International Journal of Speech Technology (2024) 27:1027–1037
https://doi.org/10.1007/s10772-024-10136-2

A transformer-based approach to Nigerian Pidgin text generation

Kabir Garba[1] · **Taiwo Kolajo**[1,2] · **Joshua B. Agbogun**[1]

Received: 15 February 2024 / Accepted: 25 August 2024 / Published online: 16 October 2024

Abstract

This paper describes the development of a transformer-based text generation model for Nigerian Pidgin also known as Naijá, a popular language in West Africa. Despite its wide use, Nigerian Pidgin remains under-resourced, particularly in areas related to text generation and natural language processing. These difficulties are primarily due to technological constraints rather than the language's fundamental attributes. There is currently a demand for Nigerian Pidgin-specific solutions because it is used in everyday communication and has a unique linguistic blend. This paper aims to close this gap by exploring the application of state-of-the-art transformer technology to develop a text generation model for Nigerian Pidgin. This work uses the public Afriberta-corpus dataset to optimize the Generative Pre-trained Transformer (GPT-2) model across a sizeable dataset. The performance evaluators, BLEU and Perplexity metrics provide a detailed breakdown of the model's text quality and predictive accuracy. Despite the difficulties caused by a limited amount of training data, preliminary evaluations show that the model can generate coherent Nigerian Pidgin text. The performance evaluation yielded perplexity scores of 43.56 for variable target reference length and 43.26 for fixed text length. BLEU scores of 0.15 for fixed max length and 0.56 for variable reference target length. This highlights the quality of generated text and the significant improvement when the generated text length is aligned with the reference target. Our work was benchmarked against African American Vernacular (AAVE) revealing that BLEU scores for AAVE are significantly lower than those for Standard American English, with BLEU given as 0.26. Our Nigerian Pidgin model, with a BLEU score of 0.56, shows a better performance. However, both results suggest that both dialects are challenging for language models. Leveraging the pre-trained transformer-based language model and evaluation metrics, we showcase the model's capacity for coherent Nigerian Pidgin text generation. For future research, the research work can serve as a good foundation for advancement and progress in the Nigerian Pidgin language generation and other low-resource languages.

Keywords Transformers · Nigerian pidgin · Controllable text Generation · Natural Language Processing · Natural Language Generation · Pre-trained Language models

1 Introduction

The evolution of Nigerian Pidgin, a low-resource language also known as Naijá is attributed to linguistic adaptation, multilingual code-switching, and code-mixing (Aji et al., 2022; Bob & Obiukwu, 2022; Saeed et al., 2024). Nigerian Pidgin is a creole that is popular in Western Africa (Okafor, 2022; Oyewusi et al., 2020). It is widely spoken and used as the language of choice for entertainment, film and social media content in Nigeria. Naijá has grown significantly over recent decades, emerging as the most widely spoken and arguably the most influential language in Nigeria (Adelani et al., 2024). Due to its mutual intelligibility with other West African Pidgins, it offers substantial potential for regional

✉ Taiwo Kolajo
taiwo.kolajo@fulokoja.edu.ng; taiwo.kolajo@up.ac.za

Kabir Garba
kabir.garba@fulokoja.edu.ng

Joshua B. Agbogun
joshua.agbogun@fulokoja.edu.ng

[1] Department of Computer Science, Federal University Lokoja, P.M.B 1154, Lokoja, Kogi State, Nigeria

[2] Department of Informatics, Faculty of Engineering, Built Environment & IT, University of Pretoria, Pretoria, Republic of South Africa (RSA)

integration and serves as a powerful tool in the pursuit of sustainable development.

Previous research attempts for Nigerian languages like Hausa, Igbo and Yoruba left significant steps in the understanding and development of tools for Nigerian Pidgin (Brasoveanu & Andonie, 2020). Due to its presence in daily communication and a distinct linguistic mix, there is now a need for Nigerian Pidgin-specific solutions (Cahyawijaya et al., 2021). Recently, the proliferation of digital technology and the internet has inspired a growing interest in leveraging natural language processing (NLP) techniques to explore and enhance the capabilities of different languages, including non-standard or low-resource ones. This increase in interest has led to improvements in the area of machine translation, sentiment analysis and text generation. However, Nigerian Pidgin, despite its widespread use and cultural significance, has been largely overlooked in the realm of NLP and language technology development (Kolajo et al., 2019, 2020).

The development of NLP tools for Nigerian Pidgin is not only a matter of linguistic curiosity but also a practical necessity. While acknowledging that progress has been made there still more needs to be done because Nigerian Pidgin serves as the lingua of choice for communication among people from different linguistic backgrounds, there is a growing demand for technology-driven solutions that can facilitate effective communication, education and information dissemination in this language (Chang et al., 2020). Transformer-based text generation serves as a development in the field of Artificial Intelligence (AI), which uses machine learning algorithms to generate new content (Dong et al., 2022). Recently, a Transformer-based language model has been developed for Natural Language Generation (NLG) which produces high-quality, coherent text in response to arbitrary input. The Transformer architecture has revolutionized the field of NLP (Khan et al., 2023; Waswani et al., 2017). Pre-trained Language models like GPT-2 are one of the recent advancements in Natural Language Generation (NLG). This Transformer-based language model enables the generation of high-quality-coherent text (Groenwold et al., 2020). Breakthroughs have been recorded due to its ability to handle long-range dependencies and to capture contextual information in machine translation, text summarization and language generation tasks (Brasoveanu & Andonie, 2020). Natural language generation (NLG), a sub-field of artificial intelligence, has brought remarkable enhancements in the generation of human-like texts (Bandi et al., 2023). This paper aims to create a specialized text generation mode by leveraging the pre-trained transformer architecture and to increase the diversity and inclusivity of the language environment while enabling more open and inclusive digital communication. Apart from enhancing communication this study also combats the lack of diversity in terms of NLP tools for low-resource languages.

The rest of this paper is structured as follows: Sect. 2 presents the background and related work underpinning the Nigerian Pidgin Language generation. The methodology to achieve Nigerian Pidgin text generation is discussed in Sect. 3. Sect. 4 presents the results and discussion while Sect. 5 concludes the paper.

2 Background and related work

This section presents the background and the related work concerning low-resource language generation, especially Nigerian Pidgin.

