Sriwijaya Journal of Informatic and Applications
Vol. 5, No. 2, October 2024, pp. 75-81
ISSN 2807-2391

Generating Indonesian Poem: A Fine-Tunning Approach Using Pretrained GPT-2 Models

Arya Mulya Kusuma [a,1], Abdiansah Abdiansah [a,2,*]

[a] Department of Informatics Engineering, Faculty of Computer Science, Universitas Sriwijaya, Indonesia
[1] aryakusuma832@gmail.com*; [2] abdiansah@unsri.ac.id
* corresponding author

ARTICLE INFO

Article history
Received 28 Feb 2024
Revised 24 July 2024
Accepted 29 August 2024

Keywords
NLP
Text Generation
Indonesian Poem
Fine Tuning
GPT-2

ABSTRACT

In recent years, text generation has become an important subfield within Natural Language Processing (NLP), gaining significant attention and focus. Over the past decade, text generation technology has expanded significantly, reaching diverse application domains, especially in creative areas such as poem. Generating poetic content is a unique challenge that requires combining linguistic knowledge, creativity, and originality to craft each poem. This study focuses on developing a text generator for Indonesian language poem, using fine-tuning methodology with the pre-trained GPT-2 model from the Flax community. The study conducted a comparative analysis, benchmarking the performance of the researcher's model against a baseline model developed by Muhammad Agung Hambali. The evaluation outcomes showed the researcher's model outperformed the baseline model, exhibiting a 73.68% improvement in perplexity value. Furthermore, the study conducted a survey involving 62 respondents to determine the reception of the generated poem. The results indicated the poem produced by the research model was marginally superior to that of the baseline model.

1. Introduction

Text generation is an important subfield in natural language processing (NLP) that has been widely applied in various applications, including creative poem text generation [1]. Poem text generation combines linguistic knowledge models, creativity, and originality, making the development of poem text generation a challenge in creativity computational linguistics over the past five decades [2]. This field is beneficial for the exploration of computational algorithms and has positive impacts in the entertainment, advertising, and education sectors [3].

Poetic text generation can be achieved by fine-tuning pre-trained models, such as GPT-2, based on transformer models [4]. GPT-2, as a leading model in text generation, has been successfully implemented to create creative poem in various languages, including Chinese, English, and Arabic [5], [6], [7]. This process involves decoders of transformer model blocks capable of generating text without specialized supervised training [8].

In the context of Indonesian, Siallagan and Alfina showed that GPT-2 performed better than SeqGan in the pantun generation, with significant differences in Structure Accuracy, Rhyme Correctness, and vocabulary [9]. Another study by Muhammad Agung Hambali[1] using GPT-2 in Indonesian from the flax community[2] achieved a perplexity of 29.4884, with a train loss of 3.104 and eval loss of 3.384, proving the successful utilization of pre-trained models for poem text generation.

[1] https://huggingface.co/ayameRushia/gpt2-medium-fine-tuning-indonesia-poem
[2] https://huggingface.co/flax-community/gpt2-medium-indonesian

2. Related Works

Relevant research comes from various scientific sources, especially research journals, and proceedings, which are the main foundation and references to support researchers in answering the problems at hand and provide additional strength to the research being conducted.

In a study by Siallagan and Alfina, they compared two generative models, SeqGAN and GPT-2, in producing Indonesian pantun poem. The experimental results showed the superiority of GPT-2 in pantun structure, rhyme, and completeness of words, although both had difficulties in following the rhyme pattern of the pantun. This research encourages a focus on improving the accuracy of rhyme patterns, evaluating the relatedness and meaning of sentences in pantun, and involving human judgment to improve the linguistic quality of the resulting poems [9].

Liao introduced a new method for generating classical Chinese poem by using pre-trained models, especially GPT, which is considered simpler and more effective than previous approaches [5]. Another study by Beheitt and Ben Haj Hmida used the pre-training method and refinement of the GPT-2 model to automatically generate Arabic poems, showing that the model can generate poems with good quality, assessed through BLEU scores and human judgment [7].

