A Study on a Joint Deep Learning Model for Myanmar Text Classification

Myat Sapal Phyu
Faculty of Computer Science
University of Information Technology
Yangon, Myanmar
myatsapalphyu@uit.edu.mm

Khin Thandar Nwet
Faculty of Computer Science
University of Information Technology
Yangon, Myanmar
khinthandarnwet@uit.edu.mm

Abstract

Text classification is one of the most critical areas of research in the field of natural language processing (NLP). Recently, most of the NLP tasks achieve remarkable performance by using deep learning models. Generally, deep learning models require a huge amount of data to be utilized. This paper uses pre-trained word vectors to handle the resource-demanding problem and studies the effectiveness of a joint Convolutional Neural Network and Long Short Term Memory (CNN-LSTM) for Myanmar text classification. The comparative analysis is performed on the baseline Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN) and their combined model CNN-RNN.

Keywords—*text classification, CNN, RNN, CNN-RNN, CNN-LSTM, deep learning model.*

I. INTRODUCTION

Text classification is one of the interesting application areas in NLP such as sentiment analysis, machine translation, automatic summarization, etc. Nowadays, information overloading is one of the important problems for today's people. People waste too much time to select their interesting information because they receive a vast amount of information from various sources of internet. A powerful text classifier can extract useful content from massive amounts of information. It can classify text into associated categories including sentiment analysis, spam detection, articles, books, and hate speech detection. It is among the most active research areas in the field of NLP. Historically, many researchers performed text classification by using various machine learning algorithm and their variant models. The use of deep learning models has achieved great interest in text classification due to their ability to capture semantic word relationships in recent years. All deep learning models typically require a lot of data to accomplish a specific task. Although most pre-trained vectors are trained on general datasets, the use of pre-trained vectors can handle data requirements. It can be used for the specific task by transferring learning with the appropriate amount of data. This paper focuses in particular on the classification of Myanmar articles. Words are considered as basic units in this paper. Since Myanmar language has no standard rule to determine word boundaries, it is important in the pre-processing phase to determine the word boundaries. As Myanmar language is a morphologically rich language, good word representation is difficult to learn because many word forms seldom occur in the training corpus. The BPE tokenizer[1] is used to determine the boundary of the word. The segmented words are converted into an embedding matrix by using the pre-trained vectors trained on Wikipedia by the Skip-gram model [2]. The effective use of pre-trained vectors is very supportive of resource-scarce languages. In this paper, CNN and LSTM models are jointly performed for Myanmar text classification. The convolution process is performed on the embedding matrix for selecting the features and the LSTM model is used to capture long term information. In this paper, the max-pooling layer is dropped in the CNN model to prevent the loss of context information.

The next sections are as follows, the works related to the text classification and other areas of application are addressed for both Myanmar and English in section 2. Section 3 describes Myanmar's background knowledge. Section 4 discusses the components of the CNN-LSTM joint model. Section 5 explains the findings of the experiment and the paper is concluded in section 6.

II. RELATED WORK

Wang et al. [8] proposed a regional CNN-LSTM model that captures regional information by CNN and predicts valence arousal (VA) by LSTM

[1] https://github.com/bheinzerling/bpemb

model and outperforms than regression-based, lexicon-based and NN-based methods on two datasets for sentiment analysis. Kwaik et al. [5] proposed a deep learning model that combines LSTM and CNN models for dialectal Arabic sentiment analysis and this model performs better than two baseline models on three datasets. Zhang et al. [9] constructed the CNN model on the top of LSTM. The output from the LSTM model is further extracted by the CNN model and it performed better in terms of accuracy than the baseline models. Kim et al. [4] conducted the experiments with variations of CNN model, CNN-rand, CNN-static and CNN-multichannel on the top of pre-trained vectors for sentence classification. These CNN models performed better in 4 out of 7 tasks than the state-of-the-art.