2.1 Nigerian Pidgin

Pidgin as a Nigerian English adaption has developed through time as a result of linguistic adaptation, multilingual code-switching, and code-mixing (Okafor, 2022). Nigerian Pidgin is a simplified form of communication used to bridge the language gap between two distinct cultures or language groups, often when English is the dominant language. West African Nigerian Pidgin is mostly spoken in Nigeria, with other varieties being spoken in various countries in Africa, such as Cameroon and Ghana.

Nigerian Pidgin has a long history in Africa, dating back to the 16th century when it was used among merchants and traders in the coastal areas of West Africa. Since then, it has evolved and changed depending on the needs of the speakers (Bob & Obiukwu, 2022). Today, West African Nigerian Pidgin is used in informal contexts, and it is a type of Creole, meaning it has its own unique rules and structure. It combines elements of English with other local languages, such as Yoruba, Hausa and Igbo. Common features include pre-verbal particles, characteristic subject pronouns, simplified verb forms and shared vocabulary.

Nigerian Pidgin is a key part of Nigerian culture, and it is also spoken in neighbouring countries. There have been efforts recently to popularize the monolingual Nigerian Pidgin as seen in the BBC Pidgin however, it remains under-resourced in terms of the available parallel corpus for machine translation (Chang et al., 2020). The Transformer model, developed by Google researchers, was unveiled in 2017 and has completely changed the field of natural language processing by enabling more precise sentence encoding. Transformer-based text production is based on this model (Okafor, 2022). This method of text production can be applied to many tasks, including question-answering, machine translation, and summarization. Additionally, it can be used to create fresh texts that look to have been

authored by people. Recent research has demonstrated that when it comes to people who are members of impoverished groups, NLG tools tend to be biased (Groenwold et al., 2020). Due to bias, when subjected to transformer-based text generators like GP2, resource-rich languages like English typically perform better and provide greater accuracy than low-resource languages.

2.2 Transformer-based text generation for Nigerian Pidgin

Transformer-based text production is a recent innovation in artificial intelligence (AI) that uses machine learning algorithms to generate fresh content (Yu et al., 2022). This technology can model the complex links between words and sentences in a language, making it ideal for producing natural language content. Because it recognises long-range dependencies in the input sequence, the transformer is ideal for producing cohesive and contextually relevant text. With enough training data and fine-tuning, the transformer model can learn to write high-quality writing that is identical to text written by a human (Topal et al., 2021).

Transformer-based text creation can be used to generate high-quality, natural-sounding text in the case of Nigerian Pidgin, an informal language that arose as a result of linguistic adaptability and code-switching. This is because the transformer architecture is intended to reflect long-term dependencies in a language, which is required for generating coherent text in Nigerian Pidgin.

The significance of transformer-based text creation for Nigerian Pidgin resides in its ability to bridge the language gap between Nigerian Pidgin speakers and speakers of other languages. Even though it is not legally recognised as a language of instruction in schools or other formal settings, Nigerian Pidgin is extensively used in Nigeria and is recognised as a valuable language for communication (Bob & Obiukwu, 2022). Recent research has revealed the circumstances in which natural language models (NLG) models exhibit bias towards particular languages even though pidgin corpus development and translation have received so much attention. The creation of a transformer-based text-generating system for Nigerian Pidgin can produce texts that are more linguistically accurate and diversified, improving language comprehension and enabling more realistic and natural discussions. It can also result in applications like increased sentiment analysis and sentiment categorization, as well as machine translation accuracy improvements.

2.3 Related work

Oyewusi et al. (2020) studied the evolution of Pidgin, a West African (especially Nigerian) English adaption using multi-language code-switching, code-mixing and inference for natural language generation (NLG) in a few-shot environment. In sequence linguistic adaptation, this output token selection across the two generators enables the adapter to take into account just task-relevant elements. The authors argue that employing direct English sentiment analysis of Nigerian social media posts is sub-optimal because it fails to capture semantic diversity and contextual change in the modern meaning of these phrases. To solve this problem, they supplement sparsely human-labelled code-changed material with an abundance of synthetic code-reformatted text and meaning. They also provide 300 VADER lexicon-compatible Nigerian Pidgin sentiment tokens and scores, as well as 14,000 gold standard Nigerian Pidgin tweets and sentiment labels.

Chang et al. (2020) developed a natural language generation technique to bridge the gap between West African Nigerian Pidgin and English. To create relevant machine translation systems and NLP datasets for Nigerian Pidgin, the authors used a monolingual Nigerian Pidgin text and a parallel English data-to-text corpus to build a system that can automatically generate Nigerian Pidgin descriptions from structured data. Dong et al. (2022) provided an in-depth look at Natural Language Generation, a subject within Artificial Intelligence and Natural Language Processing that focuses on generating language that is both understandable and coherent for humans. They explored traditional approaches, statistical models and deep learning approaches to open-domain dialogue systems and investigated the current state of research in this booming field.

Groenwold et al. (2020) created a dataset of intent equivalent parallel African American Vernacular English (AAVE)/Standard American English (SAE) tweet pairs to evaluate the effect of utilizing GPT-2 to generate text in AAVE. Sentiment analysis results demonstrate that, while AAVE text has more negative sentiment classifications than SAE, using GPT-2 enhances positive sentiment occurrences for both. The human examination also demonstrates that GPT-2 generated text has a high level of contextual rigour and overall quality.

Iqbal et al. (2022) offered an overview of deep generative modeling developments for text generation using deep learning algorithms. The authors analysed several deep learning models that are used for text production and discussed the past, present and future of text generation models in deep learning, focusing on publications published after 2015. This review also includes the many models and methodologies investigated and evaluated in various NLP application fields. Syed et al. (2021) reviewed recent research in automatic text summarizing, focusing on neural network-based abstractive summarization. The survey presented several neural network-based abstractive summarization

models as well as a proposed conceptual framework that included five important elements: encoder-decoder architecture, processes, training techniques, optimization algorithms, dataset and evaluation measure. The survey's goal was to provide a general overview of modern neural network-based abstractive text summarization models, as well as to raise awareness of the challenges and issues connected with these systems. The paper proposed the use of pre-trained language models in conjunction with neural network architecture for abstractive summarization tasks and the review revealed that transformer-based encoder-decoder architecture models are the new state-of-the-art. The analysis was performed qualitatively, and a concept matrix was used to indicate similar trends in the design of recent neural abstractive summarization systems. In the same vein, Bandi et al. (2023) investigated the fundamentals of generative AI systems focusing on requirements necessary for the implementation, taxonomy of architectural characteristics, input-output formats classification, and evaluation metrics.

