Tamil Content Generation Using Transformer[Yazhi]

Punidha[1]
Department of Artificial Intelligence and Data Science KPR
Institute of Engineering and Technology Coimbatore,
Tamil Nadu, India

Gokulachalam[2]
Department of Artificial Intelligence and Data Science KPR
Institute of Engineering and Technology Coimbatore,
Tamil Nadu, India

Karthi Prasad[3]
Department of Artificial Intelligence and Data Science KPR
Institute of Engineering and Technology Coimbatore,
Tamil Nadu, India

Ramakrishnan[4]
Department of Artificial Intelligence and Data Science KPR
Institute of Engineering and Technology Coimbatore,
Tamil Nadu, India

Abstract:- This paper presents Yazhi, a transformation model specially designed for Tamil, known for its robustness and unique language features Yazhi combines advanced transformer architecture with reinforcement learning, encoder -decoder system for Yazhi traditional model moves highly effective therefore , improving comprehensibility and generation of advanced Tamil text Represents a significant improvement in natural language processing, and offers robust solutions to computer understanding and translation challenges using available resources mastery and learning to quickly understand subtle nuances of language. Through his seminal work, Yazhi sets a new benchmark for business research on Tamil content generation and translation.

Keywords:- *Tamil Language, Transformer Architecture, Large Language Model, Yazhi, Morphological Complexity, Data Scarcity, Semantic Variability, Encoder-Decoder Structure, Reinforcement Learning, Python, PyTorch, NLP.*

I. INTRODUCTION

Tamil is known for its deep linguistic heritage and cultural significance, a systemic quality fostered by these challenges that reflects the challenges of the broader Natural Language Processing (NLP) landscape, and calls for new approaches that language this unique quality of centuries in literatureformed wealth.

In response to this need, this paper includes a comprehensive framework designed to address the multifaceted challenges of Tamil NLP. Enter Yazhi, a unique transformer model carefully designed to showcase the complexity of the Tamil language processing industry. This new approach uses advanced conceptual techniques and state-of-the-art reinforcement learning techniques, which not only aspire to improve content generation and semantic accuracy but also computational efficiency and it also balances functionality, increasing the general applicability of the model

At the core of this ambitious endeavor is a standard two-gram language model, an iteration of the Transformers design. This strategic choice not only confirms the project's aim of addressing the challenges of online text generation specific to Tamil, but also reflects a commitment to pushing the boundaries of language technology .Stemming from an iterative maintenance process fueled by real-world applications and community feedback, Yazhi will not only meet but exceed existing standards in Tamil NLP, he has said, to get things done new and superior.Yazhi, on the other hand, deals with subtle phonetic and grammatical peculiarities of Tamil, which are often stumbling blocks for traditional NLP models Yazhi's architecture lies in his ability to see and lay out nuances in a rich fabric of context its meanings have been expressed in the ancient and modern on texts of different idiomatic idioms to blend in the Tamil translation. These combinations enable Yazhi to surpass the ability of traditional models to capture the nuances of Tamil poetry, prose and language with remarkable accuracy Yazhi's development philosophy places great emphasis on creating a tool that is more compatible with Tamil speakers, especially those interested in using GPT (Generative Pre-trained Transformer) technology in their native language. The development of this initiative is closely linked with feedback and insights from a wide range of Tamil speakers including students, technology enthusiasts and everyday users Tamil it is a valuable asset.

Yazhi offers unmatched resources for Tamil speakers who are interested in using the power of GPT for creative writing, content creation, or academic research If we hear about Tamil culture and language more than knowledge knowledge base and acquired, the model opens up new possibilities for authors and researchers to more efficiently and accurately assess parameters. This capability is transformative for those working to digitize and preserve Tamil literature, making ancient texts more accessible to modern audiences and translators Moreover, teachers and students stand to benefit greatly, as Yazhi offers learning in providing personalized experiences, translating educational materials into Tamil and engaging them through interactive platforms.The international aspect of Yazhi's influence promises to break down language barriers and empower Tamil speakers globally. Yazhi generates its language based on RL. By incorporating human feedback, RLHF loop enhances Yazhi's ability to produce Tamil text. This two-pronged strategy guarantees Yazhi adjusts and improves its linguistic abilities efficiently.