The previous research works in deep learning and machine learning models for Myanmar language in speech recognition, named entity recognition, and text classification are also investigated. Aye et al. [1] aimed to enhance the sentiment analysis for Myanmar language in the food and restaurant domain by considering intensifier and objective words and improved the prediction accuracy. Hlaing et al. [3] applied LSTM-RNN in Myanmar speech synthesis. Mo et al. [6] annotated Named Entity tagged corpus for Myanmar language and evaluation was performed comparatively between neural sequence model and baseline CRF. In our previous works [7], we performed the comparative analysis of CNN and RNN both on syllable and word level by using three pre-trained vectors and also collected and annotate six Myanmar articles datasets. In this paper, a comparative analysis is performed on a joint CNN-LSTM model against baseline CNN, RNN, and their combine model by using Myanmar articles datasets.

III. MYANMAR LANGUAGE BACKGROUND

Myanmar language is the official language of the Republic of the Union of Myanmar. It is a morphologically rich language and Myanmar sentences are basically constructed as the subject, object, and verb pattern. The Myanmar script is written from left to right and the characters are rounded in appearance. There is no regular inter-word spacing in Myanmar language like in the English language. Though, spaces are used to mark phrases. Sentences are clearly delimited by a sentence boundary marker. Myanmar words are constructed by one or more syllables. There are thirteen three basic consonants, eight vowels and eleven medial. A Myanmar syllable is constructed by one initial consonant (C), zero or more medial (M), zero or more vowels (V) and optional dependent various signs. Words are considered a basic unit in this paper.

A. Pre-processing

Pre-processing basically contains two steps, 1) removing unnecessary characters, and 2) determining the word boundaries. As this study focus on the Myanmar text classification, Myanmar Unicode range between [U1000-U104F][2] is removed to ignore non-Myanmar characters. The numbers "၀-၉", [U1040-U1049] and punctuation marks "၊, ။", [U104A- U104B] are also removed. As Myanmar language has no rule to determine word boundary, it is needed to determine the boundary of words. In this work, word boundaries are determined by the BPE tokenizer.

B. Pre-trained Model

The reuse of the pre-trained model by transfer learning on a new task is very efficient because it can train the deep learning models with not much data. Typically, most of the NLP tasks don't have sufficient label data to be trained on such complex models. In this paper, we use the pre-trained vector file that was publicly released by the Facebook AI Research (FAIR) lab[3]. It was trained on Wikipedia using the fastText skip-gram model with 300 dimension. In this paper, the pre-trained model is used as the starting point instead of learning from scratch.

C. Construction of Embedding Matrix

The segmented words are transformed into word vectors by matching the vocabulary in the pre-trained vectors file. Figure 1 shows the construction of the embedding matrix. Firstly, Myanmar text data are extracted from online news websites. Secondly, unnecessary characters are removed from the extracted text. Then, word boundaries are determined by the BPE tokenizer although sometimes the results are meaningless as it is a frequency-based tokenizer. The tokenized words are matched with pre-trained word vectors file in order to construct the word embedding matrix for the embedding layer.

Table I shows the sample of Myanmar text pre-processing with the sample sentence "Corona ဗိုင်းရပ်စ်

[2] https://mcf.org.mm/myanmar-unicode/ [3] https://fasttext.cc/docs/en/pretrained-vectors.html

ကာကွယ်ရေး မြန်မာတိုးမြှင့်ပြင်ဆင်။” and non-Myanmar characters “Corona” is removed. Then, the text string is segmented into words as “ဗိုင်းရပ်စ်_ကာကွယ်ရေး_မြန်မာ_တိုးမြှင့်_ပြင်ဆင်”, “_” shows the boundary of words.

TABLE I. SAMPLE OF TEXT PRE-PROCESSING

Input Sentence	Corona ဗိုင်းရပ်စ် ကာကွယ်ရေး မြန်မာတိုးမြှင့်ပြင်ဆင်။
English Meaning	Myanmar ramps up the defense against Coronavirus
Word Segmentation	ဗိုင်းရပ်စ်_ကာကွယ်ရေး_မြန်မာ_တိုးမြှင့်_ ပြင်ဆင်

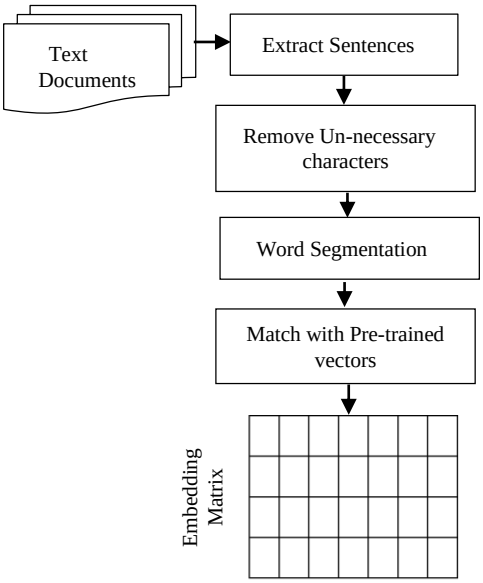

Figure 1. Construction of embedding matrix.