Wang et al. (2019a) explored Transformer designs for language models. These models have shown usefulness for many NLP tasks on large-scale corpora utilizing pre-trained language models, but they are suboptimal for language modelling itself. The authors attempted to improve language modelling by adding extra LSTM layers while retaining computational efficiency. They presented a Coordinate design Search (CAS) approach to locate an effective design through iterative refining of the model to accomplish this. Experiment findings on the WikiText2 and WikiText103 datasets revealed that CAS attained perplexity scores ranging from 20.42 to 34.11 on all tasks, representing a 12.0 perplexity unit improvement over state-of-the-art LSTMs. The source code is publicly available. Wang et al. (2019b) used a neural topic module and a variational auto-encoder-based neural sequence module to construct topic-based sentences while providing considerable flexibility to the estimated posterior of the latent code during model inference. The proposed model has been evaluated and found to outperform previous approaches for both unconditional and conditional text creation. The topic-guided variational auto-encoder model may generate semantically relevant sentences on a variety of themes.

Yu et al. (2022) proposed a few-shot generative approach for rewriting a conversational query. The method is based on rules and self-supervised learning to generate weak supervision data using large amounts of ad hoc search sessions and to fine-tune GPT-2 for the task. On the TREC Conversational Assistance Track, the weakly supervised GPT-2 rewriter improved the state-of-the-art ranking accuracy by 12%, using very limited amounts of manual query rewrites. Notably, the rewriter still gave a comparable result to previous state-of-the-art systems in the zero-shot learning setting. Analysis of the results showed that GPT-2 effectively picked up task syntax and captured context dependencies, including hard cases with group references and long-turn dependencies.

Zhang et al. (2022) addressed the Controllable Text Generation (CTG), a new topic of research in the field of natural language generation (NLG). The authors emphasized the significance of CTG for the creation of cutting-edge text generation technologies that are more believable and better able to adhere to certain limitations in real-world applications. They stressed how the most recent iteration of NLG is based on comprehensive pre-trained language models (PLMs), especially transformer-based PLMs that enable the production of more diverse and fluid text. The controllability of these methods must be guaranteed because deep neural networks have a low interpretability degree. To handle this problem, the authors used transformer-based PLMs to give a systematic critical review of the common duties, approaches, and evaluation methods in CTG. They discussed the wide variety of approaches that have developed over the last three to four years, each of which focuses on a different CTG job that might call for a different kind of controlled restriction.

The findings from studies reveal common themes that deserve attention; insufficient evaluation, limited practical application, ignoring language-specific considerations, neglecting the diversity of linguistic and cultural backgrounds, lack of knowledge integration, and limited analysis of constraints and biases (Chang et al., 2020). Many studies fail to compare and evaluate text generation techniques or models in terms of their performance and effectiveness. This is particularly true when it comes to low-resource languages like Pidgin (Groenwold et al., 2020; Floridi & Chiriatti, 2020).

3 Methodology

The research methodology borders on the development of a transformer-based text generation model for Nigerian Pidgin. In this section, the architecture along with the description of its components, the rules for text generation, and the sequence diagram, highlighting the processes and how a user interacts with the model are presented. The development process of the transformer-based text generation was in these phases: data collection, preprocessing, feature engineering, training/fine-tuning, testing, and evaluation.

3.1 Description of components of the architecture

This section presents the description of each component in the architecture as depicted in Fig. 1.