In another language context, Astigarraga introduced an automated system for generating poem in Basque using the Markov Chain. Although the quality of the results is still below that of human composers, evaluations by experts in Bertsolaritza showed a positive impression of the system [10]. Finally, Muhammad Agung Hambali also contributed with his research using the pre-trained GPT-2 model in Indonesian, producing impressive perplexity evaluation results and becoming a key reference for further research.

3. Material and Method

a. Research Phases

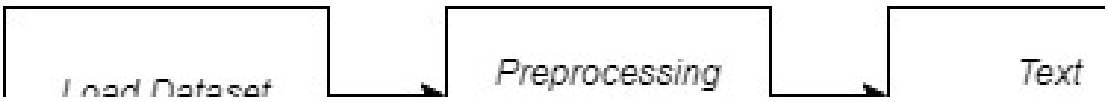

Fig. 1. Flowchart of Research Phases

The development of Indonesian poem text generation software through fine-tuning with the GPT-2 pre-trained model involved several phases. The initial phase includes the pre-processing of the dataset, which involves data cleaning, case folding, tokenization, and removal of author name information. Next, at the text representation phase, the poem text is converted into integer value representations for easy processing by the transformer model. The next major stage is the fine-tuning of GPT-2, where the model is adapted to the poem dataset through key parameter adjustments, requiring training iterations to gradually develop the model's capabilities. After the fine-tuning process, the final step involves model evaluation and analysis, where the model's performance and capabilities are evaluated to ensure the poem results meet the expected aesthetic and quality standards and are relevant to users.

b. Dataset

The type of dataset used in this research is a secondary dataset in the form of tables with Comma Separated Values (*.csv) format obtained from Kaggle[3]. The data is a collection of poems totaling 7,221 lines consisting of several columns, namely poem text, poem title, author, and a combination of poem text and poem title. However, in this research, only the poem text column will be taken and used as a training and validation dataset. The dataset division is done with a ratio of 8 to 2.

[3] https://www.kaggle.com/datasets/ilhamfp31/puisi-indonesia

c. Generative Pre-trained Transformer 2 (GPT-2)

GPT-2 is a language model from OpenAI that can be freely used without restrictions. It has special capabilities in capturing the relationships between words, so it can provide better results in various natural language processing tasks compared to publicly available LLM models [8].

In Figure 2, this study implements a medium-sized Indonesian GPT-2 pre-trained model. The model generally consists of 24 decoder layers. In each layer, there are 12 independent attention heads passed on to 144 different attention patterns. This attention-based model makes GPT-2 suitable for managing long texts [9].

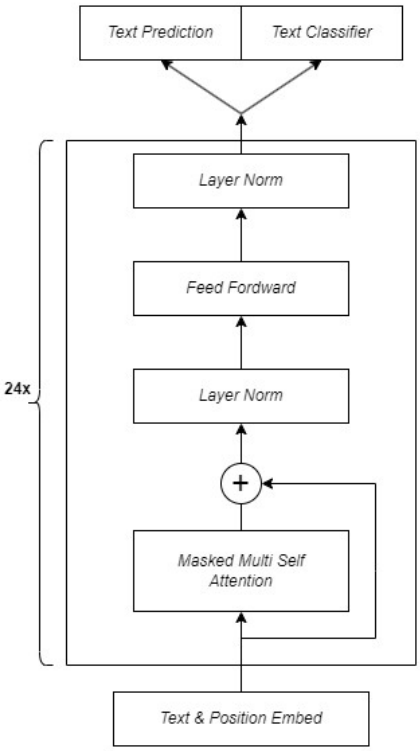

Fig. 2. GPT-2 Architecture

The process of generating a text involves log-likelihood loss reduction using the formula seen in Equation 1. In formula (1), θ refers to the model parameters, and x_i is the input token used [9].

$$P(\theta) = -\sum_i^l \log P(x_i | x_1, \dots, x_{i-1}) \quad (1)$$

After the training process, the model will generate the poem text by applying Top-K and Nucleus Sampling (Top-p) sampling methods. Top-K will take the k words with the highest probability. Meanwhile, as described in Equation 2, nucleus sampling takes the smallest combination of sequences by ensuring that the number V ≥ p [9].

