Future Technology
Open Access Journal
ISSN 2832-0379

November 2025| Volume 04 | Issue 04 | Pages 138-145

Journal homepage: https://fupubco.com/futech

https://doi.org/10.55670/fpll.futech.4.4.12

Future Publishing LLC

Article

FNet-GPT: Fourier-based lightweight transformer for emotion-aware text generation using GPT

Atul Haribhau Kachare[1,2]*, Chandrashekhar Goswami[1], Ashutosh Gupta[1], D.S. Chouhan[3]

[1]Department of Computer Science & Engineering, Sir Padampat Singhania University, Udaipur, Rajasthan, India
[2]Department of Computer Engineering, Shah & Anchor Kutchhi Engineering College, Mumbai, Maharashtra, India
[3]Department of Mathematics, Sir Padampat Singhania University, Udaipur, Rajasthan, India

ARTICLE INFO

Article history:
Received 15 June 2025
Received in revised form
25 July 2025
Accepted 13 August 2025

Keywords:
Story generation, Emotionally aware narration, FNET, GPT-2, Natural Language Processing

*Corresponding author
Email address:
atul.kachare@spsu.ac.in

DOI: 10.55670/fpll.futech.4.4.12

ABSTRACT

Neural story generation models have two significant challenges: (1) coherence over narrative structure, especially long-range dependencies, and (2) emotional coherence and consistency, generally producing redundant or incoherent narration. A new, emotionally intelligent two-stage short story generation model is presented by combining GPT-2 with a tailored FNET model, a light transformer architecture substituting standard self-attention with Fourier Transform layers to improve semantic and emotional relationship capture in text. The first stage employs GPT-2 to generate a list of input candidate sentences, a question, an answer, and an emotional state. The candidate sentences are then filtered using an emotion classifier from DistilRoBERTa to keep only those that adhere to a desired emotional tone. The filtered sentences are then fed into a fine-tuned FNET model, which examines inter-sentence relationships and enforces emotional coherence to generate a coherent and emotionally engaging narrative. An empirical comparison using three benchmark datasets demonstrates the system's superiority over earlier state-of-the-art approaches. The FNET model achieves 0.3093 in BLEU-1, outperforming Plan-and-Write (0.0953) and T-CVAE (0.2574), with an enhanced narrative quality and lexical coherence with human-written narratives. The story coherence and emotion retention accuracies are 85%, 67%, and 60% for Visual7W, ROCStories, and Cornell Movie Dialogs datasets.

1. Introduction

Text generation is a process of automatically creating text that sounds like it was written by a human, which is trending in the world of artificial intelligence (AI). It has its roots in the early days of computational linguistics research [1]. Nowadays, with modern text generation models and deep learning methods like transformers [2], we can write a text that's not only coherent but also contextual. It employs various applications, including writing and creating content for power chatbots, as well as generating code [3]. However, to move ahead, we must consider the ethical considerations [4]. Short story generation, a part of text generation, is about crafting tight narratives with gripping plots and characters. With early models like the n-gram language [5], creating stories with sense was difficult. Nevertheless, with deep learning, especially with the introduction of models like GPT-3 [6], machines can tell creative stories, complete with rich characters and surprising twists. Still, we have many issues, such as making sure stories are coherent, diverse, and used ethically [7]. Text generation covers a broad spectrum of creating different types of text sentences, paragraphs, or code while ensuring grammar and coherence. It can be used for machine translation, summarization, or question answering. Story generation, a subset of text generation, focuses explicitly on crafting narratives. It requires linguistic fluency and understanding of plot structures, character development, and thematic elements. Story generation aims to create engaging and imaginative tales that evoke emotions and captivate readers. This paper is categorized as follows: Section 1 introduces the research area. The background of the problem and a discussion of different approaches are given in Section 2, whereas Section 3 outlines a literature review. Section 4 details the proposed system and its methodology. Section 5 outlines the experimental setup, including datasets, baselines, and metrics, and presents the results obtained. The final section concludes the article by discussing the findings and potential future directions. References to relevant works are provided at the end.

