AUTOMATIC INDONESIAN POETRY GEENERATOR BASED ON GPT-2

Barly Joshua Djaja, Abba Suganda Girsang

Abstract— **This research discussed about how to utilize generative pretrained transformer-2 (GPT-2) to automatically generate Indonesian poetry. The base model of GPT-2 will be finetuned for the task of generating Indonesian poetry. This model uses a transformer type of neural network. GPT-2 is a type of generative AI model based on transformer mechanism. The finetuned GPT-2 will be trained using a dataset containing list of Indonesian poetry that contains a title, author, and the content of poetry. The model will accept title as an input and will generate a new poetry with similar format as the training dataset. This research was done to know how to utilize GPT-2 to generate Indonesian poetry and evaluate the perplexity and ROUGE-1 score. Experimental results show that on average, finetuned GPT-2 model has a perplexity score of 1.18 which is around 20.27% higher than encoder-decoder RNN model and ROUGE-1 score of 0.13 which is around 44.44% increase in comparison to the base model of GPT-2.**

Index Terms— **Poetry Generation, GPT-2, Transformer, Generative AI, Indonesian Poetry**

IJSER

I. INTRODUCTION

THE development of artificial intelligence (AI) is very rapid. In the last 5 years alone, AI development has been very fast [1]. In 2017, the world was shocked by the introduction of the concept of transformers and attention [2] with a journal entitled "attention is all you need". This journal is very revolutionary in the development of generative AI, whose applications are currently very widely used, and a lot of new research has been developed based on transformers and attention mechanisms.

Generative AI [3], one of the AI methods, can create text, images and other media based on trained models. Generative AI models learn from patterns and structures from input training data and can generate new data that is similar to the training data and has similar patterns, structures and characteristics. In early 2020, the development of deep neural networks based on transformers was very rapid. This development allows generative AI to accept human language as an input prompt. This allows chatbot models like ChatGPT, Bing Chat, Bard, & LLaMA, etc. to look like real humans. Generative AI is also currently being used in various applications and industries such as art, literature, software development, product design, health, finance, games, marketing and fashion [4][5][6][7][8].

There has not been much research on automatic generation of Indonesian poetry using GPT-2 based machine learning [9]. Seeing the potential for the development of machine learning and combining it with art and literature, makes this topic interesting and can contribute by researching further into this field using finetuned Generative Pretrained Transformers-2 (GPT-2) [26] technology where GPT-2 is one of the state-of-the-art technologies the-art [10] in the field of Natural Language Processing, especially in terms of Generative AI. Therefore, in this research, an automatic Indonesian poetry generator will be created using machine learning based on GPT-2. GPT-2 will undergo a finetuning process with publicly available datasets taken from Kaggle.

Many previous studies aimed to combine art, literature, and also machine learning, especially in terms of automatically generating Indonesian poetry. Several previous methods proposed for generating Indonesian language poetry used a constraint-satisfaction approach [9] The constraints made in the constraint-satisfaction approach are features, namely the number of lines, number of words, and rhyme. This system creates poems from a collection of templates combined with a series of words. The result of this combination is a poem that is composed and becomes a poem. If you look at it from an evaluation perspective, it is still quite subjective because a survey was conducted on 180 respondents using the Turing test method without sufficient quantitative values.

There have been several attemps [11][12] who have tried to create an automatic Indonesian poetry generator using GPT-2 based machine learning but it has not been written well in journals and the model created has not been validated for accuracy and has not been compared with other state-of-the-art models such as encoders. RNN or GPT-2 decoder before finetuning. Because of the things above, further research was carried out regarding the automatic creation of Indonesian poetry using GPT-2 based machine learning which has been finetuned from a publicly available poetry dataset taken from Kaggle [13].

This research intends to contribute to researching how generative pretrained text (GPT-2) can be finetuned to produce a high-quality Indonesian poetry. The proposed model will accept a title as an input and will produce the generated poetry with similar format as the trained dataset.

The formulation of the problem of this research is by providing input of the title of the poem as a prompt, what are the results of automatically creating Indonesian poetry using GPT-2 based machine learning which has undergone a finetuning process when compared with other state-of-the-art models, namely the RNN encoder-decoder and GPT-2 before finetuning.

Manuscript received August 31, 2020; revised January 4, 2021. This work was supported in part by Bina Nusantara Graduate program.

Barly Joshua Djaja is a master student at Computer Science Department, BINUS Graduate Program-Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480 (email: barly.joshua@binus.ac.id)

Abba Suganda Girsang is a lecturer at Computer Science Department, BINUS Graduate Program-Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480 (corresponding author, e-mail: agirsang@binus.edu).