IV. A JOINT CNN-LSTM MODEL

The proposed joint CNN-LSTM model basically consists of four components, 1) Embedding Layer, 2) Convolution Layer, 3) LSTM layer, and 4) Fully connected and output layer. Figure 2 illustrates the joint CNN-LSTM model.

A. Embedding Layer

In the embedding layer, segmented words are transformed into vector representation by matching the pre-trained vectors file. The pre-trained vectors file is like vocabulary and each word in the vocabulary attached with their corresponding vectors that can catch the context information.

B. Convolution Layer

In the convolution layer, the convolution process is performed by the ReLU activation function with stride size 1. Convolution layer selects the features and the result of the convolution process is in the form of the feature map. Max-pooling layer is discarded because it captures only very important features and ignores the un-important features and it can lead to losing the context information.

C. LSTM Layer

LSTM layer is used instead of the max-pooling layer to capture context information.

D. Fully Connected Layer and Output Layer

In the final output layer, the probability of the class is predicted by using the sigmoid activation function. In the hyper-parameter setting, Adam optimization function and binary_crossentropy loss function with 0.5 dropouts and 16 batch size on 10 epochs. Moreover, l2=0.01 is set in kernel and bias regularizer to reduce overfitting.

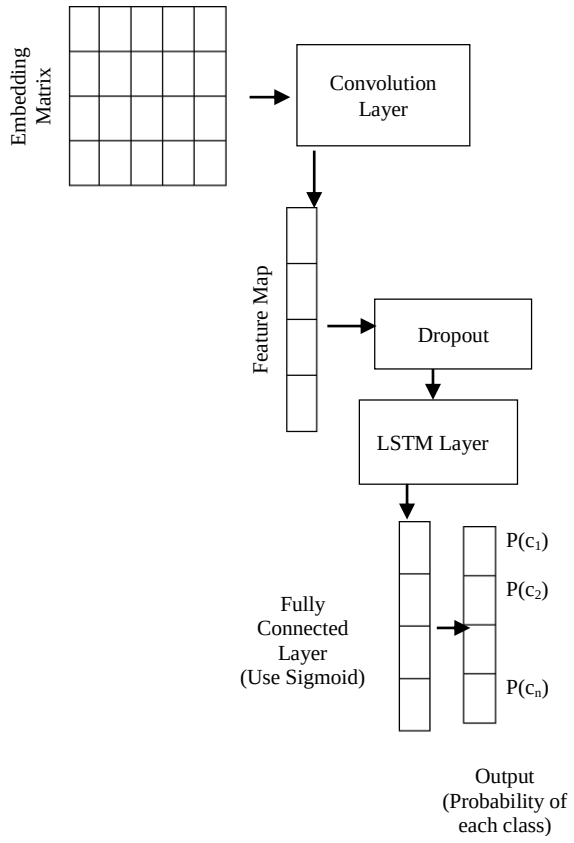

Figure 2. A joint CNN-LSTM model

V. EXPERIMENT

A. Dataset

Myanmar text data are collected from various sources of Myanmar news websites including 7Day Daily[4], DVB[5], The Voice[6], Thit Htoo Lwin[7], and Myanmar Wikipedia[8]. The text data contains five categories art, sport, crime, education, and health. Each category is collected in tab-separated values (.tsv) files and each row contains a sentence that is annotated with the corresponding label. The text data are converted from Zawgyi to Unicode by Rabbit converter[9] and shuffle and split 75% and 25% for train and test dataset for each topic and the total number of sentences for the training is 36,113 sentences and the testing is 12,037 sentences. The details of the dataset are listed in Table II.