2. Background

The historical trajectory of story generation is deeply rooted in human culture, evolving from oral traditions and written literature to computational approaches [8]. Early computer programs utilized rule-based systems, followed by more sophisticated methods like case-based reasoning and planning, exemplified by MINSTREL [9] and BRUTUS [10]. The rise of machine learning, particularly RNNs, LSTMs [11], and transformer-based models like GPT [12] and BERT [13], has significantly advanced the field. The current landscape is dominated by large language models (LLMs) like GPT-3 [3] and interactive storytelling. So, the different template-based systems [14], rule-based approaches [15], case-based reasoning [9], planning models [16], and even simulation techniques [17] were explored. With the arrival of neural language models like GPT-3, the story generation process is transformed into creative and diverse narratives. However, there are still some hurdles in controlling what these models generate and ensuring they are ethical. Looking ahead, we see a blend of strategies to craft stories that grab attention and ensure we keep those ethical considerations in mind.

3. Literature survey

The research conducted by Fan A. and Lewis M. [18] aimed to enhance fluency and coherence by utilizing a hierarchical model with self-attention, but encountered issues such as tokenization and text repetition. Xu J. and Ren X. [19] focused on making sentences connected semantically for better coherence, using a Seq2Seq model, but struggled because there were not enough human-annotated examples in the real-world datasets they were working with. Yao L. and Peng N. [20] introduced a hierarchical framework with explicit storyline planning, but the model struggled with off-topic content and inconsistencies. Wang T. and Wan X. [21] developed a model for generating coherent plots within incomplete stories using a T-CVAE, but faced limitations in generating story endings. Chen G. and Liu Y. [22] proposed generating an outline to bridge the gap between title and story, resulting in more extensive narratives, but capturing cross-sentence dependencies remained a challenge. Zhang Y. and Shi X. [23] utilized a Knowledge Graph to generate the image captioning, but acknowledged limitations due to the offline construction of the graph by computing cosine similarity. Chen G. and Liu Y. [24] create the outline from the training data with a title and story using a variational neural network, but they generate only a single-sentence outline, which restricts its ability to support complex or long-form story structures. Brahman F. and Chaturvedi S. [25] focused on generating emotionally aware stories, but the lack of large-scale annotated story corpora posed a challenge. Tan B. and Yang Z. [26] proposed a progressive generation method for long text passages, but the need to expand vocabulary while maintaining relevance and accuracy remained a limitation. Min K. and Dang M. [27] proposed generating short stories from images using RNNs and an encoder-decoder model, but with grammar and emotional expression limitations. Wu C. and Wang J. [28] introduced the ICPGN model for generating classical Chinese poems from images, but its reliance on a specific dataset limits it. Liu Y. and Huang Q. [29] proposed the SSAP model for generating story endings based on context and sentiment using ChatGPT-2. Jin Y. and Kadam V. [30] presented SCRATCHPLOT, a method for generating stories using pre-trained language models without fine-tuning, but it requires significant post-processing. Chen Y. and Li R. [31] proposed a co-creative visual storytelling method that allows user control over events and emotions, but needs better prompt formats and emotion classifiers. Khan L. and Gupta V. [32] focused on generating coherent stories using keywords and genre-based inputs by optimizing the Hugging Face GPT-2 model, but the influence of the title or keyword on the narrative remains limited. Based on the literature survey, some research gaps are identified as follows:

- Failure to manage the proper emotional context: Most models fail to guarantee that the output text is always tied to a target emotion throughout the story.
- Inadequate modelling of long-range dependencies: Current systems do not have strong mechanisms for modelling and preserving coherence across sentences or pieces of a story.
- Recurrent behaviours and narrative flow inconsistencies are commonly found in generative models because of shallow attention mechanisms or the absence of semantic feedback loops.
- Lack of a cohesive system that identifies emotions and increases coherence: Few research efforts have attempted to combine emotional filtering with semantic structuring in a single, effective system.