The evaluation method used in this research is by comparing the value of perplexity [14] and ROUGE [15] with another state-of-the-art model such as RNN [25] and base model of GPT-2

II. Related Works

Text generation has become a popular research topic in the past few years. Xing Zhang et al. in 2014 proposed a Chinese poetry generation using Recurrent Neural Networks [16] in which they argued is ideally suited to capture poetic content and form. The generator jointly performs content selection ("what to say") and surface realization ("how to say") by learning representations of individual characters, and their combinations into one or more lines as well as how these mutually reinforce and constrain each other. Poem lines are generated incrementally by considering the entire history of what has been generated so far rather than the limited horizon imposed by the previous line or lexical n-grams. Experimental results show that the model outperforms competitive Chinese poetry generation systems using both automatic and manual evaluation methods.

Yi Liao et al. in 2019 proposed effective method for generating high quality classical Chinese poetry with Generative Pre-trained Language Model (GPT) [17]. The method adopts a simple GPT model, without using any human crafted rules or features, or designing any additional neural components. While the proposed model learns to generate various forms of classical Chinese poems. Yi Liao et al. claimed that the generated poems are of high quality. Yi Liao et al. also proposed and implement a method to fine-tune the model to generate acrostic poetry. To the best of their knowledge, this is the first to employ GPT in developing a poetry generation system. They also have released an online mini demonstration program on Wechat to show the generation capability of the proposed method for classical Chinese poetry.

Zhiqiang Liu et al in 2019 proposed a Rhetorically Controlled Encoder-Decoder for Modern Chinese Poetry Generation [18]. According to Zhiqiang Liu et al. the model relies on a continuous latent variable as a rhetoric controller to capture various rhetorical patterns in an encoder, and then incorporates rhetoric-based mixtures while generating modern Chinese poetry. For metaphor and personification, an automated evaluation shows that our model outperforms state-of-the-art baselines by a substantial margin, while a human evaluation shows that our model generates better poems than baseline methods in terms of fluency, coherence, meaningfulness, and rhetorical aesthetics.

Fam Rashel et al. in 2014 proposed a constraint satisfaction-based generator of topical Indonesian poetry [9]. It scans popular news websites for articles and extracts relevant keywords that are combined with various language resources such as templates and other slot fillers into lines of poetry. It then composes poems from these lines by satisfying a set of given constraints. A Turing Test-style evaluation and a detailed evaluation of three different configurations of the system was conducted through an online questionnaire with 180 respondents. The results showed that under the best scenario, 57% of the respondents thought that the generated poems were authored by humans, and that poems generated using the full set of constraints consistently measured better on all aspects than those generated using the other two configurations. The system is now available online as a web application.

Santillan Marvin et all. in 2020 proposed a Poem Generation using Transformers and Doc2Vec Embeddings [19]. Santillan Marvin et all. propose a method of generating poems using transformers, coupled with doc2vec embeddings in order to assess the automatically generated poems. In this method, we first train a transformer and a doc2vec model using a poem dataset. Then the trained transformer takes an input text and produces several generated poems. To have an objective basis for assessing the generated poems, we present a preliminary attempt at measuring the quality of a machine-generated poem by computing the cosine similarity score, referenced to the trained doc2vec model. This score is used as a basis for choosing the final output poem. The results show that this method ensures good cohesion between the machine-generated poem and the given input text. We then also explore the implication of the transformer training to the doc2vec embeddings of its output poems, which are shown to be more like poems (documents) in the train set as training progresses. Finally, we demonstrate how transformers can learn some poetry styles by exposing them to poems of specific poets.

Mika Hamalainen et al. in 2022 proposed a Modern French Poetry Generation with RoBERTa and GPT-2 [20]. It is a novel neural model for modern poetry generation in French. The model consists of two pretrained neural models that are fine-tuned for the poem generation task. The encoder of the model is a RoBERTa based one while the decoder is based on GPT-2. This way the model can benefit from the superior natural language understanding performance of RoBERTa and the good natural language generation performance of GPT-2. The evaluation shows that the model can create French poetry successfully. On a 5-point scale, the lowest score of 3.57 was given by human judges to typicality and emotionality of the output poetry while the best score of 3.79 was given to understandability.

Jianli Zhao et al. in 2022 proposed an Automatic Generation and Evaluation of Chinese Classical Poetry with Attention-Based Deep Neural Network [21]. They present a novel Transformer-XL based on a classical Chinese poetry model that employs a multi-head self-attention mechanism to capture the deeper multiple relationships among Chinese characters. Furthermore, they e utilized the segment-level recurrence mechanism to learn longer-term dependency and overcome the context fragmentation problem. To automatically assess the quality of the generated poems, they also built a novel automatic evaluation model that contains a BERT-based module for checking the fluency of sentences and a tone-checker module to evaluate the tone pattern of poems. The poems generated using our model obtained an average score of 9.7 for fluency and 10.0 for tone pattern. Moreover, they visualized the attention mechanism, and it showed that their model learned the tone-pattern rules. All experiment results demonstrate that our poetry generation model can generate high-quality poems.