II. RELATED WORKS

The transformational effect of [1] is mediated inside the reconstruction of natural language processing components. This led to transformational transformation tactics, substantially improved the capture of contextual relationships using self-focused techniques and became essentially adaptive for subsequent developments in language models.Delving into versatile quick-shot learning capabilities, [2] suggests the surprising capability of GPT-3 to perform a variety of tasks with a limited number of training models. This is amazing. The adaptability and breadth they highlighted using large-scale language models, opening new avenues for research.The revolutionary undertaking of [3] revolutionized the previous training techniques. The pioneers of bidirectional contextual learning using a language model based mask, the section obtained groundbreaking work on natural language processing tasks, providing a new theoretical framework for contextual representation.The exceptional impact of the scaled-up model is shown in [4], a GPT that excels in brief instructions and demonstrates that it can do an impressive range of thingswithin natural language applications. The 3.5 Turbo solidifies its position as the pinnacle of the larger language models.In the landmark work on [5], the section outlines the adaptation of a more developed model. Excelling in brief instructions, GPT- 3.5 Turbo demonstrates impressive abilities in a variety of natural language applications, solidifying its position asunique among large-scale languages.Tackling the limitations of predetermined context within the original transformer architecture, [6] presents a partition tier repeat component, essentially progressing the model's capacity to handle longer groupings and conditions, contributing to a more nuanced dialect understanding[7] introduces an attentive language model with a segment level recurrence mechanism, addressing the restrictions of fixed-length and enhancing the understanding of longer sequences.[8] introduced an innovative framework, advancing multilingual language data and playing a vital role in the enchancement of neural machine translation.[9] refines BERT's pre- training methods, improving the model's robustness and overall performance across numerous natural language processing tasks. This optimization contributes to the continued success and applicability of BERT-based models.In [10], the versatility of text-to-text transformer models is confirmed, pushing the limits of transfer learning capabilities in natural language understanding and establishing new avenues for unified approaches to diverse language tasks.[11] introduced an integrated text-to-text conversion model. The section pushes the limits of transferable learning abilities in natural language understanding, presenting a new approach that opens new avenues for integrated solutions for language tasks.[12] proposed an autoregressive pretraining method to improve language understanding and modeling performance, which significantly contributed to the advanced state of the autoregressive language model.[13] explores neural reading comprehension, providing valuable insights beyond traditional language processing tasks and contributing to a deeper understanding of complex language comprehension challenges.[14] Tells about large-scale model training strategies, addressing the challenges associated with large- scale model training. This work contributes to thedevelopment of large-scale model training and efficiency, which allows further improvements in the field.[15] Examines generative pretraining to improve language understanding, and sheds light on advances in language model development and implementation.[16] Enhances text generation in Tamil by expanding LLaMA's vocabulary with 16,000 Tamil tokens, using LoRA for efficient training, and making significant strides in large language models' performance in Indian languages.

III. PROPOSED SOLUTION

➢ *Tamil Alpaca Dataset*

In order to achieve exceptional accuracy and precision in Tamil Natural Language Processing (NLP) using our Yazhi model, we prioritize the use of well-curated and optimized Tamil data sets. These datasets form the cornerstone of our learning model, and a key step involves data cleaning, where sophisticated techniques, such as tokenization, lemmatization, and part-of- speech tagging, are used to remove noise and information inconsistencies from known for sufficient compound words and complicated syntactic rules.