These identified gaps highlight the need for a unified story generation system to simultaneously manage emotional control and narrative coherence using efficient, scalable techniques. Although deep learning models like GPT-2 have revolutionized natural language generation, they remain limited in their ability to generate stories that are both emotionally consistent and narratively coherent. Current systems often produce emotionally neutral content and exhibit weak cross-sentence dependencies, leading to repetitive or disconnected storylines. Moreover, the lack of integration between emotion classification and sequence modeling further restricts the expressive quality of generated narratives. These limitations create a gap in developing AI-generated stories that genuinely resonate with human emotions and maintain a logical narrative flow. Bridging this gap requires a hybrid approach that fuses emotional filtering with semantically aware text generation. To address these limitations, this paper proposes a two-stage hybrid architecture that combines the generative power of GPT-2, the emotion recognition capability of DistilRoBERTa, and the semantic modeling efficiency of FNET. The key objectives of this research are:

- To ensure emotional alignment by filtering GPT-2-generated sentences using DistilRoBERTa.
- To enhance narrative coherence by modeling global dependencies through a Fourier-based FNET architecture.
- To empirically validate the system on benchmark datasets (Cornell Movie Dialogs, Visual7W, and ROCStories).
- To evaluate performance using BLEU and METEOR metrics and emotion retention accuracy.

4. Proposed system

Having performed an extensive review of existing literature, we identify two long-standing issues in the domain of automatic narrative generation:

- Repetition and Inconsistency: The majority of modern systems lack consistent character behavior and consistent story progression, instead generating repetitive or contradictory content.
- Lack of Cross-Sentence Dependency Modeling: The current models are limited by their failure to capture long-distance semantic and affective relations between sentences, impacting narrative coherence.

We have developed a hybrid model, as shown in Figure 1, that combines GPT-2, DistilRoBERTa, and a specially crafted FNET model to tackle specific challenges. FNET is a Transformer encoder that mixes via the Fourier transforms instead of using the usual self-attention method. It helps effectively capture those global dependencies.

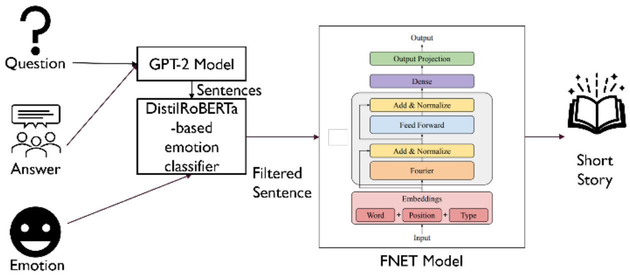

Figure 1. Proposed system for short story generation

The reason to use FNET:

- Global Context Awareness: Many attention mechanisms become complex as the input grows. Nevertheless, FNET's Fourier transform encoder is more scalable. It can handle complete sequences.
- Improved Consistency: Another great thing about FNET is that it keeps the narrative on track. It minimizes those annoying moments where the story drifts or characters seem inconsistent. It does this by holding onto those long-range dependencies.
- Reduced Computational Expense: FNET skips the whole attention mechanism altogether. It operates with a time complexity of O(n log n), which is significantly better than the O(n^2) we see in standard Transformers. The system, as proposed, works in two phases.

A comparative analysis of FNet with Linformer, Longformer, and Performer is presented in Table 1, which outlines the trade-offs across computational efficiency, memory footprint, and real-world performance. It supports the design decision to employ FNet for coherent and emotionally grounded narrative generation. FNet bypasses self-attention by applying a two-dimensional Discrete Fourier Transform (DFT) across token and embedding dimensions, enabling fast global mixing. Linformer, while highly efficient with O(n) complexity, uses low-rank projections that may weaken its ability to retain subtle narrative transitions or emotional nuances required in story generation. Longformer is optimized for document-level tasks using a sliding window and sparse global attention, which can be less effective for capturing inter-sentence dependencies in shorter narratives. Performer approximates self-attention using kernel-based methods and random projections, offering scalability but with increased computational cost. FNet thus provides a compelling balance: faster than self-attention, semantically expressive, and well-suited for maintaining long-range dependencies, which are crucial for generating coherent and emotionally aligned narratives.

4.1 Stage 1: Emotionally aware sentence generation

- The system is presented with a question, a corresponding answer, and a target emotional state (e.g., happiness, fear, anger).
- Sentence Generation: A pre-trained GPT-2 model produces some potential sentences based on the inputs provided.
- Emotion Filtering: Every sentence is passed through a DistilRoBERTa-based emotion classifier. Sentences belonging to the target emotion only are kept.
- Selection Mechanism: A beam search algorithm produces diverse sentence forms with maximum semantic and emotional appropriateness.