IJSER

III. Proposed Method

In this study, the researchers proposed a finetuned model of GPT-2 to generate Indonesian poetry. The proposed method of generating Indonesian poetry generation consists of some steps. The number of poetries used as many as 7223 poetries in Indonesian language from Kaggle dataset provided by IlhamFP [13]. Figure 1 shows an overview of how the flow of this research was conducted.

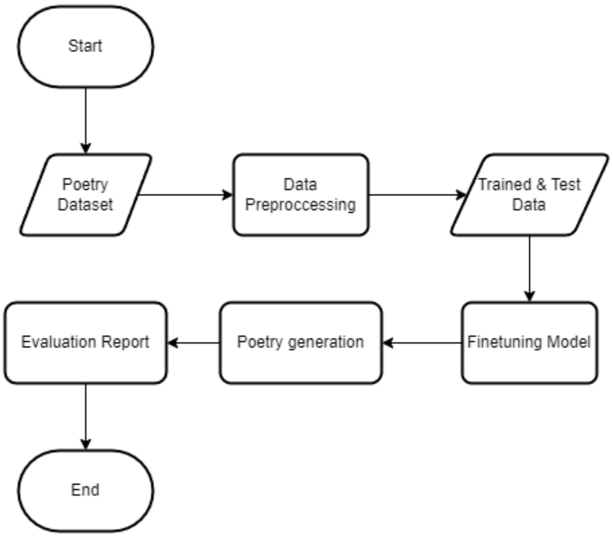

Figure 1. Research Method

A. Data Collecting

The dataset of Indonesian poetry is retrieved from IndoSum dataset. Kaggle is a corpus dataset for various dataset. 7300 poetries were crawled from various websites and compiled. Each poetry in the Kaggle dataset is supplied with title, author, puisi (poetry), and puisi_with_header (poetry_with_header). The Kaggle dataset is stored in the .csv (comma-separated values) file format, and no further conversion needs to be done since this file format is supported by pandas python library.

IJS

B. Data Preprocessing

In the first stage, preprocessing, the dataset available to the public is still in raw form, so additional processes need to be carried out to modify the existing dataset into data that is ready to be used for finetuning the GPT-2 model. The poetry dataset used consists of 7223 unique poems on Kaggle obtained [13] and is available for public consumption and will go through the preprocessing stage first. The purpose of preprocessing on this dataset is to separate the dataset into two parts, namely train dataset and test dataset, and each poem will be inserted, namely the string <BOS> at the beginning of the poem and <EOS> at the end of the poem. The purpose of inserting at the beginning and end of a poem is to function as a flag (marker) that the poem has started or ended. Figure 2 illustrates the flow of the proposed preprocessing process.

Figure 2 explains of the flow of how preprocessing is carried out. In this flow there are several stages that need to be passed so that the preprocessing process runs well, namely splitting dataset to train & test which is 90% and 10% respectively and inserting special token <BOS> and <EOS> on every dataset.

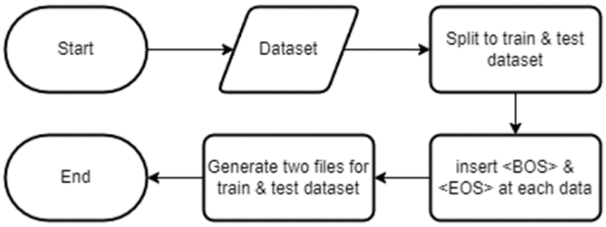
Start
Dataset
Split to train & test dataset
insert <BOS> & <EOS> at each data
Generate two files for train & test dataset
End

Figure 2. Preprocessing Flo

C. Model Definition

The proposed method is using gpt-2 and finetuning it with poetry dataset. The reason for using gpt-2 is because it is the state-of-the art model and is open-source therefore the code can be modified openly. GPT-2 stands for "Generative Pretrained Transformer 2," and it is a variant of the Transformer model architecture developed by OpenAI. The GPT-2 model is known for its capability to generate human-like text and has been pretrained on a large corpus of text from the internet. It has a deep architecture with a large number of parameters, making it suitable for various natural language processing tasks, including text generation.

The architecture used is attention layer. In the context of deep learning and natural language processing (NLP) is a key component for focusing on specific parts of the input sequence when making predictions or encoding information.

Input and output shapes, the input expected to be poetry title and output expected to be the generated poetry.

D. Fine tuning model

Fine-tuning a language model like GPT-2 involves training the model on a specific dataset or task to adapt it for that task. In the context of poetry generation, fine-tuning means training the GPT-2 model on a dataset of poetry to make it better at generating coherent and contextually relevant poems.

The finetuning process includes 3 main processes, splitting the dataset into training and evaluation, training the model to mimic the preprocessed dataset, and finally generate the poetry with similar style given in the dataset.

The first step is to split the dataset into 2 parts, train and evaluation data. The test size of the dataset is 0.1 or a10% of the total dataset. The random state given in this experiment is 31. While splitting the dataset, we also add <BOS> which acts as a flag for beginning of sentences (BOS) and we add <EOS> which acts as a flag for end of sentences (EOS). This technique is implemented so that it can be known where to start the poem, and where to end the poem.

