4. Fine-Tuned Model

Language model pre-training has been shown to be effective in improving many NLP tasks [12]-[15]. These include sentence-level tasks such as natural language inference [16], [17] and paraphrasing [18], which aim to predict the relationships between sentences by analyzing them holistically, as well as token-level tasks such as named entity recognition and question answering, where models are required to produce fine-grained output at the token level [19], [20]. There are two existing strategies for applying pre-trained language representations to downstream tasks: feature-based and fine-tuning. The feature-based approach, such as ELMo [21], uses task-specific architectures that include the pre-trained representations as additional features. The fine-tuning approach, such as the Generative Pre-trained Transformer (OpenAI GPT) [15], introduces minimal task-specific parameters, and is trained on the downstream tasks by simply fine-tuning all pre-trained parameters.

In this paper, we improve the fine-tuning-based approaches by applying BERT: Bidirectional Encoder Representations from Transformers. BERT is conceptually simple and empirically powerful. For this research, we used a fine-tuned model, VANILLA-BERT [22], aiming to improve performance scores. Fine-tuning is straightforward since the self-attention mechanism in the Transformer allows BERT to model many downstream tasks— whether they involve single text or text pairs—by swapping out the appropriate inputs and outputs. For applications involving text pairs, a common pattern is to independently encode text pairs before applying bidirectional cross attention [23], [24].

$$BERT(X) = Transformer\big(Embeddings(X)\big) \tag{7}$$

where $BERT(X)$ represent the output embedding for the input sequence X, X can be a concatenation of a query and a document, $Embeddings(X)$: This involves converting the input tokens into embedding. $Transformer()$: This refers to the transformer architecture, which consists of multiple layers of self-attention and feedforward mechanisms as in Eq. (7).

5. Evaluation Metrics

We used MAP, MRR, P@1, and P@3 evaluation metrics to assess the performance of IR systems by comparing their retrieved results to the ground truth relevance assessments. These metrics are commonly used in IR evaluation to assess the quality of ranking systems. Higher values for these metrics indicate better-performing systems. These performance metrics are commonly used for evaluating neural networks in IR and recommendation tasks: MAP (Mean Average Precision), MRR (Mean Reciprocal Rank), P@1 (Precision at 1), and P@3 (Precision at 3). These equations provide a quantitative measure of the performance of a ranking system based on different aspects such as precision, average precision and reciprocal rank. They are used in IR to assess the quality of ranked lists of documents. Evaluation metrics results ranges from 0 to 1.

6. Experiments and Results

In this work, we trained different deep neural rankings models on the Myanmar news dataset as mentioned in Section 2. The detailed information of the Myanmar news dataset is presented in Table 4. We applied the DRMM [7], MP [8], Duetl [9], KNRM [1], PACRR [10], CONV-KNRM [2], MZ-CONV-KNRM [11] models in advance datasets.

	Number of documents	Number of sentences	Number of words
Training Set	90,607	47,964,418	1,122,242,776
Testing Set	13,940	3,784,204	93,569,044
Validation Set	13,939	2,885,793	67,448,335

Table 4. Statistics of training, testing and validation the Myanmar news dataset.

The results obtained from the experiments can be seen in Fig. 1-4. The comparisons of neural ranking performance on the Myanmar news dataset are illustrated in the following figures: Fig. 1 shows the performance measured by MAP, Fig. 2 shows the performance measured by MRR, Fig. 3 shows the performance measured by P@1, and Fig. 4 shows the performance measured by P@3. It can be observed that CONV-KNRM performs better than other neural ranking models. This demonstrates the versatility and adaptability of CONV-KNRM in addressing the retrieving task across various contexts and the similarity scores of different deep neural ranking models using the Myanmar news dataset.