We employed the Emotion English DistilRoBERTa-base model [33], classifying text into seven categories: anger, disgust, fear, joy, neutral, sadness, and surprise. It is built by fine-tuning DistilRoBERTa-base on a balanced subset (~20k samples) drawn from six diverse English-language datasets, including GoEmotions, Crowdflower, ISEAR, MELD, etc. Each emotion category is represented by approximately 2,811 examples, with 80% used for training and 20% for evaluation. On held-out data, the model attains an accuracy of 66%.

4.2 Stage 2: FNET-based coherent story generation

The FNet encoder replaces the standard self-attention mechanism of transformers with a 2D Discrete Fourier Transform (DFT2), achieving efficient global mixing with reduced computational complexity. This section formally describes the mathematical operations that govern the encoding process. Given a sequence of tokens $x = [x_1, x_2, \dots, x_n]$. Each token is embedded in a vector, $E_i = Embed(x_i) \in R^d$. So the input matrix becomes:

$$X \in R^{n \times d} \tag{1}$$

Where n: sequence length, d: embedding dimension, each row x_i correspond to the embedding of the ith token.

To preserve the order of tokens in the input sequence, we add fixed sinusoidal positional encodings $P \in R^{n \times d}$and type encodings $T \in R^{n \times d}$

$$X' = X + P + T \tag{2}$$

These encodings allow the model to distinguish between tokens at different positions without learning position-specific parameters. Instead of using quadratic-complexity self-attention, FNet applies a 2D Discrete Fourier Transform (DFT2) across both the token and embedding dimensions of the input:

$$Z = Re\left(FFT2(X')\right)R^{n \times d} \tag{3}$$

Table 1. Comparative analysis of efficient transformer variants across time complexity, memory usage, parallelizability, and practical applicability

Model	Time Complexity	Memory Usage	Parallelizability	Real-World Use
FNET	O(n log n)	Low	High	High-speed on GPUs; minimal memory
Lin former	O(n)	Low	Moderate	Needs pre-defined projection; sensitive to rank
Long former	O(n) (locally), O(n^2) for global tokens	Medium-High	Medium	Scales well on long docs, less for short sequences
Per former	O(n)	Medium	High	Random projections increase training overhead

Here, FFT2(.) is a 2D Fourier Transform applied to the matrix (1st DFT across Rows: Sequence dimension and then 2nd DFT across columns: Feature dimension), and Re(.) extracts the real part to keep the result compatible with downstream layers. This operation transforms the input into the frequency domain, enabling global interaction between all tokens via frequency components. The core idea is that Fourier mixing captures long-range dependencies through global frequency patterns without explicitly computing pairwise attention scores. It replaces the standard attention computation with an efficient and non-learned operation that reduces the time complexity from $O(n^2)$ to $O(n\log n)$, as demonstrated in the original FNet work [34]. At this stage, we combine the output of the Fourier Transform block Z with the original embedding input X'. It helps retain the original input signal and makes the model more stable during training. After this addition, a Layer Normalization operation is applied to the result. LayerNorm standardizes the summed output along the feature dimensions to improve convergence and avoid internal covariate shift.

$$Y = LayerNorm\,(Z + X') \tag{4}$$

The normalized output is passed through a standard position-wise feedforward network:

$$H = ReLU(YW_1 + b_1)W_2 + b_2 \tag{5}$$

Where, $W_1, W_2 \in R^{n\times d}$ are learnable weight matrices and $b_1, b_2 \in R^d$ are biases. ReLU adds non-linearity to improve representational power.

Finally, another residual connection is added between the input to the Feedforward block and the output of the Feedforward block. It helps the model combine local (feedforward) and global (Fourier) features. After adding them, LayerNorm is applied again to stabilize and standardize the representation.

$$H' = LayerNorm\,(Y + H) \tag{6}$$

Once the final representation H' is ready and mapped to the vocabulary space using a linear output projection. It is done by multiplying it with a learnable weight matrix:

$$Logits = H'W_0 + b_0 \tag{7}$$

The model applies the softmax function to the logits to generate the next token and get probabilities. Then, it uses argmax to choose the most likely token:

$$\hat{y}_t = \arg\max\big(Softmax(Logits)\big) \tag{8}$$

The final output is a short story composed of emotionally consistent and semantically connected sentences generated from the filtered GPT-2 outputs and refined through FNET's encoding-decoding mechanism.