The second step is for training and finetuning the existing model. For this step, transformers library from hugging face was used and accelerate library to fasten the training process. This part consists of several functions to train the data properly. First the train and eval were loaded and set a tokenizer from the pretrained model. In this case pretrained model by Wirawan [22] called gpt2-small-indonesian-522M was used. Data collator was also used in this experiment. Trained data also used a special token for <BOS> and <EOS>. The training process took approximately 2 hours of computing time, and the model was generated successfully.

The third and last step is generating the poem based on the finetuned trained model. Transformer library was also used to generate poems. This part also consists of several smaller steps, that is loading the finetuned model, loading the tokenizer, and generating the text. Model generation accepts several inputs such as id, max length, pad token id, top k, top p, and do sample. The generated poem will then be printed in the terminal. Figure 3 shows the flow of model finetuning.

IJSER

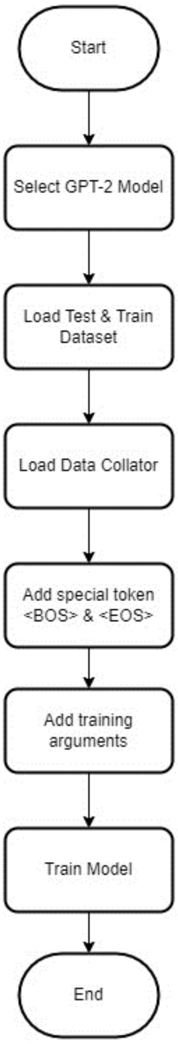

Figure 2. Flow of model finetuning

E. Poetry Generation

The last step is for generating the Indonesian poetry. With the model ready and the finetuning process has been carried out, the poetry generator is ready to be used. This final process is the stage of using a finetuned model to create poetry. By providing an input sequence, namely the title, the model will imitate the poetry pattern from the existing dataset. In the process of generating poetry, there are also several steps needed to create poetry from a predetermined model. Figure 3 shows the flow of the poetry generation process.

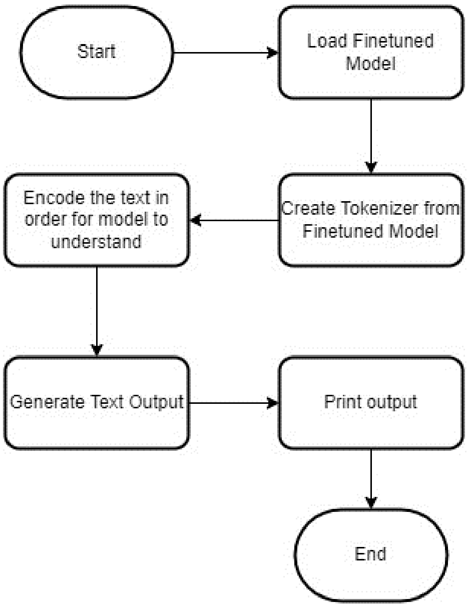
Start
Load Finetuned Model
Create Tokenizer from Finetuned Model
Encode the text in order for model to understand
Generate Text Output
Print output
End

Figure 3. Flow of Poetry Generation

F. Evaluation Method

The evaluation method used to evaluate the score of finetuned model is by using 2 methods, that is: Perplexity and *Recall-Oriented Understudy for Gisting Evaluation* (ROUGE). Perplexity (PPL) is one of the most common metrics used to evaluate the language model. It is defined as exponentiated average negative log-likelihood of a sequence, calculated with exponent base e. In simpler terms, PPL is how much a model is surprised by seeing new data. The lower the PPL amount, the better the training is. PPL can be calculated by exponent of the loss obtained from the model. The full formula for PPL can be seen in Eq. (1)

$$PPL(X) = exp\left\{-\frac{1}{t}\sum_{i}^{t}\log p_{\theta}(x_i|x_{<i})\right\} \qquad (1)$$

ROUGE evaluation model is a collection of metrics and software designed to evaluate automatic summarization but can also be used for machine translation. This metric compares the results created by the model to those generated by humans. In this research, we will use the value of ROUGE-1, namely the overlap of unigrams (each word) between the model output and the reference. ROUGE uses recall and precision to compare model output with human-generated references. It is calculated how many n-grams in the reference are in the model. The formulation for calculating ROUGE value can be seen in Eq. (2), Eq. (3), & Eq. (4)

$$Recall = \frac{Number\ of\ matching\ n-grams}{Number\ of\ n-grams\ in\ the\ reference} \qquad (2)$$

$$Precision = \frac{Number\ of\ matching\ n-grams}{Number\ of\ n-grams\ in\ the\ Candidate} \qquad (3)$$

$$ROUGE - 1 = 2 * \frac{Precission * Recall}{Precision+Recall} \qquad (4)$$

IV. Results and Discussion

A. System Specification

System specifications consist of hardware and software used to perform, develop, test and run modelsdevelop, test and run models. The hardware and software used in this research can be seen in Table I dan Table II.