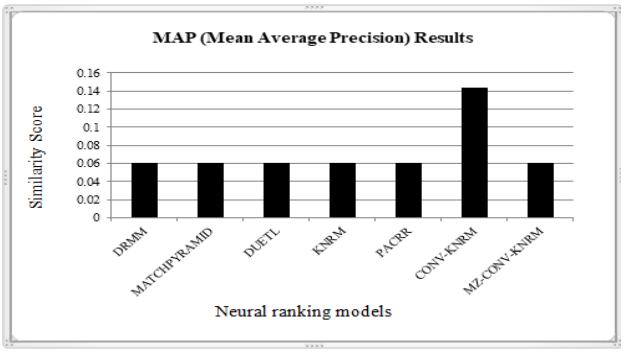

Fig. 1. Comparison of neural ranking performance on the Myanmar news dataset measured by MAP

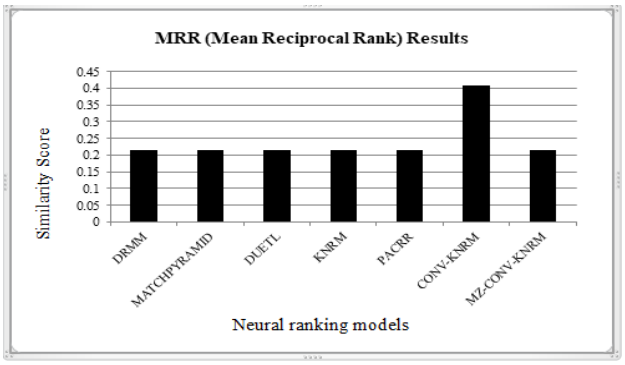

Fig. 2. Comparison of neural ranking performance on the Myanmar news dataset measured by MRR

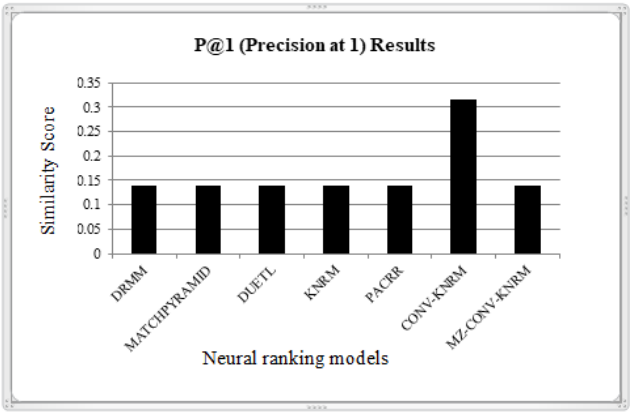

Fig. 3. Comparison of neural ranking performance on the Myanmar news dataset measured by P@1

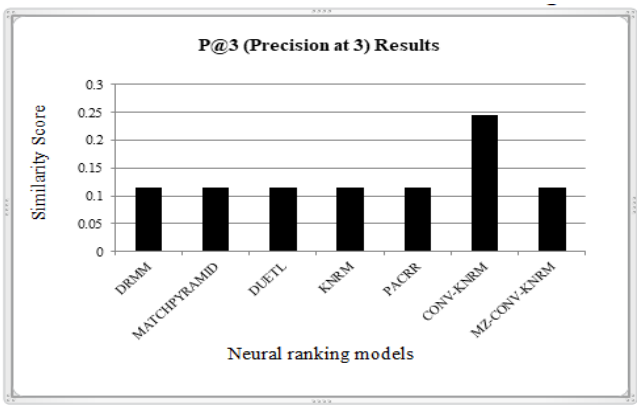

Fig. 4. Comparison of neural ranking performance on the Myanmar news dataset measured by P@3

The best neural ranking model “CONV-KNRM” has been used as a baseline model for our research work. The comparison was done between the fine-tuned ranking model and CONV-KNRM. During fine-tuning, we applied

VANILLA-BERT fine-tuned model to improve the performance of ranking.