$$p = -\sum_{x \in V(p)} P(x_{1:i-1}) \geq p. \quad (2)$$

This study used Flax Community's GPT-2 model, which is the same as the model in the reference (baseline) study, to ensure an "apples to apples" comparison. The choice of a similar

model aims to eliminate additional variables, allowing for a more in-depth evaluation of the effects of the selected parameters.

d. Fine-Tuning

Pretrained models are initially trained with large, purpose-specific data sets, allowing them to learn important features and patterns. Fine-tuning, the process of adjusting a pre-trained network for a new task and utilizing existing knowledge becomes a more efficient solution than training from scratch [11]. A common practice in fine-tuning involves copying all layers of the previous model except the last layer, replaced with new layers according to the number of classes in the new target domain [12]. In GPT-2, initially, the layers retrieve features to generate general text; then through fine-tuning some layers or only the last layer, the model adapts to generate poem text specifically.

e. Likert Scale

Data collection in survey research is a critical stage that requires data collection instruments, such as questionnaires, to collect information from respondents [13],[14]. The questionnaire is used by submitting structured written questions to respondents about their responses to the variables under study [15]. This data collection method is considered effective for obtaining the data needed for further analysis.

The Likert scale is a measurement method used to assess the perceptions, attitudes, or opinions of individuals or groups regarding an event [16]. This scale has positive and negative questions with different scores, where positive questions are scored 5 to 1, while negative questions are scored 1 to 5. With the Likert scale, researchers can describe the level of pro or con respondents to a concept or statement.

According to Naibaho, the steps in using a Likert scale involve collecting relevant questions, presenting them to groups of respondents, scoring by respondents, calculating the total score, identifying the highest and lowest scores, and calculating the Percentage Index for a more measurable interpretation of the results [17].

4. Results and Discussion

In this study, various variations of hyperparameters were experimented with to consider the best hyperparameters. The parameters include batch size and number of epochs, while the sequence length and learning rate were fixed from the beginning at 128 and $1e^{-4}$. The sequence length was chosen based on the average length of poems in the dataset, while the learning rate was optimized with the lr_find() method. The batch sizes tested were 4, 8, and 16, with the number of epochs ranging from 10 to 15. From these experiments, the best hyperparameter combinations were selected, as shown in Table 1.

Table 1. Hyperparameter Configuration

Configuration	Number
Epoch	15
Batch Size	16
Sequence Length	128
Learning Rate	$1e^{-4}$

After fine-tuning the GPT-2 model with the hyperparameter combination, the following evaluation results were obtained: training loss of about 2.052, validation loss of about 2.048, and perplexity of about 7.756, which can be seen in detail in the accuracy graph in Figure 3.

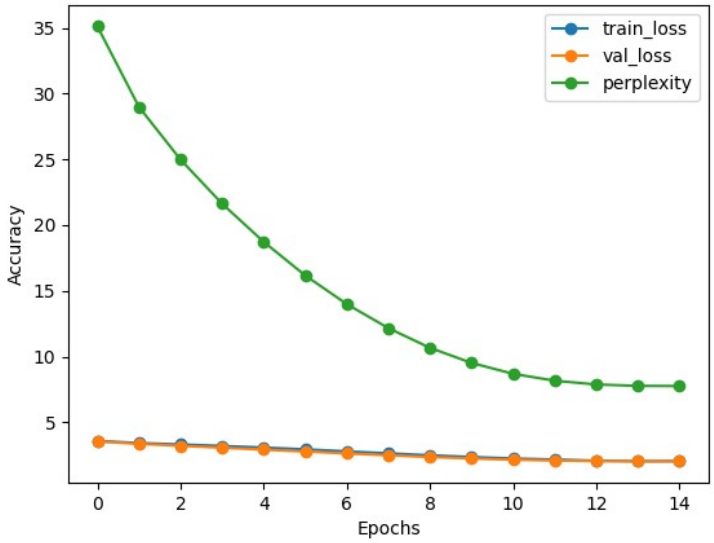

Fig. 3. Model Accuracy Graph

Graphical analysis of the decrease in train loss, validation loss, and perplexity indicates the ability of the model to understand and generalize the data effectively. This performance improvement provides confidence that fine-tuning was performed with positive results and the model has the potential to produce superior poems as expected. This research was also successful in improving the evaluative performance of the model. The model developed by the researcher successfully outperformed the reference model (baseline) in terms of evaluation scores, including train loss, valid loss, and perplexity. This fact reflects that the research effort to optimize the hyperparameter configuration had a significant positive impact on the model's performance, even surpassing the baseline model used as a benchmark. The table 2 compares the evaluation values between the research model and the baseline model.