TABLE I
HARDWARE USED IN THIS RESEARCH

Hardware	Specification
Operating system	Windows 11 Pro 64-bit (10.0, Build 22621)
Memory	32GB
Graphics card	AMD Radeon RX 6800XT
Processor	12th Gen Intel® Core ™ i7-12700KF (20 CPUs), 3.6GHz

TABLE II
SOFTWARE USED IN THIS RESEARCH

Software	Version
Python	3.9.13
Transformer	4.33.3
Pandas	2.1.1
Numpy	1.26.0

B. Finetuning results

In this research, the GPT-2 model is used for the finetuning process. The title of the poem is what will be used as input. The maximum number of text lengths is 200. The GPT-2 model automatically creates poetry and limits the number of text lengths to 200. The results of generating poetry from the GPT-2 model which has undergone a finetuning process can be seen in table 4.3. In Table 4.3, it can be seen that the results of the poems generated have a writing format that is similar to the dataset in the poetry_with_header column. This is the result of a finetuning process carried out so that the resulting poetry format imitates the writing format in the source dataset. Table III shows some examples of the generated poetry with given title.

As shown on Table III, the model is capable of generating a good quality Indonesian Poetry and follows the same exact formatting as the trained dataset. This shows that GPT-2 is a good model for generating Indonesian poetry as well as a good model for generative AI model since it is outputting an exact text formatting.

Table IV shows the learning rate and loss to the number of steps. In this experiment, the hyperparameters are epoch number of 10 and train batch size of 8. This resulted in total steps of 3290 and logging was done for every 500 steps.

IJSER

TABLE III

GENERATED POETRY IN BAHASA INDONESIA

Title	Generated Poetry
PENYESALAN YANG MENYAKITKAN	PENYESALAN YANG MENYAKITKAN Oleh Devi Bella Efendi Di pagi hari yang sunyi ini Angin berhembus dari balik atap rumahku Aku merenungi Terasa hampa bila aku tak mengingatmu Dan bertanya tentang diriku sendiri Aku tak akan pernah memberitahu tentang diriku Aku tak akan pernah lupa akan ucapanmu yang hanya sandiwara Aku tidak akan meningalkan aku ketika itu Aku hanya bisa menangis saat itu
RATAPKU MENATAP REMBULAN	RATAPKU MENATAP REMBULAN Oleh Dhymas Fajri Terpaku aku berdiri tak mampu ku lepas tanpa ku dapati aku merindu dan tak dapat ku lepas ketika kau terluka karenamu namun ku dapat kembalikan setelah kau pergi kabulkanlah semua harapanku agar aku dapat kembali hidupku penuh dengan harapan tak terkatupkah Engkau hanya menunggu ku hanya akan tersenyum saat bersamamu selamanya
PENANTIAN YANG PANJANG	PENANTIAN YANG PANJANG Oleh Agus Kurniawan Di setiap hari ku berjalan Di setiap hari ku menatap ke langit Dimana itu ada yg terasa berbeda Ada yang merasa berbeda Ada yang merasa sama Ada yang merasa tidak merasa sama Ada yang merasa dirinya tidak mau berbicara dengan mu Ada yang merasa sama Ada yang merasa merasa sendiri Ada yang merasa tidak merasa sama Ada yang merasa sama Ada yang merasa sama Ada yang merasa sama

TABLE IV

LEARNING RATE AND LOSS TO STEPS

Steps	Learning Rate	Loss
500	4.24e-05	3.756
1000	3.48e-05	3.0902
1500	2.72e-05	2.5348
2000	1.96e-05	2.1573
2500	1.20e-05	1.8227
3000	4.41e-06	1.6348

Table IV shows that the higher steps will result in lower learning rate and lower loss. The learning rate progress is linear with the increase of steps and the loss is log function of the number of steps. This can be seen in Figure 4 and Figure 5 respectively.

