Utilizing RoBERTa Intermediate Layers and Fine-Tuning for Sentence Classification

Eaint Thet Hmu Soe
Natural Language Processing Lab
University of Computer Studies
Yangon, Myanmar
eaintthethmusoersc@gmail.com

Win Pa Pa
Natural Language Processing Lab
University of Computer Studies
Yangon, Myanmar
winpapa@ucsy.edu.mm

Abstract*—Text classification becomes more and more challenging due to a scarcity of standardized labeled data in the Myanmar NLP domain. The majority of the existing Myanmar research has relied on models of deep learning that significantly focus on context-independent word embeddings, such as Word2Vec, GloVe, and fastText, in which each word has a fixed representation irrespective of its context. Meanwhile, context-based pre-trained language models such as BERT and RoBERTa recently revolutionized the state of natural language processing. In this paper, we conducted the experiments to enhance the performance of classification in sentiment analysis by utilizing the transfer learning ability of RoBERTa. Existing pretrained model based works only utilize the last output layer of RoBERTa and ignore the semantic knowledge in the intermediate layers. This research explores the potential of utilizing RoBERTa intermediate layers to enhance the performance of fine-tuning of RoBERTa. To show the generality, we also compared Myanmar pretrained RoBERTa model (myanBERTa)[1] and multilingual pretrained model (XLM-roBERTa)[2]. The effectiveness and generality of intermediate layers were proved and discussed in the Experimental results.

***Keywords*—Transfer Learning, Fine Tuning, RoBERTa, myanBERTa, Sentiment Analysis, Myanmar Language, Pretrained Model**

1. INTRODUCTION

The success of deep learning relies on the huge amount of labeled data in many applications. Large quantities of tagged data are typically difficult or expensive to gather. Researchers have turned to transfer learning to solve this problem. Transfer learning takes into account the situation where we have little labeled data from the target domain for a particular task but have many relevant tasks with a lot of data from other domains (also known as out-of-domain data). The objective is to transfer knowledge from the high-resource domains to the low-resource target domain.

The lack of resources is the main issue in resolving new NLP research in the Myanmar language, a low resource language. Analysis task of classification is labeled by domain experts and this manual labeling process is intensively expensive. Pre-trained language models can leverage large amounts of unlabeled data to learn the universal language representations, which provide an effective solution for the above problem. Pre-trained transformer-based masked language models such as BERT, RoBERTa and ALBERT had a dramatic impact on the NLP landscape in recent years. A pre-trained model is often trained on a supervised downstream dataset for a few epochs, which is known as fine-tuning, in the common recipe for using such models.

In this research, we explore the best approach for the classification using the pre-trained contextualized language model RoBERTa and compare the results with the language specific model and multilingual model. The proposed architecture can be used in any low resource classification problem to increase the accuracy than the traditional machine learning and data intensive deep learning approach.

2. RELATED WORK

Many fine-tuning models have been proposed for the sentiment analysis task. Weibo Text Sentiment Analysis Based on BERT and Deep Learning[3], this paper used BERT to represent the text with dynamic word vector, BiLSTM to extract the contextual feature and CNN architecture to extract the important local sentiment features on the COVID-19 weibo text dataset. Another approach investigated the effectiveness of the BERT embedding component on the task of End-to-End aspect-Based Sentiment Analysis (E2E-ABSA)[4] .They explore to couple the BERT embedding component with various neural models and conduct extensive experiments on two benchmark datasets.

3. METHODOLOGY

Main contributions of this research can be summarized as follows:

I. Two pooling approaches - LSTM pooling and weighted pooling
II. Two pre-trained models - language specific MyanBERTa and multilingual xml-RoBERTa-base

The proposed architecture compares with the baseline approach.

3.1 Baseline approach

The baseline approach utilizes the CLS token of the final layer.

3.2 LSTM pooling approach

Hidden state representation hCLS is a unique sequence: an abstract-to-specific sequence. Because the LSTM network is inherently suitable for processing sequential information, we use it to connect all intermediate representations of the [CLS] token, and the output of the last LSTM cell is used as the final representation.

3.3 Weighted pooling approach

Intuitively, attention operations can learn the contribution of each CLS. We use a dot-product attention module to dynamically combine the last four intermediates.

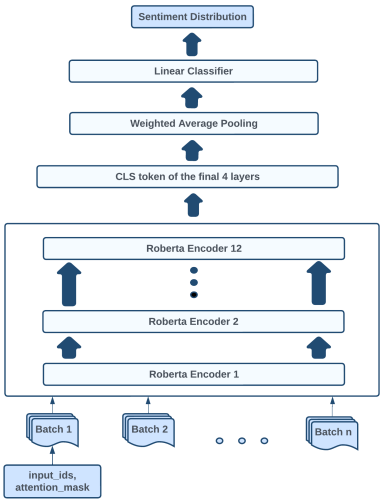


Figure 1: Overview of the proposed RoBERTa weighted Pooling model. Pooling Module is responsible for connecting the intermediate representations obtained by Transformers of RoBERTa

4. Experiments and Results

4.1. Dataset Information

The dataset is crawled from the comments of the Myanmar Celebrity facebook page. After crawling, this dataset is tagged by humans manually to get the positive, neutral and negative. The following sample is from the sentiment dataset. The dataset contains 72K sentences.