Ranking and fine-tuned models	MAP	MRR	P@1	P@3
CONV-KNRM	0.1439	0.4066	0.3150	**0.2450**
VANILLA-BERT	**0.1472**	**0.4415**	**0.3700**	0.2433

Table 5. Comparison of performance on CONV-KNRM and fine-tuned model on the Myanmar news dataset measured by evaluation metrics.

Ranking and fine-tuned models	MAP	MRR	P@1	P@3
CONV-KNRM	0.2031	0.5902	0.4800	0.3717
VANILLA-BERT	**0.2801**	**0.7101**	**0.5950**	**0.4967**

Table 6. Comparison of performance on CONV-KNRM and fine-tuned model on the Antique news dataset measured by evaluation metrics.

As in Table 5, fine-tuning using VANILLA-BERT is found to be better than CONV-KNRM in all evaluation metrics except P@3 on the Myanmar news dataset. Specifically, the MRR results were 0.4066 and 0.4415, which is the best statistically significant difference score results on other evaluation metrics (MAP, P@1 and P@3), whereas for CONV-KNRM and VANILLA-BERT. As this result, we studied that the score results are significantly different in MAP and MRR because MAP measures the average precision at different recall levels, providing an overall assessment of a ranking model's ability to retrieve relevant items across the entire list and MRR calculates the average of the reciprocal ranks of the first relevant items in the ranked lists, emphasizing the model's effectiveness in placing relevant items high in the list.

As in Table 6, the Antique datasets [25] is also used to see the clear performance of our ranking model in the experiments. The Antique datasets consists of 89M questions and answers–pair datasets collection. According to our experiments and results, we observed that the fine-tuned model outperforms CONV-KNRM with the best score of 0.4415 in the Myanmar news dataset and 0.7101 in the Antique dataset in terms of MRR. It can be clearly seen in Tables 5 and 6 that the fine-tuned model achieved better performance than the CONV-KNRM on the Myanmar news dataset, specifically, 0.0349 MRR value higher than the CONV-KNRM, and the fine-tuned model achieved better performance than the CONV-KNRM on the Antique dataset, specifically, 0.1199 MRR value higher than the CONV-KNRM. The experimental results provide interesting results while comparing the performance of different deep neural rankings on the Myanmar news dataset. The results suggest that the choice of fine-tuned technique can significantly impact the performance of the deep neural ranking models.

7. Conclusion

This paper focused on Information Retrieval (IR) for the Myanmar news dataset which contained caption and contents. Different experiments have been conducted, with a wide variety of fine-tuned models and deep neural ranking models. It was observed that the best-performing model is VANILLA-BERT, fine-tuned in this work. The statistical significance of the superior performance has been confirmed by comparing the results of the baseline CONV-KNRM and the fine-tuned model on the Myanmar news and Antique datasets. Our experiments also indicate that the use of fine-tuning techniques can result in significant improvements in the performance of deep neural ranking models for different datasets. The experiment results suggest that fine-tuning approach can potentially be extended to other retrieval applications. Concerning further research as future work, it would be interesting to investigate the effect of adding more features to the textual data. This study adds valuable insights to the ongoing discussions within the field, paving the way for future research endeavors aimed at optimizing models to address a spectrum of challenges in IR.

References

[1] Chenyan Xiong, Zhuyun Dai, James P. Callan, Zhiyuan Liu, and Russell Power. (2017). *End-to-End Neural Ad-hoc Ranking with Kernel Pooling*. In SIGIR.

[2] Zhuyun Dai, Chenyan Xiong, Jamie Callan, and Zhiyuan Liu. (2018). *Convolutional Neural Networks for Soft-Matching N-Grams in Ad-hoc Search*. In WSDM.

[3] Pa, Win Pa, Ye Kyaw Thu, Andrew Finch, and Eiichiro Sumita. 2008. "Wordboundary identification for Myanmar text using conditional random fields." In International Conference on Genetic and Evolutionary Computing.

[4] Mohamed Trabelsi, Zhiyu Chen, Brian D. Davison, and Jeff Heflin, NEURAL RANKING MODELS FOR DOCUMENT RETRIEVAL. ArXiv: 2102.11903v1 [cs.IR] 23 Feb 2021.