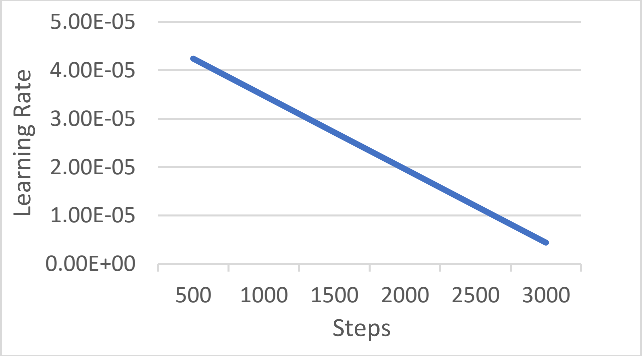

Figure 4. Learning Rate to Steps

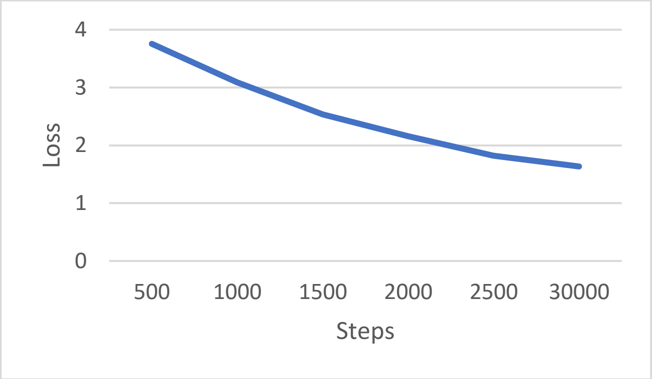

Figure 5. Model Loss to Steps

C. *Evaluation of model performance*

Evaluation of the proposed model will be carried out using perplexity and ROUGE metrics. The finetuning GPT-2 model will be compared with the RNN encoder-decoder architecture as well as the GPT-2 baseline to create automatic poetry.

RNN encoder-decoder architecture in automatic poetry generation, the code is also written and the model is trained using a bidirectional long-short term memory (BiLSTM) [23] layer and FastText [24] as word embedding. Due to device limitations as in Table I, the computer was unable to carry out training for the RNN with 7223 poetry datasets. Therefore only 4127 datasets were used. The training process with the RNN encoder-decoder architecture takes up to 8 days.

Based on the GPT-2 model, the poetry creation stage is directly carried out by the poetry generation process without

IJSER

any finetuning process. The GPT-2 base model used is the Indonesian version [22] which is given the name "Wirawan/gpt2-small-indonesian-522M". This is the GPT-2 base model used as a comparison for evaluation models. The results of the comparison of the evaluation of perplexity value metrics for titles that are not in the dataset can be seen in Table V. The results of the comparison of the ROUGE-1 value metric evaluation from the test dataset can be seen in Table VI

TABLE V
RESULTS OF PERFORMANCE COMPARISON OF PERPLEXITY MODEL VALUES WITH VARIOUS TITLES OUTSIDE THE TRAIN DATASET

Title	Perplexity		
	GPT-2 Finetuned	Encoder-Decoder RNN	Base Model GPT2
PENYESALAN YANG MENYAKITKAN	1.09	1.42	1.46
RATAPKU MENATAP REMBULAN	1.28	1.15	1.61
PENANTIAN YANG PANJANG	1.16	1.20	1.45
SELEMBAR PUISI UNTUK KEKASIH	1.08	1.35	1.38
ALASAN UNTUK BERPISAH	1.16	1.41	1.58
KEMARAHAN, KETIDAKBERDAYAAN DAN UANG	1.32	1.60	1.37
Mean	1.18	1.36	1.48

Table VI shows the ROUGE value of different models with generated poetry as prediction and human creation as reference

TABLE VI
RESULTS OF PERFORMANCE COMPARISON OF PERPLEXITY MODEL VALUES WITH VARIOUS TITLES OUTSIDE THE TRAIN DATASET

Title	ROUGE-1		
	GPT-2 Finetuned	Encoder-Decoder RNN	Base Model GPT2
PENYESALAN YANG MENYAKITKAN	0.196	0.288	0.110
RATAPKU MENATAP REMBULAN	0.196	0.218	0.054
PENANTIAN YANG PANJANG	0.142	0.183	0.070
SELEMBAR PUISI UNTUK KEKASIH	0.143	0.21	0.140
ALASAN UNTUK BERPISAH	0.10	0.182	0.078
KEMARAHAN, KETIDAKBERDAYAAN DAN UANG	0.148	0.122	0.090
Mean	0.13	0.20	0.09

In Table V it can be seen that the average perplexity value of the finetuned GPT-2 model is 1.18, the perplexity value of the RNN encoder-decoder model is 1.36, and the perplexity value of the GPT-2 base model is 1.48. In general, a lower perplexity value indicates that the model has a better ability to calculate the next word that may appear in a sentence sequence. Even in the context of writing poetry, the word order that appears can be very varied and does not follow a previously existing pattern. However, in general it can be seen in Table V. that the finetuned GPT-2 model has the lowest perplexity value among the other two comparison models. The perplexity value has an increase of 13.23% compared to the RNN encoder-decoder model and an increase in the perplexity value of 20.27% when compared to the GPT-2 base model.