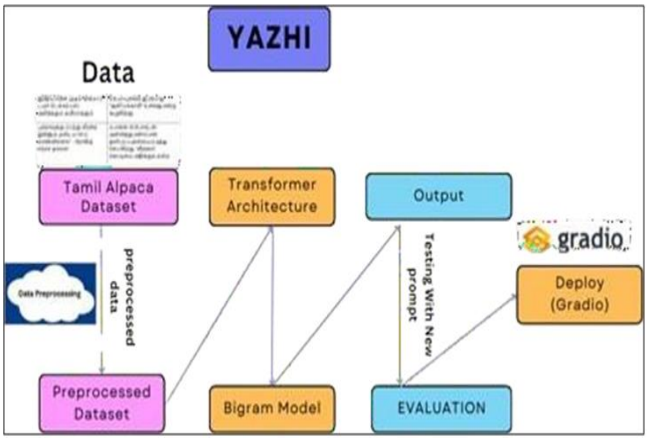


Fig 1 Working of Yazhi

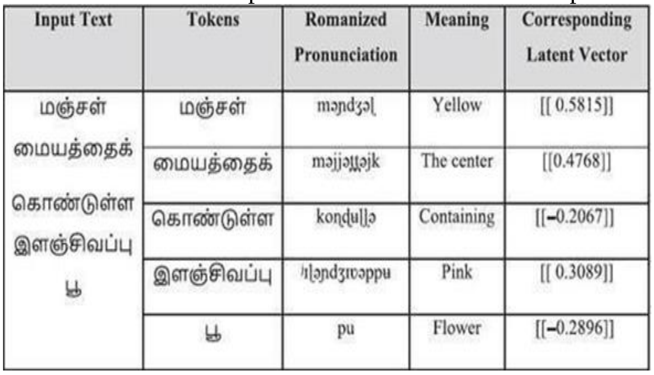
➢ *Preprocess Data:*

Yazhi's special model has a custom transformer architecture purpose-built to nuance Tamil words and idioms. Graphical models developed for bigram analysis play an important role in dissecting deep contextual linguistic structures within Tamil, affecting predictions in various linguistic scenarios. Other advances include sophisticated feedforward layers in other neural network layers. Dynamically prioritizing different segments of input data on the basis, thus facilitating further learning This process is further optimized by advanced optimization techniques and hyperparameter tuning.

➢ *Transformer Architecture:*

At the heart of Yazh's language processing capabilities is a custom transformer architecture [YAZHI]. This purpose-built design ensures a nuanced understanding of the Tamil language, contributing to the model's high accuracy and efficiency in text understanding and generation. The Transformer architecture is a key component for handling complex language scenarios and is optimized for incremental learning using advanced neural techniques.

Table 1 Vector Representation of Words - Example



Input Text	Tokens	Romanized Pronunciation	Meaning	Corresponding Latent Vector
மஞ்சள் மையத்தைக் கொண்டுள்ள இளஞ்சிவப்பு பூ	மஞ்சள்	mənd͡ʒəɭ	Yellow	[[0.5815]]
	மையத்தைக்	məjjəʈʈəjk	The center	[[0.4768]]
	கொண்டுள்ள	koɳɖuɭɭə	Containing	[[−0.2067]]
	இளஞ்சிவப்பு	ʰɪɭənd͡ʒɪʋəppu	Pink	[[0.3089]]
	பூ	pu	Flower	[[−0.2896]]

➢ *Output:*

The output of the Yazhi model in Tamil Natural Language Processing (NLP) encompasses various aspects, offering specified insights into linguistic evaluation and generation.This comprehensive overview includes tokenized representations, lemmatized output, element-of-speech tagging, generated text, contextual predictions, structural evaluation facts, and assessment metrics effects.

➢ *Evaluation and Gradio Deployment :*

To ensure the effectiveness of Yazhi's language model, rigorous assessment strategies are in place, assessing metrics which include accuracy, precision, and contextual understanding to validate performance. This ongoing evaluation lets in for non-stop improvement and addition to evolving linguistic patterns and consumer desires.Our innovative language technology framework seamlessly integrates into the Gradio platform for consumer-pleasant deployment. This integration guarantees clean admission to and utilization of Yazhi's advanced capabilities in diverse packages, providing a transformative experience in Tamil communications. Gradio deployment, with its person-pleasant interface, opens avenues for broader adoption and effect.

In Yazhi’s case, the bigram model is the foundational issue of our probabilistic language version. Unlike traditional models, the Yazhi bigram model is now not the handiest work on the letter stage but also displays the complicated morphological structure of Tamil, making it more effective for computational linguistics and herbal language processing offerings inclusive of speech generation and text prediction.

Trained on an extensive Tamil script, Yazhi's bigram model absolutely understands the frequencies and shapes in the language and in the schooling process generates a letter transition matrix, where each row and column represents a letter of Tamil characters. The dots inside the discern show one color after some other. Once trained, Yazhi uses this matrix to expect the following letter in the sequence with the aid of looking at the following most likely letter for a given ‘modern’ letter. Yazhi’s use of the bigram model no longer ensures linguistic accuracy but also stands proud for its computational efficiency.

This performance makes Yazhi's Bigram Model ideal for real- time applications, along with predictive textual content entered for Tamil keyboards or real-time translation offerings. Yazhi's Bigram Model has become an important asset for content material generation in Tamil. It allows the automatic introduction of coherent and contextually applicable Tamil text, proving valuable for applications ranging from automatic responses in chatbots to producing predictive textual content pointers. Moreover, this version serves as a strong basis for adding superior language processing obligations within Yazhi, inclusive of system translation or speech reputation, wherein a deep expertise in the language's shape is paramount.