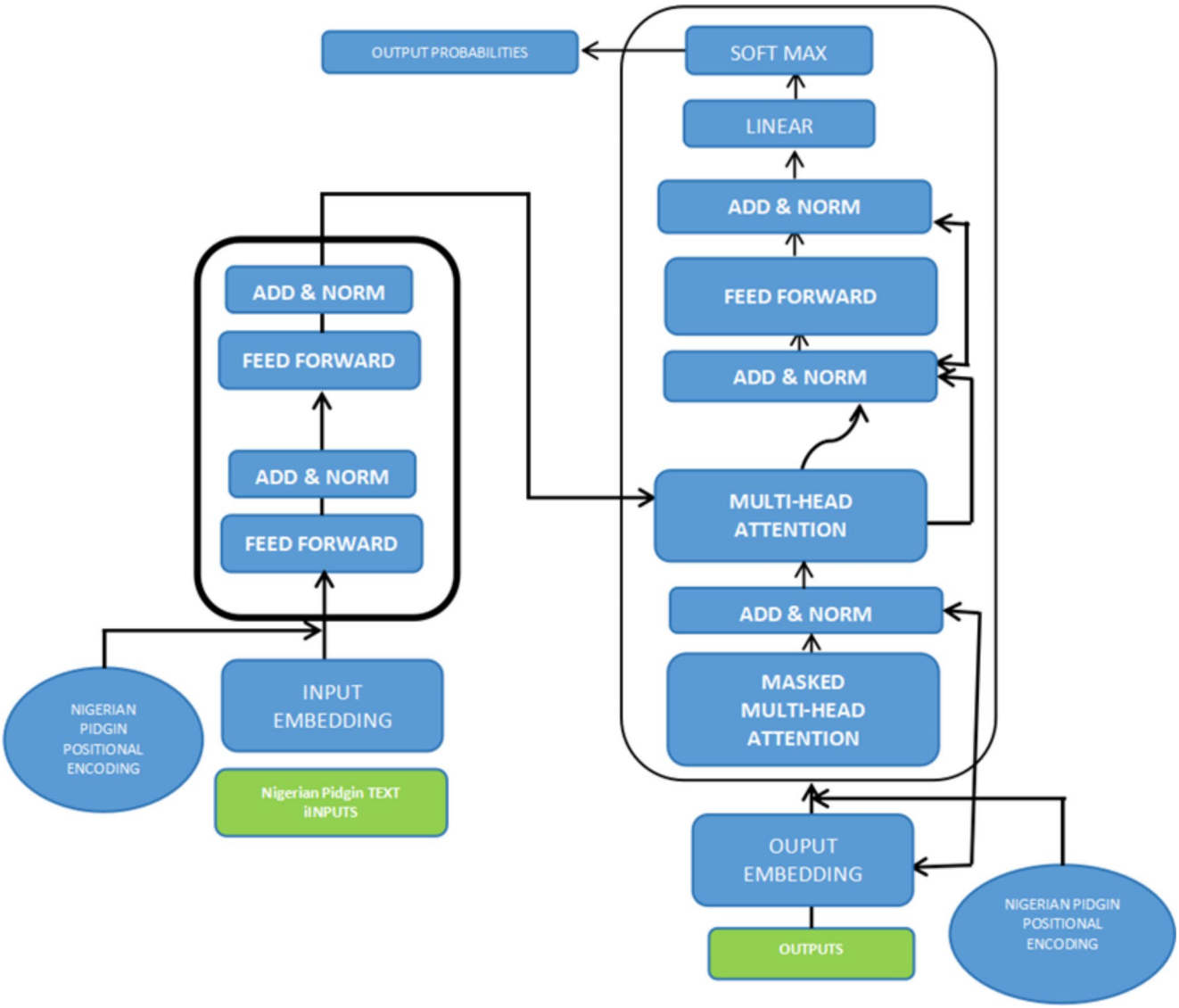

Fig. 1 Architecture of the transformer based text generation for Nigerian pidgin

a. **Nigerian Pidgin text inputs**: This component represents the raw Nigerian Pidgin text data that is fed into the model. It is the initial input that undergoes various transformations throughout the architecture to generate meaningful outputs.
b. **Input embedding**: The input text is transformed into dense vectors (embeddings) in this layer. These embeddings are numerical representations of the input words or tokens, capturing their semantic meaning in a high-dimensional space. This is crucial for enabling the model to process and understand the text.
c. **Nigerian Pidgin positional encoding**: Since transformer models do not inherently understand the order of tokens, positional encoding is added to the embeddings to provide information about the position of each word in the sequence. This helps the model maintain the sequential nature of the text.
d. **Masked multi-head attention**: In this layer, the model focuses on different parts of the input sequence to generate predictions. The "masked" aspect ensures that the model only considers previous words when predicting the next word in a sequence, which is essential for text generation tasks.
e. **Add & norm (add and normalize)**: This layer applies residual connections (adding the input of the layer to its output) followed by normalization. This process stabilizes and speeds up the training process by preventing the vanishing or exploding gradient problem, ensuring the model learns effectively.

f. **Feed forward**: After the attention mechanism, the data passes through a feed-forward neural network. This network further processes the information and helps the model capture complex patterns and relationships in the text.
g. **Multi-head attention**: Similar to the masked attention layer, but this time, the model can attend to any part of the input sequence. It allows the model to consider different parts of the sequence simultaneously, capturing relationships between words that are far apart in the text.
h. **Output embedding**: This layer converts the processed data back into a dense vector format, which can then be used to generate the final output sequence. It essentially reverses the input embedding process but in the context of generating text.
i. **Nigerian Pidgin positional encoding (for output)**: Positional encoding is again applied here to ensure that the order of generated tokens is taken into account when producing the output sequence. This maintains the coherence and grammatical structure of the generated text.
j. **Linear transformation**: This is a simple linear transformation applied to the output embeddings before generating the final predictions. It maps the high-dimensional embeddings back to the vocabulary space, preparing the data for the final SoftMax layer.
k. **SoftMax**: The SoftMax function is applied to the output of the linear layer to convert the raw scores into probabilities. These probabilities represent the likelihood of each token in the vocabulary being the next word in the sequence.
l. **Output probabilities**: The final output of the model is a set of probabilities for each possible word in the vocabulary. The word with the highest probability is selected as the model's prediction for the next word in the sequence.

3.2 Rules of text generation

a. **Length limits**: In this text generation process, inputs and outputs are by setting specific length constraints. This is achieved through parameters such as max_length for text generation and block_size for data preparation. These constraints ensure that the generated text remains within practical and manageable limits (Li et al., 2024).
b. **Vocabulary restrictions**: The text generation model relies on a predefined managed vocabulary that is established during the tokenization phase. This vocabulary is crucial for the model's operation, as it defines the set of tokens the model can use, thereby managing computational efficiency and ensuring relevant text generation.
c. **Formatting**: Text formatting is handled through preprocessing steps that clean and prepare the data before it is fed into the model. This includes removing non-alphanumeric characters and converting text to lowercase, ensuring that the generated text adheres to expected formatting norms.
d. **Neural networks**: The core of the text generation process involves GPT-2, a neural network-based transformer model. Unlike simpler statistical models, GPT-2 leverages deep learning to understand and generate text based on complex patterns and dependencies learned from extensive datasets (Pandey, 2024).
e. **Prompt design**: Effective prompt design (prompt engineering) is crucial for guiding the model's output. The prompts used in this text generation process are carefully crafted to elicit desired responses from the model. This involves designing prompts that provide clear context and direction, ensuring the generated text meets specific requirements and objectives (Pandey, 2024).

3.3 Process workflow

In this section, the workflow which includes data collection, data preprocessing, feature engineering, model training/fine-tuning, and model evaluation is presented (see Fig. 2).

3.4 Data collection

The dataset used for this research was acquired via Twitter which is a popular social media platform and publicly available datasets referred to as the castorini/afribertacorpus, accessible on the Huggingface website's datasets section. The primary source of textual data was the castorini/afriberta corpus. The Twitter API provided access to Twitter's vast data repository particularly BBCPidgin, allowing for the collecting of real-time and user-generated material in Nigerian Pidgin. We used pre-existing datasets freely offered via the Hugging Face website, a centre for natural language processing tools, in addition to Twitter data. These datasets presented unique challenges due to the various types and degrees of linguistic noise. This combination allowed us to create a broad and complete corpus by combining real-time Twitter posts with selected text sources.

3.5 Data preprocessing

To prepare it for model training, the obtained textual data underwent considerable preprocessing. Tokenization and lemmatization were part of the preparation workflow.