In table VI it can be seen that the average ROUGE-1 value of the finetuned GPT-2 model is 0.13, the ROUGE-1 value of the RNN encoder-decoder model is 0.20, and the ROUGE-1 value of the GPT-2 base model is 0.09. The ROUGE-1 value ranges between 0-1, where the closer the value is to 1, the better the model output results are against the reference provided. It can be seen in table 4.7 that the RNN encoder-decoder model has the highest ROUGE-1 value, namely 0.20, this indicates that the RNN model has higher predictive ability than the GPT-2 base model and the finetuned GPT-2 model. However, the finetuned GPT-2 model has an increased ROUGE-1 value when compared to the base GPT-2 model. The increase in ROUGE-1 value from basemodel GPT-2 to finetuned GPT-2 is 44.44%. Of course, this is a significant improvement, although it still cannot compete with the ROUGE-1 value of the RNN encoder-decoder model. More research needs to be done on this.

In general, it can be seen that the examples of poetry generation results created by the three models, finetuned GPT-2 has the most consistent format with the poetry dataset provided. Every poem created by the finetuned GPT-2 model will always have a title, poet, and content, but the RNN encoder-decoder model and GPT-2 base model will not always follow this format. In the RNN encoder-decoder model, only 4 of the 6 sample poems generated have the poet's name. This indicates that finetuned GPT-2 follows the writing format of the training dataset better than the other two models. Of course, the GPT-2 base model will have difficulty following the training dataset format because no finetuning process has been carried out on the GPT-2 base model.

V. CONCLUSION AND FUTURE WORK

Evaluation of the proposed model, namely finetuning the GPT-2 base model for the needs of automatically generating Indonesian poetry, has been carried out. The finetuned GPT-2 model will accept the title of the poem as input and will generate the poem as output. To get a finetuned GPT-2 model, the base model of GPT-2 will be trained on poetry datasets that are available in open source [13]. There are 2 evaluation metrics carried out on this model, namely the perplexity value and the ROUGE-1 value. The proposed model will then be compared with 2 other models, namely the RNN encoder-decoder and the GPT-2 base model. For the RNN encoder-decoder model, the BiLSTM layer is used, and word embedding is FastText. For the GPT-2 base model, use the GPT-2 model which has been pretrained in Indonesian with the model’s name Wirawan/gpt2-small-

IJSER

indonesian-522M (Wirawan, 2020) which is available in the HuggingFace library.

Model training was carried out with 10 epochs, 8 batch sizes. This brings the total steps to 3290. Logging of training results is carried out every 500 steps. The results of the training model show several things, namely the step value will be inversely proportional to the learning rate and also loss. The learning rate value decreases linearly as steps increase and the loss value decreases logwise with the number of steps.

Evaluation of the model was carried out by comparing 2 other models, namely the RNN encoder-decoder and the GPT-2 base model. The experimental results show that the finetuned GPT-2 model has the best perplexity value at 1.18. This figure is around 13.23% better than the RNN encoder-decoder model and 20.27% better than the GPT-2 base model. The evaluation value of ROUGE-1 shows that the RNN encoder-decoder model has the highest number, namely 0.20 and is followed by finetuned GPT-2 which has a ROUGE-1 value of 0.13 and followed by the base model GPT-2 which has a ROUGE-1 value of 0.09. The ROUGE-1 value of finetuned GPT-2 has an increase of 44.44% when compared to the base model GPT-2. This dizziness is certainly significant. The finetuned GPT-2 model also shows a more consistent format because all the poetry examples generated have the same format as the dataset, namely title, poet's name, and poetry content. Other models are inconsistent in this case where the RNN encoder-decoder model only has the names of the authors of 4 of the 6 poems generated.

In the future, the researchers will conduct experiments with a wider scope of experiments for the classification of texts and other combinations of parameters from term weighting.

Based on the conclusions given, there are several suggestions that can be made for future research, namely improving the current dataset because the poetry dataset is less than optimal. Poetry can be divided based on genre to make it more specific and the resulting poetry has content that is more relevant to the title. In addition, combining with other transformer architectures such as BERT or RoBERTa can be another interesting option to improve model performance.

IJSER

REFERENCES

[1] Chui, Michael, et al. “The State of AI in 2023: Generative AI’s Breakout Year.” McKinsey & Company, McKinsey & Company, 1 Aug. 2023, www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year#/.

[2] Vaswani, Ashish, et al. “Attention Is All You Need.” arXiv.Org, arXiv, 2 Aug. 2023, arxiv.org/abs/1706.03762.

[3] Andrej, Karpathy, et al. Generative Models, OpenAI, 16 June 2016, openai.com/research/generative-models.

[4] Beheitt, Mohamed El, and Moez Ben Haj Hmida. “Automatic arabic poem generation with GPT-2.” Proceedings of the 14th International Conference on Agents and Artificial Intelligence, 2022, https://doi.org/10.5220/0010847100003116.

[5] Eapen, Tojin T., et al. “How Generative AI Can Augment Human Creativity.” Harvard Business Review, Harvard Business Review, 13 June 2023, hbr.org/2023/07/how-generative-ai-can-augment-human-creativity

[6] Economist. “The Race of the AI Labs Heats Up.” The Economist, The Economist Newspaper, 30 Jan. 2023, www.economist.com/business/2023/01/30/the-race-of-the-ai-labs-heats-up.