This system leverages FNET to pick up on the subtle meanings and feelings, along with GPT-2's amazing storytelling abilities. Such a multi-step journey; craft short stories that resonate with emotion while responding to specific questions and answers. The first stage is all about creating sentences that are aware of emotions. It begins with taking in a question, an answer, and an emotional context. GPT-2 works to weave a narrative that feels relevant and coherent. Then, the emotion classifier, which has been trained on data tagged with various emotions, helps pinpoint the specific feelings in the sentences that get generated. The beam search algorithm explores all kinds of word combinations, figuring out which arrangements could form sensible sentences. After generating all those options, the system goes through them again with the emotion classifier to ensure that the final picks match the intended emotional tone. In the latter part of the system, we have used an advanced FNET model as shown in Figure 2, in which a typical self-attention layer is replaced by a Fourier Transform layer [35] to run smoother and quicker. Self-attention is a mechanism that helps the model understand how different words relate to each other in a sentence [36]. The Fourier Transform method breaks down a signal into frequency components, helping us identify long-range correlations in the input sequence, which can be beneficial. The model's structure has two key parts to focus on: the encoder and the decoder. The encoder takes in the input sequence and, at the end of its job, comes up with a context vector. Then, the decoder takes over, grabs that context vector, and turns it into the output sequence. Positional embeddings are added to the input and output sequences to help us understand where each word belongs.

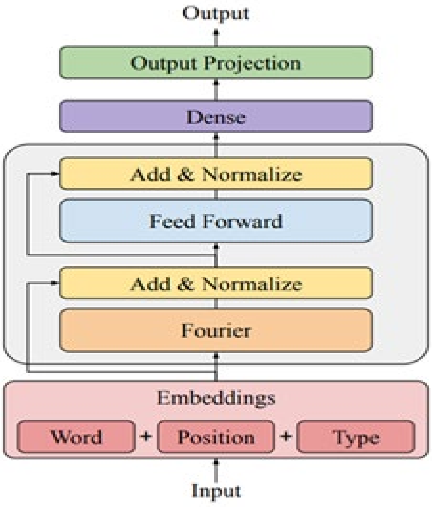

Figure 2. FNET model

5. Evaluation of the proposed system

We evaluate the system to see how well it could write an emotionally engaging short story. We examine three different benchmark datasets, each one having some unique challenges. The system was evaluated against some established baseline models to measure the quality and coherence of the stories. In the following sections, we will delve into the setup of our experiments, the results we obtained, and the insights gained from those findings.

5.1 Dataset

So, we decided to work with three benchmark datasets for our experiments to see how well the proposed system performs. First up, we have the Cornell Movie-Dialogs Corpus [37]. It is a pretty massive collection, with all these fictional conversations pulled from movie scripts, totaling around 220,579 exchanges between 10,292 pairs of characters. Next is the Visual7W dataset [38], which comes from COCO images and is packed with 327,939 question-answer pairs. Plus, it features 1,311,756 human-generated answer choices and 61,459 object groundings. While Visual7W is originally a visual QA dataset, in our framework, it plays a crucial role in enabling visually grounded story generation. We adapt the dataset by using its image-question-answer triples to train our Visual Question Answering and Visual Question

Generation (parts of a larger system), producing prompts and emotions that are used to guide the narrative generation process. It allows the system to generate stories that are not only coherent but also contextually aligned with visual content. Furthermore, the emotional relevance of each image is inferred via Visual Sentiment Analysis, helping to tailor the tone and affective quality of the generated narrative. Lastly, we looked at the ROCStories dataset [39], a resource for NLP research, with 98,162 five-sentence stories about everyday life. It has been designed to help models understand and generate stories by capturing causal and temporal relationships.