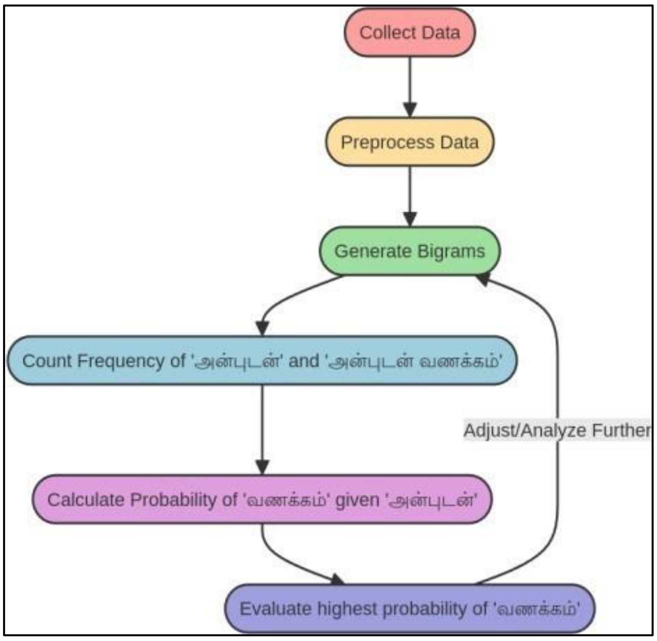

Collect Data
Preprocess Data
Generate Bigrams
Count Frequency of 'அன்புடன்' and 'அன்புடன் வணக்கம்'
Adjust/Analyze Further
Calculate Probability of 'வணக்கம்' given 'அன்புடன்'
Evaluate highest probability of 'வணக்கம்'

Fig 2 Workflow of Bigram Model

IV. BIGRAM MODEL

Yazhi innovatively integrates bigram probabilities into its transformer-based architecture, achieving a nuanced understanding of Tamil text. Unlike conventional bigram models that only consider the immediate predecessor, "Yazhi" analyzes broader context, significantly enhancing its predictive capabilities.Now, let's look at a practical example from the 'Alpaca' dataset to see how Yazhi works.

➢ *Case Study - "அன் புடன் வணக்கம் "*

The phrase "2468 4 " ("Greetings with love"), derived from the "Alpaca" dataset, serves to illustrate "Yazhi's" proficiency. "Yazhi" predicts the likelihood of " " (greetings) following " " (with love) by evaluating both the statistical probability and the context surrounding the phrase, showcasing a significant advancement over traditional models.Bigram models calculate the likelihood of a word based on its prompt forerunner utilizing the equation:

$$P(Wn \mid Wn\text{-}1) = C(Wn\text{-}1, Wn)/C(Wn\text{-}1)$$

Where C(Wn-1, Wn) is the count of how many times the bigram occurs, and C(Wn-1) is the count of prefix word. Incorporating contextual information, "Yazhi" adjusts bigram probabilities for the phrase "அன் புடன் வணக்கம் " from the "Alpaca" dataset. Assuming " 60 " appears 50 times and the complete phrase 30 times, the base probability is:

$$P(B|A)\ Count(A)\ /Count(A,B)$$

P(வணக்கம் | அன் புடன்) =30/50 =0.6. "Yazhi" enhances this calculation by factoring in additional contextual information from the sequence, using transformer architecture to dynamically adjust the probability, reflecting a more nuanced understanding of language use.

V. EVALUATION

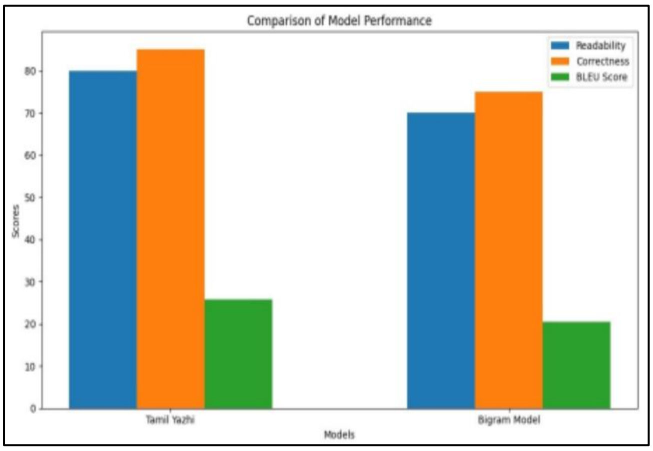


Fig 3 Evaluation of Yazhi

The Tamil Yazhi model demonstrates exceptional readability and grammatical correctness scores for its language generation capabilities. Despite a lower BLEU score against reference human translations, its high readability and correctness metrics indicate robust linguistic modeling and fluent text generation. In future development, the model's performance will likely be further improved to better match human references and increase the BLEU score.