TABLE II. MYANMAR TEXT DATASET FOR FIVE TOPICS

Class	Train	Test	Total
Sport	9,919	3,242	13,161
Art	9,483	3,173	12,656
Crime	6,718	2,184	8,902
Health	4,743	1,632	6,375
Education	5,250	1,806	7,056
Total	**36,113**	**12,037**	**48,150**

B. Comparison Models

In this work, the comparative analysis is performed on a joint CNN-LSTM model with baseline CNN, RNN, and their combined model CNN-RNN.

1) CNN: CNN is a feed-forward neural network and a basic CNN contains three layers, convolution layer, pooling layer, and fully connected layer. The ReLU activation is mostly used in the convolution layer, max-pooling is mostly used in the pooling layer and Softmax function is mostly used in the fully connected layer. In this paper, the simple CNN with one convolution layer and sigmoid function is used in the experiment.

2) RNN: RNN is an artificial neural network with an internal memory that keeps information to persist. It learns from the previous data and performs the same function for all input data. RNN produces the output y_t as in equation[10] (1).

$$y_t = f(W_y h_t) \quad (1)$$

$$h_t = \sigma(W_h h_{t-1} + W_x x_t) \quad (2)$$

3) CNN-RNN: In this combined model, the CNN model is used for feature extraction and RNN is used to make a prediction by past words. The drawbacks CNN model is its locality and it requires many convolution layers to capture long term information. RNN is a bias model and it predicts the semantic of words by past words and reduces performance when predicting long text. LSTM model is an extension of the RNN model to be better preservation of long term dependency problem.

C. Experimental Setup

The experiment is implemented on Google Colaboratory (Colab) [11] that provides the Jupyter notebook environment [12], executes code in Python 3.6.9. No setup is required and runs in the cloud and have a maximum lifetime for each user. The key advantage of Colab is the support of the Tesla K80 GPU accelerator. Gensim[13] 3.6.0 is used to load the pre-trained word vector file. Keras[14] 2.2.5 is run on the TensorFlow backend. It enables fast experimentation of deep learning models and can be run on both CPU and GPU. Large size data such as pre-trained word vector files can be stored in Google drive to reduce file uploading time. The sklearn.metrics[15] 0.22.1 is used to evaluate the classification performance of the models.

D. Experimental Result

The performance of the proposed model, CNN-LSTM model is compared with comparison models described in section 5.2 as listed in Table III. In this paper, the experiment is performed on Myanmar news articles text data that contains five topics and the detail of the dataset is listed in Table II. The experiment contains three main parts pre-processing, word embedding and text classification. In the pre-processing phase, we remove unnecessary characters and segment words by BPE tokenizer. Word embedding matrix is constructed by a pre-trained model via the Gensim library and the CNN, RNN, CNN-RNN and CNN-LSTM models are trained by using Keras, model-level library. All models are trained using 10 epochs, 16 batch size, 0.5 dropout rate, 0.01 l2 bias and kernel regularizer, Adam

[4] https://7daydaily.com/
[5] http://burmese.dvb.no/
[6] http://thevoicemyanmar.com/
[7] http://www.thithtoolwin.com/
[8] https://my.wikipedia.org/wiki/
[9] https://www.rabbit-converter.org/
[10] https://en.wikipedia.org/wiki/Recurrent_neural_network
[11] https://colab.research.google.com/
[12] https://jupyter.org/install.html
[13] https://pypi.org/project/gensim/3.6.0/
[14] https://keras.io/
[15] https://scikit-learn.org/stable/

optimizer and binary_crossentropy loss function. The performance of each model is measured with scikit-learn's classification metrics that report precision, recall, f1-score measured on each class and the highest scores on each evaluation metrics precision, recall and f1-score are highlighted in bold. According to the results of the experiment, a joint CNN-LSTM model performs better in F1-score than the comparison models in all classes. CNN model equally performs better on crime and education domains. The training time for each model is also measured in this paper.

According to the measurement results, the CNN model requires the minimum training time (6 min 1 sec) because only one convolution layer is usedin the experiment. Although the CNN-LSTM model performs better than CNN in three topics in term of F1-score, the CNN model gets at least 4x faster in training time than CNN-LSTM (24 min 5sec) model and at least 3x faster than the remaining models, CNN-RNN (12 min 56 sec) and RNN (13 min 15 sec) model in most datasets. To sum up, the joint CNN-LSTM model outperforms than CNN, RNN and CNN-RNN models in most domains but it requires much training time than other models. The use of simple CNN with one layer convolution faster in training time although the accuracy of the model is slightly degraded in some topics than the CNN-LSTM model.