[7] Koullias, Theodora, et al. “Generative AI: Unlocking the Future of Fashion (Guest Blog by McKinsey QuantumBlack).” The UK’s Technology Trade Association, The UK’s technology trade association, 19 Apr. 2023, www.techuk.org/resource/aiweek2023-mckinsey-quantum-black-wed.html.

[8] Solaiman, Irene, et al. “Release Strategies and the Social Impacts of Language Models.” arXiv.Org, Cornell University, 13 Nov. 2019, arxiv.org/abs/1908.09203.

[9] Rashel, Fam, and Ruli Manurung. “Poetry Generation for Bahasa Indonesia Using a Constraint Satisfaction Approach.” Universitas Indonesia, 1 Jan. 2013, scholar.ui.ac.id/en/publications/poetry-generation-for-bahasa-indonesia-using-a-constraint-satisfa.

[10] Eapen, Tojin T., et al. “How Generative AI Can Augment Human Creativity.” Harvard Business Review, Harvard Business Review, 13 June 2023, hbr.org/2023/07/how-generative-ai-can-augment-human-creativity.

[11] Putra, Ilham. “Pembangkitan Puisi Otomatis.” Kaggle, Kaggle, 23 Dec. 2020, www.kaggle.com/code/ilhamfp31/pembangkitan-puisi-otomatis.

[12] Rushia, Ayame. “Ayamerushia/GPT2-Medium-Fine-Tuning-Indonesia-Poem · Hugging Face.” ayameRushia/Gpt2-Medium-Fine-Tuning-Indonesia-Poem · Hugging Face, 2021, huggingface.co/ayameRushia/gpt2-medium-fine-tuning-indonesia-poem.

[13] Putra, Ilham. “Puisi Indonesia.” Puisi Indonesia, 2020, Accessed 11 Nov. 2023.

[14] Jelinek, F, et al. “Perplexity—a Measure of the Difficulty of Speech Recognition Tasks.” Pubs.Aip.Org, AIP Publishing, Dec. 1977, pubs.aip.org/asa/jasa/article/62/S1/S63/642598/Perplexity-a-measure-of-the-difficulty-of-speech.

[15] Lin, Chin-Yew. “Rouge: A Package for Automatic Evaluation of Summaries.” ACL Anthology, July 2004, aclanthology.org/W04-1013/.

[16] Zhang, Xingxing, and Mirella Lapata. “Chinese poetry generation with Recurrent Neural Networks.” Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, https://doi.org/10.3115/v1/d14-1074.

[17] Liao, Yi, et al. “GPT-Based Generation for Classical Chinese Poetry.” arXiv.Org, 5 Sept. 2019, arxiv.org/abs/1907.00151.

[18] Liu, Zhiqiang, et al. “Rhetorically controlled encoder-decoder for modern Chinese poetry generation.” Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, https://doi.org/10.18653/v1/p19-1192.

[19] Santillan, Marvin C., and Arnulfo P. Azcarraga. “Poem Generation using transformers and doc2vec embeddings.” 2020 International Joint Conference on Neural Networks (IJCNN), 2020, https://doi.org/10.1109/ijcnn48605.2020.9207442.

[20] Hämäläinen, Mika, et al. “Modern French Poetry Generation with Roberta and GPT-2.” arXiv.Org, 6 Dec. 2022, arxiv.org/abs/2212.02911.

[21] Zhao, Jianli, and Hyo Jong Lee. “Automatic generation and evaluation of Chinese classical poetry with attention-based deep neural network.” Applied Sciences, vol. 12, no. 13, 2022, p. 6497, https://doi.org/10.3390/app12136497.

[22] Wirawan, Cahya. “Cahya/GPT2-Small-Indonesian-522m · Hugging Face.” Cahya/Gpt2-Small-Indonesian-522M · Hugging Face, 2020, huggingface.co/cahya/gpt2-small-indonesian-522M. [23] S. S. Ge, “An Improved TF-IDF Algorithm Based on Class Discriminative Strength for Text Categorization on Desensitized Data,” pp. 39–44, 2019.

[23] Aziz Sharfuddin, Abdullah, et al. “A deep recurrent neural network with BILSTM model for sentiment classification.” 2018 International Conference on Bangla Speech and Language Processing (ICBSLP), 2018, https://doi.org/10.1109/icbslp.2018.8554396.

[24] Bojanowski, Piotr, et al. “Enriching word vectors with subword information.” Transactions of the Association for Computational Linguistics, vol. 5, 2017, pp. 135–146, https://doi.org/10.1162/tacl_a_00051.

[25] Rumelhart, David E., et al. Learning Internal Representations by Error Propagation, 1985, https://doi.org/10.21236/ada164453.

[26] Radford, Alec, et al. "Language Models are Unsupervised Multitask Learners." OpenAI Blog, 14 Feb. 2019, openai.com/blog/better-language-models/.