXTUAL EMBEDDINGS FOR SINHALA TEXT CLASSIFICATION

G.V. Dhananjaya

218039D

Thesis submitted in partial fulfillment of the requirements for the degree
Master of Science (Major Component Research)

Department of Computer Science and Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

May 2022

DECLARATION

I declare that this is my own work and this Thesis 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. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: Date: 2023/04/20

The supervisors should certify the Thesis with the following declaration.

The above candidate has carried out research for the Master of Science (Major Component Research) Thesis under our supervision. We confirm that the declaration made above by the student is true and correct.

Name of Supervisor: Dr. Surangika Ranathunga

Signature of the Supervisor: Surangika Ranathunga Digitally signed by Surangika Ranathunga Date: 2023.05.17 00:19:18 +12'00' Date:

Name of Supervisor: Prof. Sanath Jayasena

Signature of the Supervisor: Date: 22/05/2023

DEDICATION

I dedicate this research work to all the individuals who supported me in academics, including my family, friends, and teachers.

ACKNOWLEDGEMENT

I owe gratitude to my supervisors, Dr. Surangika Ranathunga and Prof. Sanath Jayasena for the immense support, guidance and supervision they provided to complete this work. I would also like to thank Mr. Piyumal Demotte who overtook the task of pre-training SinBERT language models.

ABSTRACT

Language models that produce contextual representations (or embeddings) for text have been commonly used in Natural Language Processing (NLP) applications. Particularly, Transformer based, large pre-trained models are popular among NLP practitioners. Nevertheless, the existing research and the inclusion of low-resource languages (languages that primarily lack publicly available datasets and curated corpora) in these modern NLP paradigms are meager. Their performance for downstream NLP tasks lags compared to that of high-resource languages such as English. Training a monolingual Language model for a particular language is a straightforward approach in modern NLP but it is resource-consuming and could be unworkable for a low-resource language where even monolingual training data is insufficient. Multilingual models that can support an array of languages are an alternative to circumvent this issue. Yet, the representation of low-resource languages considerably lags in multilingual models as well.

In this work, our first aim is on evaluating the performance of existing Multilingual Language Models (MMLM) that support low-resource Sinhala and some available monolingual Sinhala models for an array of different text classification tasks. We also train our own monolingual model for Sinhala. From those experiments, we identify that the multilingual XLM-R model yields better results in many instances. Based on those results we propose a novel technique based on an explicit cross-lingual alignment of sentiment words using an augmentation method to improve the sentiment classification task. There, we improve the results of a multilingual XLM-R model for sentiment classification in Sinhala language. Along the way, we also test the aforementioned method on a few other Indic languages (Tamil, Bengali) to measure its robustness across languages.

Keywords: Multilingual language models, Multilingual embeddings, Text classification, Sentiment analysis, Low-resource languages, Sinhala language

TABLE OF CONTENTS

Declaration of the Candidate & Supervisor i
Dedication ii
Acknowledgement iii
Abstract iv
Table of Contents v
List of Figures viii
List of Tables ix
List of Abbreviations ix
1 Introduction 1
1.1 Research problem 2
1.2 Research objectives 2
1.3 Contributions 3
1.3.1 Thesis organization 3
2 Theoretical Background 4
2.1 Overview 4
2.2 Sinhala Language 4
2.3 Transformer-based pre-trained models 4
2.3.1 Encoder of the Transformer 6
2.3.2 Transformer decoder module 8
2.3.3 Overall Transformer architecture 8
2.3.4 Tokenizers 9
2.4 Transformer based deep-learning language models 9
2.4.1 BERT 10
2.4.2 mBERT 10
2.4.3 RoBERTa 10
2.4.4 XLM-R 10
2.4.5 LaBSE 11