TABLE III. COMPARISON OF TEXT CLASSIFICATION PERFORMANCE ON FIVE TOPICS

Class	Precision				Recall				F1-score			
	CNN-RNN	RNN	CNN	CNN-LSTM	CNN-RNN	RNN	CNN	CNN-LSTM	CNN-RNN	RNN	CNN	CNN-LSTM
Sport	0.91	0.91	0.88	**0.93**	0.93	0.93	**0.95**	0.94	0.92	0.92	0.92	**0.94**
Art	0.85	0.86	0.87	**0.88**	**0.91**	0.87	0.90	0.90	0.88	0.87	0.88	**0.89**
Crime	0.87	0.86	**0.89**	0.88	0.91	0.92	0.90	**0.93**	0.89	0.89	***0.90***	***0.90***
Health	**0.94**	0.92	0.91	0.91	0.87	0.88	0.89	**0.90**	0.90	0.90	0.90	**0.91**
Education	0.91	0.90	**0.93**	0.91	0.85	0.84	0.84	**0.86**	0.88	0.87	***0.88***	***0.88***

VI. CONCLUSION

This paper performs the comparative experiments on the joint CNN-LSTM model with CNN, RNN and CNN-RNN models on five categories including sport, art, crime, health, and education. We initially experimented with many convolution layers in the CNN model to get higher performance and to catch long term information, but the performance was not increased as expected. Moreover, the max-pooling layer of the CNN model led to the loss of the local context information. So, we use only one convolution layer to extract features and the LSTM layer instead of the max-pooling layer in order to catch long term information and to reduce the loss of context information. According to the experiment, the joint CNN-LSTM model performs better than CNN, RNN, and CNN-RNN models in most domains, but it takes much training time than the remaining models.

ACKNOWLEDGMENT

We deeply thank the anonymous reviewers who give their precious time for reviewing our manuscript. We greatly thanks all of the researchers who shared pre-trained words vectors publicly and their works very helpful to accomplish our works and very useful for resource-scarce languages. We would like to thank a friend who assists to collect and annotate Myanmar text datasets.

REFERENCES

[1] Aye YM, Aung SS. Enhanced Sentiment Classification for Informal Myanmar Text of Restaurant Reviews. In 16th International Conference on Software Engineering Research, Management, and Applications (SERA), IEEE, 2018: 31-36.

[2] Grave E, Bojanowski P, Gupta P, Joulin A, Mikolov T."Learning Word Vectors for 157 Languages". In Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC-2018, 2018 May.

[3] Hlaing AM, Pa WP, Thu YK. Enhancing Myanmar Speech Synthesis with Linguistic Information and LSTM-RNN. In Proc. 10th ISCA Speech Synthesis Workshop, 2019: 189-193.

[4] Kim Y., Convolutional Neural Networks for Sentence Classification. In Proceedings of the 2014 Conference on Empirical Methods in

Natural Language Processing (EMNLP), 2014: 1746-1751.

[5] Kwaik KA, Saad M, Chatzikyriakidis S, Dobnik S. LSTM-CNN Deep Learning Model for Sentiment Analysis of Dialectal Arabic. In International Conference on Arabic Language Processing 2019 Oct 16 (pp. 108-121). Springer, Cham.

[6] Mo HM, Soe KM, Myanmar named entity corpus and its use in syllable-based neural named entity recognition, International Journal of Electrical and Computer Engineering (IJECE), 2020.

[7] Phyu SP, Nwet KT. Article Classification in Myanmar Language. In the Proceeding of 2019 International Conference on Advanced Information Technologies (ICAIT), IEEE, 2019: 188-193.

[8] Wang J, Yu LC, Lai KR, Zhang X. Dimensional sentiment analysis using a regional CNN-LSTM model. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) 2016 Aug (pp. 225-230).

[9] Zhang J, Li Y, Tian J, Li T. LSTM-CNN Hybrid Model for Text Classification. In2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC) 2018 Oct 12 (pp. 1675-1680). IEEE