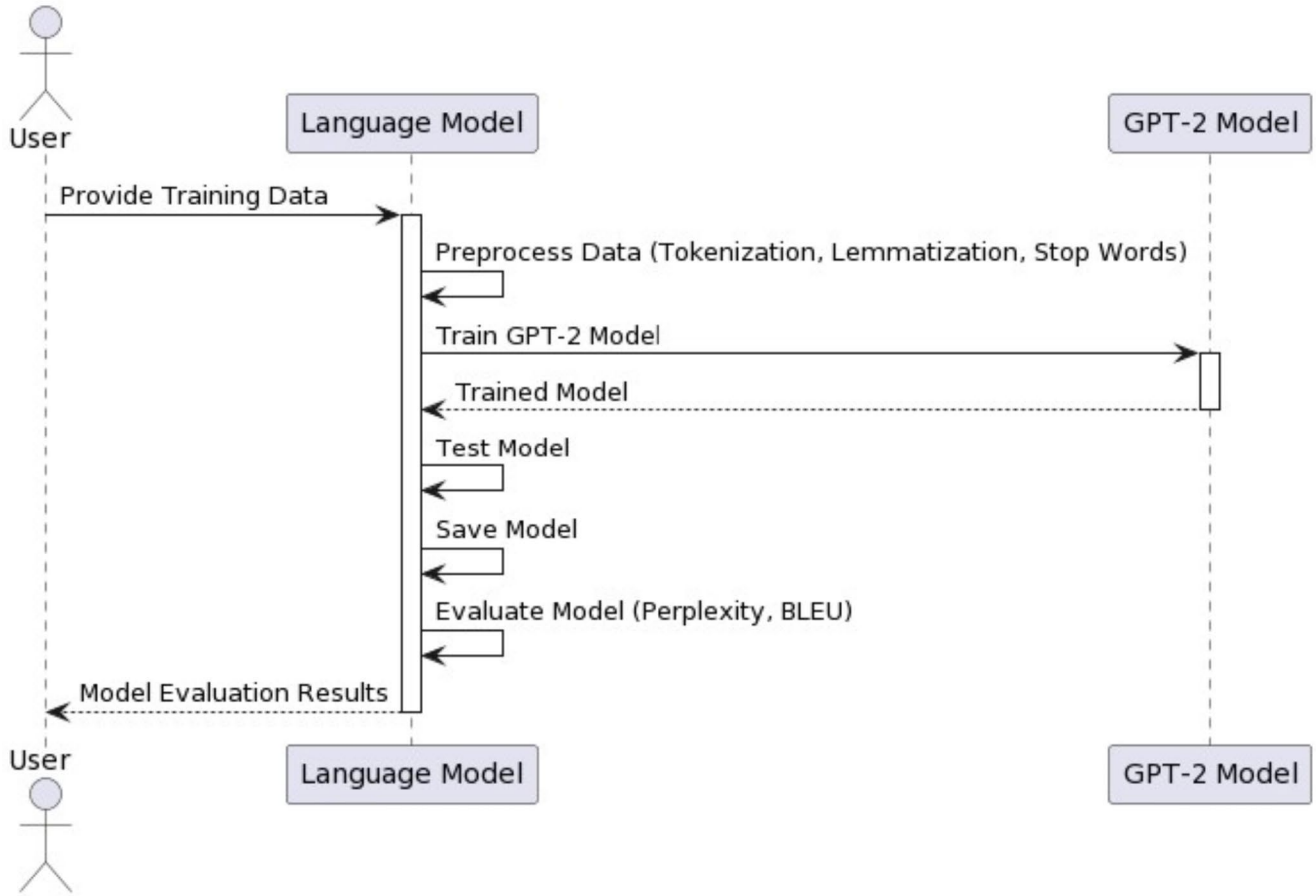

Fig. 2 Sequence diagram of the transformer based text generation for Nigerian pidgin

Tokenization was used to separate the raw text into individual tokens or words while maintaining the grammatical structure of Nigerian Pidgin phrases. Lemmatization, on the other hand, sought to reduce terms to their base or dictionary form, reducing dataset dimensionality and guaranteeing vocabulary consistency. Stopword removal was carried out on the data to remove unwanted noise and special characters which are irrelevant and unneeded. We utilized about 15,744 words and 2249 unique words. These pretreatment measures were critical in reducing the issues given by Nigerian Pidgin's informal language usage, slang, and non-standard orthographic variances. This will make the model to discover significant patterns and create cohesive text by transforming the data into a structured and standardized format.

3.6 Feature engineering

In this case, feature engineering is essentially related to the creation of input features sufficient for training our transformer-based model, namely GPT-2. As input, the model needed token sequences, which were produced by sliding a fixed-size window across the tokenized and lemmatized data. Furthermore, particular care was taken to create attention masks that highlight the places of real tokens in a given sequence while masking out padding tokens. This allowed the model to concentrate solely on the relevant information, which was important to the training process's success.

3.7 Model training/fine-tuning

The pre-processed datasets were used for training and fine-tuning our transform-based model. GPT-2 was used as the starting point for transfer learning approaches. GPT-2, a cutting-edge language model, was pre-trained on a massive corpus of text spanning several languages and disciplines.

On the Nigerian Pidgin datasets, we fine-tuned this model by changing its weights to our particular linguistic environment, allowing it to create coherent and contextually meaningful Nigerian Pidgin text. About 161,842 Nigerian Pidgin tweets were gotten from castorini/afriberta-corpus on the hugging face website, while over 200,000 tokens were used for fine-tuning for 100 epochs. An increase in the quality of text was observed at a temperature of 0.7. We applied

techniques such as gradient clipping, adaptive learning rate scheduling, and early stopping to aid in fine-tuning, model convergence, and stability. We tweaked the model's hyperparameters to increase its performance; the hyperparameters are the maximum sequence length (max length) parameter, number of return sequences and temperature. The max length hyper-parameter controls the maximum length of generated sequences, and the number of sequences parameter generates multiple diverse outputs enhancing robustness and creativity. Lastly, the Temperature hyperparameter regulates the randomness of word selection which improves the model's fluency and versatility.

3.8 Model evaluation

The effectiveness and quality of our Transformer-based text generation model for Nigerian Pidgin was extensively evaluated as part of this project. Perplexity scores and BLEU scores were used as evaluation measures.

Perplexity scores evaluated the model's ability to predict the next token in a sequence, giving details about its language comprehension and generation abilities. BLEU scores test the model's fluency and coherence in contrast to human-produced Nigerian Pidgin by measuring the similarity between generated text and reference texts. It is important to note that the model was finetuned for 100 epochs, while over 200,000 tokens were utilised. We conducted additional experiments to examine the effect of increasing the finetuning duration and training data size. We started finetuning from 50 to 100.