2.4.6 LASER 11
2.5 Utilizing multilingual models for target tasks 12
2.5.1 Vanilla fine-tuning 12
2.5.2 Other fine-tuning methods 12
2.5.3 Transfer Learning in Transformer models 13
2.5.4 Continual Learning 14
2.6 Measuring the performance of the classification models 14
3 Literature review 16
3.1 Monolingual and multilingual pre-trained models 16
3.2 Improving text classification results on Transformer based pre-trained language models 20
3.2.1 Text-classification of low-resource languages 21
3.2.2 Sinhala Text Classification 23
4 Evaluating Transformer-based encoder models 25
4.1 Task description 25
4.2 SinBERT models 25
4.3 Related work 26
4.3.1 Other existing Sinhala monoligual models 26
4.4 Datasets and sub-tasks 27
4.4.1 Sentiment analysis 27
4.4.2 News category classification 27
4.4.3 News source classification 27
4.5 Methodology 28
4.5.1 Fine-tuning method used in our work 29
4.6 Results 31
4.7 Discussion 34
5 Improving Encoder-based Language Models for Sinhala Text Classification 35
5.1 Task Description 35
5.2 Alternative methods 35
5.2.1 Translation of input text from Sinhala to English 35
5.2.2 Instance weighting 36

5.2.3 Using different prediction heads 36
5.2.4 Sequential Multi-task learning 36
5.2.5 Data augmentation 37
5.2.6 Adapter based methods 37
5.2.7 Ensembling model weights 37
5.2.8 Injection of external vector embeddings 39
5.2.9 Initial results 40
5.3 Proposed Technique 40
5.3.1 Motivation and related work 40
5.3.2 Steps of the proposed technique 41
5.4 Implementation details 42
5.4.1 Datasets and Lexicons 42
5.4.2 Experimentation setup and Baseline 44
5.5 Results 45
5.5.1 Ablation studies 45
5.6 Adapter based experiments 47
5.7 Discussion 49
6 Conclusion and Future work 51
References 53

LIST OF FIGURES

Figure	Description	Page
Figure 2.1	Traditional RNN architecture	5
Figure 2.2	Transformer model architecture [1]	6
Figure 4.1	Utilizing the CLS token representation	30
Figure 4.2	Classifier head composition for Transformer models	31
Figure 4.3	macro-F1 for sentiment classification task with varying dataset size using SinBERT and XLM-R-base	32
Figure 4.4	macro-F1 for news source classification task with varying dataset size using SinBERT and XLM-R-base	33
Figure 4.5	macro-F1 for news category classification task with varying dataset size using SinBERT and XLM-R-base	33
Figure 4.6	macro-F1 for writing style classification task with varying dataset size using SinBERT and XLM-R-base	34
Figure 5.1	Adapters introduced with the MAD-X adapter framework [2]	38
Figure 5.2	Injecting external embedding vectors	39
Figure 5.3	Selecting APs.	43
Figure 5.4	Two stage fine-tuning method proposed.	45
Figure 5.5	macro-F1 scores for the results by changing the no. of APs and no. of word	46
Figure 5.6	Visualization of XLM-R-base word embeddings without applying our technique (circle markers) and after applying our technique (triangle markers). The blue markers represent positive sentiment words and red markers represent negative sentiment words.	48

LIST OF TABLES

Table	Description	Page
Table 3.1	Sizes of corpora used for pre-training several monolingual language models	16
Table 3.2	Portion of Sinhala included in model pre-training of a few multilingual models	17
Table 4.1	Parameters of the SinBERT models	26
Table 4.2	Statistics of the pre-training corpus	26
Table 4.3	Statistics of the 4 classification datasets	28
Table 4.4	Hyperparameters for the experiments for evaluating language models	29
Table 4.5	Results for the model evaluation on the four classification tasks	32
Table 5.1	Results obtained from the weight ensembling experiment	38
Table 5.2	Initial results obtained using each alternative method	40
Table 5.3	Parameters used for each dataset	45
Table 5.4	macro-F1 scores of experiments comparing our method against each language's baseline on XLM-R (base)	46
Table 5.5	Results for experiments with varying attributes of the APs for Sinhala sentiment dataset.	47
Table 5.6	Macro-F1 scores from adapter based experiments with the proposed method	48

LIST OF ABBREVIATIONS

Abbreviation	**Description**
AP	Auxiliary phrases
CL	Continual Learning
MLM	Masked Language Modeling
MMLM	Multilingual Language Models
NER	Named Entity Recognition
NLG	Natural Language Generation
NLI	Natural Language Inference
NLP	Natural Language Processing
NLU	Natural Language Understanding
NSP	Next Sentence Prediction
POS	Part-of-Speech tagging
RNN	Recurrent Neural Network
TLM	Translation Language Modelling