Table 2. Comparison of Researcher & Baseline Model Evaluation Values

Evaluation Metrics	Researcher Model	Baseline Model
Train Loss	2,052	3,104
Valid Loss	2,048	3,384
Perplexity	7,759	29,488

While these metrics provide a solid statistical picture, qualitative aspects also play an important role. Therefore, a qualitative assessment of the text output needs to be done carefully to ensure that not only the statistics are satisfactory but also that the poem text output reaches an optimal level and is in line with the development goals of the model. This holistic evaluation provides the foundation for understanding the true impact of fine-tuning on the model's ability to create quality poem.

The next step is to use the model to generate poem texts. This process involves a comparison with the poem text of the baseline model, which belongs to Muhammad Agung Hambali. The main purpose of this comparison was to evaluate the quality of the poems generated by the researcher's model compared to the baseline model. The score comparison chart can be seen in the figure 4.

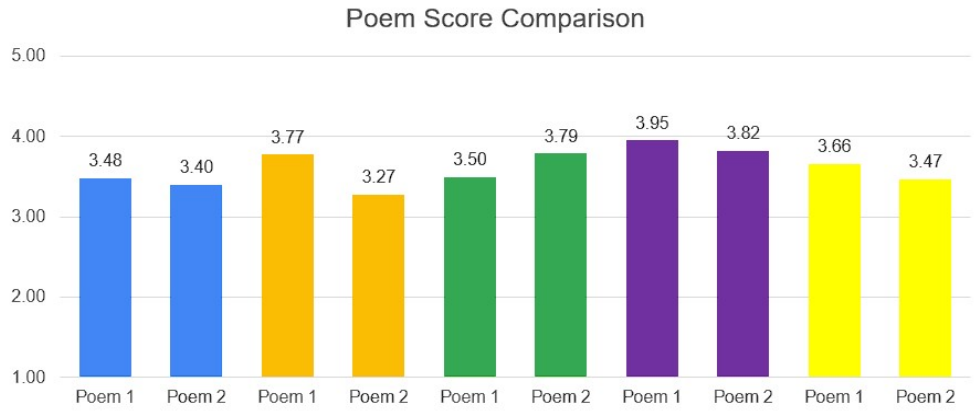

Fig. 4. Poem Score Comparison Graph

Description:
- Poem 1: Researcher's Poem
- Poem 2: Baseline Poem
- One visual color represents one comparison

In the results of the poem comparison survey, the participation of 62 respondents provided a strong basis for evaluation. A total of 4 out of 5 poems produced by the research model scored slightly higher than those of the baseline model. Although the scores were higher, the difference tended to be slight, indicating that statistically, the difference in quality between the two may not be that significant.

The poem score comparison graph illustrates that all poems, whether from the research model or the baseline, fell within the 3.0 to 3.9 range. This indicates the competency of both models in creating reasonably satisfactory poems, with respondents expressing preferences aligned with comparable levels of satisfaction. However, a distinctive advantage for baseline poems emerged in the third comparison, particularly in certain aspects. Elements such as word diversity, poem structure, and language style influenced the scoring when juxtaposing the researcher's poems with the baseline. The researcher's poems tended to showcase more creativity in terms of word variation and structure, while the baseline poems leaned toward direct narrative descriptions. The reader's decision to favor a more creative or simpler poem may affect the scoring of both poems.

5. Conclusion

The test results show that the evaluation of the model developed by the researcher successfully outperformed the reference model (baseline) in terms of evaluation values, such as train loss, valid loss, and perplexity. This finding reflects that the research effort to optimize the hyperparameter configuration significantly impacts model performance, even surpassing the reference model. This success can be used to conclude that fine-tuning the GPT-2 model with optimized parameters can improve the model's ability to generate poem texts.