4 Result and discussion

We discuss our model's performance focusing on the quantitative metrics and qualitative insights gained from the evaluation. A publicly available dataset derived from the work of Ogueji et al. (2021) termed the Afriberta-corpus which is available on GitHub and Hugginface was employed for finetuning the GPT-2 model. The training data made up of over 200,000 tokens was used to finetune the GPT2 pretrained model for 100 epochs. The evaluation was carried out with Bleu and Perplexity as the evaluation metrics. Although the limitations of BLEU and Perplexity in assessing text generation accuracy for low-resource languages are notable. Automated metrics like BLEU were initially designed to evaluate translation systems and may not fully capture the quality of text generated in low-resource languages. For instance, BLEU scores can overestimate the performance of statistical models compared to rule-based ones and may fail to account for linguistic nuances or variations inherent in low-resource languages (Babych, 2014; Mokander et al., 2023). These metrics often rely on n-gram matching, which may not align with the unique syntactic and lexical properties of such languages, leading to discrepancies between automated scores and human judgments. Perplexity, while useful for evaluating language model performance, can also be misleading when applied to low-resource languages due to limited training data and vocabulary coverage. As highlighted by Lee et al. (2023) and Al-Khalifa et al. (2024), the effectiveness of automated metrics is context-dependent, and their ability to accurately reflect translation or generation quality varies across different languages and domains. Table shows the initial evaluation of the model.

Table 1 The evaluation score for the Model

Evaluation Metric	Score (Fixed text length)
Perplexity	43.26
BLEU	0.15

Table 2 Effect of increased epochs and Training data on evaluation Metrics

Metric	50 Epochs, 50 K Tokens	100 Epochs, 100 K Tokens	100 Epochs, 200 K Tokens
Perplexity (Fixed-Length)	46.72	43.66	43.26
BLEU (Fixed-Length)	0.02	0.08	0.15
Perplexity(Variable target)	47.12	43.96	43.56
Bleu(Variable Target)	0.21	0.38	0.56

From Table 1, we obtained a perplexity score of 43.26 which suggests that the model was not so certain in its prediction. This could be due to the limited size of the training data. A lower perplexity is an indication of a better prediction. For the BLEU metric, the average score of 0.15 indicates that the generated text is not too similar to the reference text.

Further tests were carried out to examine the impact of increasing the finetuning duration and training data size. As demonstrated in Table 2, increasing the finetuning period from 50 to 100 epochs reduced Perplexity from 46.72 to 43.66 while increasing BLEU from 0.02 to 0.08. Similarly, extending the dataset from 100,000 to 200,000 tokens reduced Perplexity further to 43.26 and improved BLEU to 0.15, suggesting that more data and training time enhance model performance. Additionally, when the evaluation was performed using a variable target reference length, Perplexity showed a slight increase (47.12 to 43.56) compared to the fixed-length evaluation, indicating slightly higher uncertainty. However, BLEU scores significantly improved, reaching 0.56 when the model was trained with 200,000 tokens over 100 epochs, underscoring the importance of aligning the length of generated text with the reference target for better performance.

Table 3 Sample of generated text based on the input

Evaluation prompt	Reference Target	Generated Text
I chop food sotay belle full, I no fit chop again.	I chop food sotay belle full, I no fit chop again; I just find space siddon make I rest small.	I chop food sotay belle full, I no fit chop again. correct guy wey sabi dey
I tell them say I no go fit come, dem talk say e go dey alright	I tell them say I no go fit come, dem talk say e go dey alright; dem say make I nor worry, dem go hold my own	I tell them say I no go fit come, dem talk say e go dey alright. bros abeg fit follow go

Table 4 Comparative scores for perplexity and BLEU

Metric	Current Study	Groenwold et al. (2020) (AAVE)
Perplexity	43.56	N/A
BLEU	0.56	0.26
Rogue	N/A	0.81

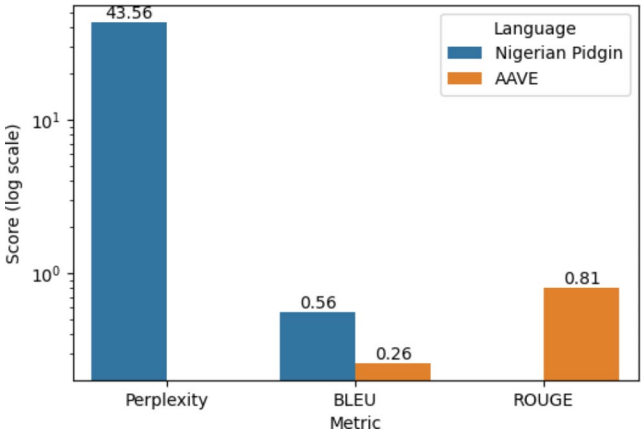

Fig. 3 Comparison of BLEU, ROUGE, and perplexity for pidgin English and AAVE

A good Perplexity score is typically lower, indicating higher certainty in the model's predictions, while a good BLEU score is higher, reflecting closer similarity to the reference text. Perplexity doesn't require a reference text but is affected by the `max_length` function in GPT-2, which determines the number of tokens the model is to generate. Table 3 presents the sample of generated text based on the input.

From Table 3, each generated text begins with the prompt given to the model. It is good to note that the reference target is only applicable to BLEU for evaluation as perplexity does not require a reference target.

We compared our Perplexity and BLEU scores with those found in a study on African American Vernacular English (AAVE) by Groenwold et al. (2020) to put the performance of our model in perspective (see Table 4). As with our work with Nigerian Pidgin, their study demonstrated the difficulties in utilising GPT-2 to generate text in a non-standard English dialect.

The BLEU score for AAVE was much lower than the BLEU score for Standard American English (SAE), with AAVE achieving a BLEU score of 0.26. Our Nigerian Pidgin model, with a BLEU-1 score of 0.56, performs better. However, the results, suggest that both dialects remain challenging for language models to process effectively. Although Groenworld et al. (2020) utilized BLEU and ROUGE, with a ROGUE score of 0.81, the authors did not utilise the Perplexity metric, our Perplexity score is 43.56. The results indicate that these non-standard English variations, including AAVE and Nigerian Pidgin, present similar difficulties in predicting the next token, suggesting consistent challenges across these non-standard English varieties. The comparison is presented in Fig. 3.