[5] B. Croft, D. Metzler, and T. Strohman. Search Engines: Information Retrieval in Practice. Addison-Wesley Publishing Company, 1st edition, 2009.

[6] L. Wang, J. Lin, and D. Metzler. A cascade ranking model for efficient ranked retrieval. In Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '11, pages 105–114, Beijing, China, 2011. ACM.

[7] Jiafeng Guo, Yixing Fan, Qingyao Ai, and William Bruce Croft. 2016. A Deep Relevance Matching Model for Ad-hoc Retrieval. In CIKM.

[8] Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, and Xueqi Cheng. 2016. A Study of MatchPyramid Models on Ad-hoc Retrieval. In

NeuIR @ SIGIR.
[9] Bhaskar Mitra, Fernando Diaz, and Nick Craswell. 2016. Learning to Match using Local and Distributed Representations of Text for Web Search. In WWW.
[10] Kai Hui, Andrew Yates, Klaus Berberich, and Gerard de Melo, PACRR: A Position-Aware Neural IR Model for Relevance Matching. ArXiv: 1704.03940v3 [cs.IR] 21 Jul 2017.
[11] Jiafeng Guo, Yixing Fan, Xiang Ji and Xueqi Cheng. 2019. Match-Zoo: A Learning, Practicing, and Developing System for Neural Text Matching. In Proceedings of the 42nd Int'l ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'19), July 21–25, 2019, Paris, France. ACM, NY, NY, USA
[12] Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In ACL. Association for Computational Linguistics.
[13] Andrew M Dai and Quoc V Le. 2015. Semi-supervised sequence learning. In Advances in neural information processing systems, pages 3079–3087.
[14] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018a . Deep contextualized word representations. In NAACL .
[15] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. Technical report, OpenAI
[16] Adina Williams, Nikita Nangia, and Samuel R Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In NAACL.
[17] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In EMNLP. Association for Computational Linguistics.
[18] William B Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In Proceedings of the Third International Workshop on Paraphrasing (IWP2005).
[19] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 2383–2392.
[20] Erik F Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll - 2003 shared task: Language-independent named entity recognition. In CoNLL.
[21] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018a. Glue: A multi-task benchmark and analysis platform for natural language understanding. In Proceedings of the EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 353–355.
[22] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv: 1810.04805v2 [cs.CL] 24 May 2019.
[23] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In ICLR.
[24] Ankur P Parikh, Oscar Tackstrom, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In EMNLP.
[25] Helia Hashemi, Mohammad Aliannejadi, Hamed Zamani, and W. Bruce Croft, ANTIQUE: A Non-Factoid Question Answering Benchmark. ArXiv:1905.08957v2 [cs.IR] 19 Aug 2019.

Authors Profile

Hay Man Oo, is a Ph.D candidate in Natural Language Processing Lab at University of Computer Studies, Yangon (UCSY) and a lecturer at Faculty of Information Science, UCSY, Myanmar. She got her B.C.Sc (Hons) in 2006, followed by M.C.Sc in 2010, respectively. Her current doctoral thesis research focuses on Information Retrieval of Myanmar Language. She is interested in the research area of Natural Language Processing (NLP), and Deep Learning.

Win Pa Pa, is a professor at Natural Language Processing Lab, Faculty of Computer Science, University of Computer Studies, Yangon. She received Ph.D.(IT) in 2009 from University of Computer Studies, Yangon, Myanmar. She is working at Natural Language Processing and Speech Processing research, supervising Master and PhD students.

Aye Mya Hlaing, is an associate professor at Natural Language Processing Lab, Faculty of Computer Science, University of Computer Studies, Yangon. She received Ph.D.(IT) in 2020 from University of Computer Studies, Yangon, Myanmar. She is working at Natural Language Processing and Speech Processing research, supervising Master and PhD students.