Through the survey results with the participation of 62 respondents, the poems produced by the researcher's model scored slightly superior to those of the baseline model. Although the difference is slight, the comparison graph shows that all poems are within the score range of 3.0 - 3.9, indicating that both models can produce satisfactory poems.

References

[1] J. Li, T. Tang, W. X. Zhao, J.-Y. Nie, and J.-R. Wen, "Pre-trained language models for text generation: A survey," *ACM Comput. Surv.*, vol. 56, no. 9, pp. 1–39, 2024.
[2] S. Colton, J. Goodwin, and T. Veale, "Full-FACE Poetry Generation.," in *ICCC*, 2012, pp. 95–102.
[3] H. Chen, X. Yi, M. Sun, W. Li, C. Yang, and Z. Guo, "Sentiment-Controllable Chinese Poetry Generation.," in *IJCAI*, 2019, pp. 4925–4931.
[4] A. Vaswani, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, 2017.
[5] Y. Liao, Y. Wang, Q. Liu, and X. Jiang, "Gpt-based generation for classical chinese poetry," *arXiv Prepr. arXiv1907.00151*, 2019.
[6] J. Wang, X. Zhang, Y. Zhou, C. Suh, and C. Rudin, "There once was a really bad poet, it was automated but you didn't know it," *Trans. Assoc. Comput. Linguist.*, vol. 9, pp. 605–620, 2021.
[7] M. E. G. Beheitt and M. B. H. Hmida, "Automatic Arabic Poem Generation with GPT-2.," in *ICAART (2)*, 2022, pp. 366–374.
[8] A. Radford *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
[9] E. A. Siallagan and I. Alfina, "Poetry Generation for Indonesian Pantun: Comparison Between SeqGAN and GPT-2," *J. Ilmu Komput. dan Inf.*, vol. 16, no. 1, pp. 59–67, 2023.
[10] A. Astigarraga, J. M. Martínez-Otzeta, I. Rodriguez, B. Sierra, and E. Lazkano, "Markov text generator for basque poetry," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2017, vol. 10415 LNAI. doi: 10.1007/978-3-319-64206-2_26.
[11] N. Tajbakhsh *et al.*, "Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?," *IEEE Trans. Med. Imaging*, vol. 35, no. 5, 2016, doi: 10.1109/TMI.2016.2535302.
[12] E. Cetinic, T. Lipic, and S. Grgic, "Fine-tuning Convolutional Neural Networks for fine art classification," *Expert Syst. Appl.*, vol. 114, 2018, doi: 10.1016/j.eswa.2018.07.026.
[13] Subandi, D. Anubhakti, and B. Vallendito, "Rancang Bangun Kuesioner Survey Berbasis Web," *Semin. Nas. Teknol. Inf. dan Apl.*, vol. 9, 2017.
[14] I. Ismail and F. P. AlBahri, "Perancangan E-Kuisioner menggunakan CodeIgniter dan React-Js sebagai Tools Pendukung Penelitian," *J-SAKTI (Jurnal Sains Komput. dan Inform.*, vol. 3, no. 2, 2019, doi: 10.30645/j-sakti.v3i2.152.
[15] M. Muchlis, A. Christian, and M. P. Sari, "Kuesioner Online Sebagai Media Feedback Terhadap Pelayanan Akademik pada STMIK Prabumulih," *Eksplora Inform.*, vol. 8, no. 2, 2019, doi: 10.30864/eksplora.v8i2.215.
[16] S. Bahrun, S. Alifah, and S. Mulyono, "Rancang Bangun Sistem Informasi Survey Pemasaran dan Penjualan Berbasis Object Oriented Programming," *TRANSISTOR Elektro dan Inform.*, vol. 2, no. 2, 2018.
[17] F. Rio Naibaho, "Sistem Pendukung Keputusan Dalam Penentuan Dosen Terbaik Di IAKN Tarutung Dengan Menggunakan Kombinasi Metode Likert dan Metode VIKOR," *Semin. Nas. Sains Teknol. Inf.* , 2019.