Sentence	label
ဖြူစင် တယ် စကားပြော ကြည့် တာ နဲ့ သိ တယ်	positive
အောင်မလေး၊ ချစ်စရာ လေး၊	positive
ကြော်ငြာ လေး၊ လျှော့ ပေး ပါ လား၊	neutral
စာလုံးပေါင်း မှန် အောင် ရေး ပါ အစ်မ	neutral
ကိုယ် ထမင်း၊ ကိုယ် စား၊ ပီး သူများ ဝေဖန် နေ တာ မင်းတို့ရှုပ် တွေ့ အရင် မှန် ထဲ ပြန် ကြည့် အလကား၊ အကုသိုလ် ကောင် တွေ့	negative
စောက်ရူးကောင် ကြောင်တောင်တောင် နဲ့	negative
ကျက်သရေ အဖြာဖြာ တုံး တယ်	negative

Table 1: Visualization of the dataset example after preprocessing

To analyze the nature of the dataset, we used word2vec and PCA (Principal component analysis). We select 1K sentences from each class and the amount of information from word2vec 100 dimension feature space reduces with PCA components 3. We separate the classes with colors; yellow as positive, green as negative and purple as neutral.

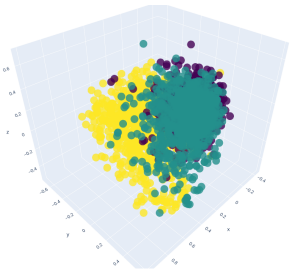


Figure 2: Visualization with Word2vec embedding into PCA over the selected sentiment dataset

4.2. Experiment Setting

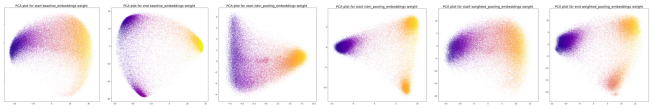
All experiments are conducted with MyanBERTa and xlm-RoBERTa-base with different pooling strategies. In order to compare the different models, we divided our dataset 20% on a test set and we used the same training set and testing set for all experiments. For training parameters, we set up Adam optimizer and linear learning rate scheduler.

Pretrained Model	Model Type	Layers	Optimizer	Scheduler	Learning Rate	Epochs	loss function
myanBERTa	baseline	last	Adam	linear	2.00E-05	5	Cross Entropy
	lstm pooling	all	Adam	linear	2.00E-05	5	Cross Entropy
	weighted pooling	4 layers	Adam	linear	2.00E-05	5	Cross Entropy
xlm-RoBERTa-base	baseline	last	Adam	linear	2.00E-05	5	Cross Entropy
	lstm pooling	all	Adam	linear	2.00E-05	5	Cross Entropy
	weighted pooling	4 layers	Adam	linear	2.00E-05	5	Cross Entropy

Table 2: Model parameters setting

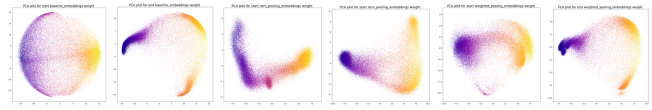
4.3. Visualization of the intermediate layers

We use principal component analysis (PCA) to visualize the intermediate representations of the [CLS] token to show how the different pooling strategies benefit from sequential representations of intermediate layers. Because the task-specific information is primarily extracted from the last layers of RoBERTa, we only depict the layer that gets from the pooler. We compare the initial epoch and final epoch to make it visible the effect of the transfer learning ability with a small amount of epoch. It is simple to conclude that the model divides distribution into three sentiment classes.



(a) (b) (c)

MyanBERTa + baseline , LSTM pooling and weighted pooling (left Side initial epoch and right side final epoch)



(d) (e) (f)

XLM-RoBERTa-base + baseline , LSTM pooling and weighted pooling (left Side initial epoch and right side final epoch)

Figure 3: Visualization of RoBERTa based three pooling strategies at the end of the first epoch and fifth epoch. Among PCA results, (b) demonstrates myanBERT+LSTM pooling coverages better than other myanBERT models, (f) xlm-RoBERTa+weighted pooling cluster each class of data more dense and discriminative

Pretained Model	Model Type	Accuracy	Recall	Precision	F1-Score
myanBERTa	baseline	84.935662	84.935662	85.166369	85.027487
	lstm pooling	84.779412	84.779412	84.923734	84.837495
	weighted pooling	84.246324	84.246324	84.403939	84.314696
xlm-RoBERTa-base	baseline	86.047794	86.047794	86.033751	86.037324
	lstm pooling	83.823529	83.823529	84.08864	83.935806
	weighted pooling	86.424632	86.424632	86.494176	86.445854

Table 3: Comparison on the evaluation matrices

5. Conclusion

In this work, we compare and contrast the utilization of the intermediate representation of the RoBERTa model. The effectiveness of transfer learning and fine-tuning are analyzed. We explore the effectiveness of the RoBERTa model in the generalization and the pooling strategies to make the model perform better. The LSTM pooling method increases accuracy in the language specific RoBERTa than the multilingual model. By examining the evaluation table, the xlm-RoBERTa-base and weighted pooling strategy outperform the others.

Acknowledgment

We thank Bagan Innovation Technology for providing the dataset.

References

[1] Aye Mya Hlaing, Win Pa Pa, MyanBERTa: A Pre-trained Language Model For Myanmar, 2022

[2] Alexis Conneau* Kartikay Khandelwal*, Naman Goyal Vishrav Chaudhary Guillaume Wenzek Francisco Guzm an,Edouard Grave Myle Ott Luke Zettlemoyer Veselin Stoyanov, "Unsupervised Cross-lingual Representation Learning at Scale", 2020

[3] Youwei Song, Jiahai Wang *, Zhiwei Liang, Zhiyue Liu, Tao Jiang, "Utilizing BERT Intermediate Layers for Aspect Based Sentiment Analysis and Natural Language Inference", 2020

[4] Xin Li1, Lidong Bing2, Wenxuan Zhang1 and Wai Lam , "Exploiting BERT for End-to-End Aspect-based Sentiment Analysis", 2019