Although the accuracy and quality of text generated are not of the standard intended, the finetuning duration and substantial training data offer promise for capturing language patterns. This demonstrates the model's potential and provides a foundational framework for future research in this area, offering insights and opportunities for improvement. In the future, Few-shot learning is essential for picking up on linguistic nuances and model performance can be improved by fine-tuning on the small Nigerian Pidgin dataset. Although not optimal due to English bias, Bleu and Perplexity were employed as temporary standards, highlighting the necessity for customized metrics in future evaluations. Overall, the model can be said to be conclusively successful in generating Nigerian Pidgin text.

5 Conclusion and further work

In this study, we addressed the pressing need for the creation of a transformer-based text production model for Nigerian Pidgin, a widely spoken language in West Africa. Nigerian Pidgin deserves more attention, especially in the fields of text production and natural language processing. This restriction serves as a problem for generative AI technology or AI-assisted creative writing, formal publications, and blog posting. Our research aimed to close this gap and enhance the utility of Nigerian Pidgin across domains such as business, entertainment, social media, and machine translation. The overall goal of this research work is to develop a transformer-based text generation for Nigerian Pidgin that is coherent and linguistically accurate. To achieve this, we leveraged pre-trained transformer models and built a language model that was tested and evaluated. We were able to collect and preprocess Nigerian Pidgin texts. In addition, we built a model utilizing the GP2 pre-trained model for fine-tuning.

We successfully evaluated the data using evaluation metrics like BLEU and PERPLEXITY. Additionally, we examined the effect of finetuning duration on the model's performance when the size of training data and epoch are increased. The finetuning experiment examined the effect of increased epochs and data on evaluation metrics. Perplexity decreased slightly with more epochs and data, indicating improved model performance. BLEU scores improved significantly with additional tokens, suggesting that more training data enhanced translation quality. The result showed considerable progress and a stepping stone towards addressing the dearth of sophisticated language technology despite the limited resources. The average BLEU score of 0.15 (fixed length), 0.56 (variable length based on reference target), and the perplexity of 43.265 demonstrates the rigorous evaluation process which sets a standard (a baseline) for future evaluations. Our Nigerian Pidgin model, with a BLEU score of 0.56, performs better than 0.26 of AAVE in Groenwold et al. (2020). Though they used ROUGE instead of Perplexity, our score of 43.56 is comparable, highlighting similar prediction challenges across these varieties. The study also demonstrates the model's capacity to generate Nigerian Pidgin text whose quality and coherence are expected to significantly increase with increment in the size of training data and epoch. The model can be further utilized for interactive educational content creation and storytelling and to break down language barriers and promote understanding between different cultures.

From the evaluation scores obtained for both Perplexity and Bleu metrics, the built model has shown promise in addressing the challenges of developing transformer-based text generation models for under-resourced languages as illustrated in the case of Nigerian Pidgin. A better result in terms of coherence, quality and evaluation scores can be achieved with further fine-tuning at elevated epochs and much larger training data. In addition, evaluating Nigerian Pidgin text generation models with traditional evaluation metrics like Bleu and Perplexity comes with its challenges. These metrics may not fully capture the linguistic nuances, unique vocabulary and syntactic variations of Nigerian Pidgin. Currently, there are no sufficient benchmark datasets for Nigerian Pidgin making human evaluation preferable and ideal for evaluating such models. For future work, we recommend exploring the use of Named Entity Recognition (NER) as a complementary evaluation metric for low-resource language models especially for entity identification, particularly in the context of Nigerian Pidgin English. NER has been effectively applied in other low-resource languages, such as Tigrinya (Yohannes & Amagasa, 2022), where it provided a more thorough evaluation of the model's capabilities. By incorporating NER, future studies could ensure that entities are accurately identified in generated outputs, offering a more nuanced understanding of how well a text generation model preserves the integrity of entity recognition and categorization. This approach is especially crucial in low-resource settings, where limited training data can hinder a model's ability to correctly identify and classify entities. NER could thus enhance the robustness of evaluations by providing a detailed analysis of the model's performance in generating contextually accurate text.

Funding We would like to disclose that this research received no external funding. We, the authors, independently conducted this study without financial support from any organization or entity. We affirm that these statements accurately reflect our circumstances and commitments regarding competing interests and funding for the research presented in this paper.
Open access funding provided by University of Pretoria.

Declarations

Competing interests We, the authors, declare that we have no competing interests, financial or non-financial, that are directly or indirectly related to the work submitted for publication.

References

Adelani, D. I., Doğruöz, A. S., Shode, I., & Aremu, A. (2024). Which Nigerian-Pidgin does generative AI speak? Issues about representativeness and bias for multilingual and low resource languages. *arXiv preprint arXiv:2404.19442*.

Aji, A., Winata, G., Koto, F., Cahyawijaya, S., Romadhony, A., Mahendra, R., Kurniawan, K., Moeljadi, D., Prasojo, R. E., Baldwin, T., Lau, J. H., & Ruder, S. (2022). One Country, 700+languages: NLP challenges for underrepresented languages and dialects in Indonesia. In *Proceedings of the 60th annual meeting of the association for computational linguistics (Vol 1: Long Papers)* (pp. 7226–7249). Dublin, Ireland, ACL. https://doi.org/10.18653/v1/2022.acl-long.500

Al-Khalifa, H., Al-Khalefah, K., & Haroon, H. (2024). Error analysis of pretrained language models (PLMs) in English-to-Arabic machine translation. *Human-Centric Intelligent Systems*, *4*, 206–219. https://doi.org/10.1007/s44230-024-00061-7

Babych, B. (2014). Automated MT evaluation metrics and their limitations. *Tradumàtica Tecnologies de la Traducció*, *12*, 464–470.

Bandi, A., Adapa, P. V. S. R., & Kuchi, Y. E. V. P. K. (2023). The power of generative AI: A review of requirements, models, input-output

formats, evaluation metrics, and challenges. *Future Internet*, *15*(8), 260. https://doi.org/10.3390/fi15080260

Bob, P. O., & Obiukwu, E. N. (2022). Exploring the linguistic status of the Nigerian Pidgin. *PREORC Journal of Arts and Humanities*, *7*(1), 173–186.