VI. CONCLUSION

Yazhi stands as a groundbreaking advancement in the realm of Tamil Natural Language Processing, offering a beacon of possibility for Tamil speakers worldwide. By using the power of transformer architecture and reinforcement learning, Yazhi transcends conventional limitations, fostering a new erawhere digital communication and content creation in Tamil are not just feasible but flourish with unprecedented ease and accuracy. This transformative model has the potential to democratize technology for Tamil speakers, making it accessible and relevant to their linguistic heritage. For the Tamil-speaking community. Additionally, Yazhi can revolutionize customer service and support for Tamil-speaking users, providing them with assistance in their native language, thus enhancing user experience and satisfaction. In the domain of content creation, Yazhi paves the way for Tamil writers and content creators to produce more content efficiently, breaking language barriers and reaching wider audiences. It could significantly reduce the effort and time required for translation and content generation, encouraging the preservation and propagation of Tamil culture and literature. Moreover, Yazhi's capabilities can be leveraged in developing more intuitive and user-friendly interfaces for digital platforms, ensuring that Tamil speakers can navigate, interact, and benefit from technology in their native language. This inclusivity not only promotes digital literacy among Tamil communities but also ensures that the benefits of technology are equitably distributed. Furthermore, Yazhi could play a crucial role in enhancing accessibility for Tamil-speaking populations, offering tools for speech recognition and generation that cater specifically to their linguistic needs. This will be especially useful for people with inabilities, giving them with the implies to communicate and get to data more openly. In essence, Yazhi embodies a significant stride toward bridging the digital divide, ensuring that Tamil speakers are not left behind in the digital revolution. By enabling more effective communication, content generation, and interaction in Tamil, Yazhi contributes to the preservation of linguistic diversity in the digital age. Its development is a testament to the potential of specialized NLP models to enrich the lives of specific linguistic communities, showcasing a path forward for similar initiatives in other languages. Through Yazhi, the Tamil- speaking community can look forward to a future where technology speaks their language, literally and figuratively, enhancing their daily lives and cultural engagement.

REFERENCES

[1]. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is All You Need. Advances in Neural Information Processing Systems.

[2]. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language Models are Few-Shot Learners. arXiv preprint arXiv:2005.14165.

[3]. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Few-Shot Learners. arXiv preprint arXiv:1905.05583.

[4]. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprintarXiv:1810.04805.

[5]. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). The GPT-3.5 Turbo Language Model.

[6]. Dai, Z., Yang, Z., Yang, F., Carbonell, J., Le, Q. V., & Salakhutdinov, R. (2019). Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context. arXivpreprint arXiv:1901.02860.

[7]. Radford, A., & Dailey, D. (2018). "Enhancing Dialect Comprehension through Generative Pretraining. OpenAI Technical Report."

[8]. Vaswani, A., Bengio, S., Brevdo, E., Chollet, F., Gomez, A. N., Gouws, S., ... & Uszkoreit, J. (2018). Tensor2Tensor for Neural Machine Translation. arXiv preprint arXiv:1803.07416.

[9]. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Approach. arXiv preprint arXiv:1907.11692.

[10]. Brown, T. B., Hellwig, L., Tu, Z., Langlois, V., Neelakantan, A., Ng, A. Y., & Bellemare, M. G. (2019). Language Models are Few-Shot Learners. arXiv preprint arXiv:1901.08103.

[11]. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2019). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv preprint arXiv:1910.10683.

[12]. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. arXiv preprint arXiv:1906.08237.

[13]. Chen, T. Q., So, D. R., Li, C., & Liang, P. (2017). Neural Reading Comprehension and Beyond. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.

[14]. Shoeybi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J., & Catanzaro, B. (2019). Megatron: Large-scale Model Training with 3D Parallelism. arXiv preprint arXiv:1909.08053.

[15]. Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Advancing comprehension of dialects through generative pretraining

[16]. Abhinand Balachandran, "TAMIL-LLaMA: A NEW TAMIL LANGUAGE MODEL BASED ON LLAMA 2," arXiv:2311.05845v1 [cs.CL], Nov. 2023.

[17]. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E.Sayed. Mistral7b, 2023.

[18]. Ainslie, J. Lee-Thorp, M. de Jong, Y. Zemlyanskiy, F. Lebrón, and S. Sanghai. Gqa: Training generalized multi- query transformer models from multi-head checkpoints, 2023.

[19]. Caswell, T. Breiner, D. van Esch, and A. Bapna. Language id in the wild: Unexpected challenges on the path to a thousand-language web text corpus, 2020.

[20]. T. Kudo and J. Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing, 2018.