Brasoveanu, A. M. P., & Andonie, R. (2020). Visualizing transformers for NLP: A brief survey. In *2020 24th international conference information visualisation (IV)*, (257–266). https://doi.org/10.1109/IV51561.2020.00051

Cahyawijaya, S., Winata, G., Wilie, B., Vincentio, K., Li, X., Kuncoro, A., Ruder, S., Lim, Z. Y., Bahar, S., Khodra, M., Purwarianti, A., & Fung, P. (2021). IndoNLG: Benchmark and resources for evaluating Indonesian natural language generation. In *Proceedings of the 2021 conference on empirical methods in natural language processing (EMNLP)* (pp. 8875–8898). https://doi.org/10.18653/v1/2021.emnlp-main.699. Online and Punta Cana, Dominican Republic. ACL.

Chang, E., Adelani, I., Shen, D., X., & Demberg, V. (2020). Unsupervised pidgin text generation by pivoting English data and self-training. In *2020 International conference on learning representation (ICLR 2020)*. Addis Ababa, Ethiopia.

Dong, C., Li, Y., Gong, H., Chen, M., Li, J., Shen, Y., & Yang, M. (2022). A survey of natural language generation. *ACM Computing Surveys*, *55*(8), 173. https://doi.org/10.1145/3554727

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, *30*(4), 681–694. https://doi.org/10.1007/s11023020-09548-1

Groenwold, S., Ou, L., Parekh, A., Honnavalli, S., Levy, S., Mirza, D., & Wang, W. Y. (2020). Investigating African-American vernacular English in transformer-based text generation. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)* (pp. 5877–5883). ACL. https://aclanthology.org/2020.emnlp-main.473

Iqbal, T., & Qureshi, S. (2022). The survey: Text generation models in deep learning. *Journal of King Saud University-Computer and Information Sciences*, *34*(6), 2515–2528.

Khan, M., Ullah, K., Alharbi, Y., Alferaidi, A., Alharbi, T. S., Yadav, K., Alsharabi, N., & Ahmad, A. (2023). Understanding the research challenges in low-resource language and linking bilingual news articles in multilingual news archives. *Applied Sciences*, *13*(15), 8566. https://doi.org/10.3390/app13158566

Kolajo, T., Daramola, O., & Adebiyi, A. (2019). Sentiment analysis on Naija-tweets. In *Proceedings of the 57th annual meeting of the association for computational linguistics: Student research workshop (ACL)* (pp. 338–343). 28 July– 2 August, Florence, Italy. https://doi.org/10.18653/v1/P19-2047

Kolajo, T., Daramola, O., Adebiyi, A., & Seth, A. (2020). A framework for preprocessing of social media feeds based on integrated local knowledge base. *Information Processing & Management*, *57*(6), 102348.

Lee, S., Lee, J., Moon, H., Park, C., Seo, J., Eo, S., Koo, S., & Lim, H. (2023). A survey on evaluation metrics for machine translation. *Mathematics*, *11*(4), 1006. https://doi.org/10.3390/math11041006

Li, J., Tang, T., Zhao, W. X., Nie, J. Y., & Wen, J. R. (2024). Pre-trained language models for text generation: A survey. *ACM Computing Surveys*, *56*(9), 1–39.

Mokander, J., Schuett, J., R Kirk, H., & Floridi, L. (2023). Auditing large language models: A three-layered approach. *AI Ethics*, *2023*. https://doi.org/10.1007/s43681-023-00289-2

Ogueji, K., Zhu, Y., & Lin, J. (2021). Small data? No problem! Exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st workshop on multilingual representation learning* (pp. 116–126). Punta Cana, Dominican Republic. ACL.

Okafor, A. Y. (2022). Nigerian Pidgin as a national tool for communication. *Journal of Emerging Trends in Educational Research and Policy Studies*, *13*(1), 3946.

Oyewusi, W., Adekanmbi, O., & Akinsande, O. (2020). Semantic enrichment of Nigerian Nigerian Pidgin for contextual sentiment classification. In *2020 International conference on learning representation (ICLR 2020)*. Addis Ababa, Ethiopia.

Pandey, R., Waghela, H., Rakshit, S., Rangari, A., Singh, A., Kumar, R., & Sen, J. (2024). Generative AI-based text generation methods using pre-trained GPT-2 model. *arXiv preprint arXiv:2404.01786*.

Saeed, M., Bourgonje, P., & Demberg, V. (2024). Implicit discourse relation classification for Nigerian pidgin. *arXiv preprint arXiv:2406.18776*.

Syed, A. A., Gaol, F. L., & Matsuo, T. (2021). A survey of the state-of-the-art models in neural abstractive text summarization. *IEEE Access: Practical Innovations, Open Solutions*, *9*, 13248–13265. https://doi.org/10.1109/ACCESS.2021.3052783

Topal, M. O., Bas, A., & van Heerden, I. (2021). Exploring transformers in natural language generation: GPT, BERT, and XLNet. In *International conference on interdisciplinary applications of artificial intelligence (ICIDAAI)*. 21–23 May, Kongre Tarihi.

Wang, C., Li, M., & Smola, A. J. (2019a). Language Models with Transformers. ArXiv. /abs/1904.09408 https://arxiv.org/abs/1904.09408

Wang, W., Gan, Z., Xu, H., Zhang, R., Wang, G., Shen, D., & Carin, L. (2019b). Topic-guided variational autoencoders for text generation. In *NAACL HLT 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies-proceedings of the conference* (pp. 166–177). Association for Computational Linguistics (ACL).

Waswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kasier, L., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems, (NIPS)*. 4–9 December, Long Beach, CA, USA.

Yohannes, H. M., & Amagasa, T. (2022). Named-entity recognition for a low-resource language using pre-trained language model. In *Proceedings of the 37th ACM/SIGAPP symposium on applied computing (SAC '22)* (pp. 837–844). https://doi.org/10.1145/3477314.3507066

Yu, W., Zhu, C., Li, Z., Hu, Z., Wang, Q., Ji, H., & Jiang, M. (2022). A survey of knowledge-enhanced text generation. *ACM Computing Surveys*, *54*(11s), 1–38. https://doi.org/10.1145/3512467

Zhang, H., Song, H., Li, S., Zhou, M., & Song, D. (2022). A survey of controllable text generation using transformer-based pre-trained language models. *ACM Computing Surveys*, *56*(3), 64. https://doi.org/10.1145/3617680