5.2 Baseline models

We compare our models with the following: 1) T-CVAE [21] leverages transformers and a variational autoencoder to learn story patterns and generate new, coherent stories, with the ability to incorporate additional input for guidance. 2) Plan&Write [20] operates in two stages: planning a storyline and generating the story text based on that plan, offering better control over story structure and coherence. The comparison aims to showcase the effectiveness and advancements of the proposed approach in story generation.

5.3 Metrics

We have chosen to utilize the BLEU (Bilingual Evaluation Understudy) [40] and METEOR [40] metrics to evaluate the quality of the generated stories. BLEU is one of the most popular machine translation metrics that determines the similarity between a human reference translation and a machine translation. BLEU verifies the similarity between the n-grams (set of n words) in the target text and the reference text—the more similar n-grams, the greater the BLEU score, which shows greater content preservation.

The BLEU score formula is derived based on precision and a penalty for brevity. The formula is:

$$BLEU = BP \times \exp\left(\frac{1}{n}\sum_{i=1}^{n} \log p_i\right) \quad (9)$$

However, BLEU does not consider the stems and synonyms of words. To overcome these limitations, METEOR uses a weighted F1-score and penalty function.

$$F_{SCORE} = \frac{10\left(\frac{matched\ unigram}{unigram\ in\ hypotheis}\right)\left(\frac{matched\ unigram}{unigram\ inreference}\right)}{\left(\frac{matched\ unigram}{unigram\ inreference}\right)+9\left(\frac{matched\ unigram}{unigram\ in\ hypotheis}\right)} \quad (10)$$

$$Penalty = 0.5\ \times \left(\frac{Number\ of\ chunks\ in\ Hypotheis}{Number\ of\ matched\ unigrams}\right) \quad (11)$$

$$METEOR = F_{SCORE}\ \times\ (1 - Penalty) \quad (12)$$

5.4 Experimental setup

We utilize a pre-trained GPT-2 model to generate a wide range of sentences. It will take a question and its answer, then churn out several ways to say it. Now, the emotion classifier will look at all those candidate sentences and determine which hits the emotional mark for the story we are crafting. This way, our sentences will help shape the narrative's overall vibe. At the core of our story-generating system is the FNET model — a transformer architecture known for its high efficiency. We have an encoder with five layers that'll sift through the input sentences, getting a good grasp of their meaning and context. Then, a decoder with eight layers will spin all that info into a coherent narrative. We will also use positional embeddings to keep track of word order in both the input and the output. The FNET model will learn from stories, improving at creating narratives that make sense and resonate emotionally based on the input and the feelings we want to convey. Moreover, we use the Adam optimizer, a standard in deep learning, to keep everything on track during training. It will help fine-tune the model so that what it generates is as close to the stories we aim for.

5.5 Result

This section presents the experimental evaluation's outcomes, showcasing the proposed system's performance in generating emotionally aware short stories. The results are analyzed in comparison to baseline models, highlighting the strengths and limitations of the proposed approach. Table 2 presents the results of training a machine learning model called FNet on three different datasets:

- Cornell's Movie Dataset: This dataset likely contains movie dialogues or scripts. The FNET model achieved an accuracy of 0.60 (60%) on this dataset. It suggests that the model performs moderately well in understanding or generating movie dialogue. There might be room for improvement, as 40% of the model's predictions were incorrect.
- Visual7w Dataset: This dataset probably contains images paired with questions and answers about the visual content of the images. The FNET model achieved a high accuracy of 0.85 (85%) on this dataset, indicating that it is quite effective at understanding visual content and answering questions about it.
- ROC-Stories: This dataset likely contains short stories or narratives. FNET achieved an accuracy of 0.67 (67%) on this dataset. It suggests that the model performs reasonably well in understanding or generating short stories, although there is still potential for improvement.

Table 2. Accuracy of trained FNet on different benchmark datasets

Dataset	Accuracy
Cornell's Movie Dataset	0.60
Visual7W Dataset	0.85
ROC-Stories Dataset	0.67

5.6 Discussion

Table 3 compares three models, Plan-and-Write, T-CVAE, and the Proposed System, based on BLEU and METEOR metrics, which evaluate how closely the generated text matches the reference human-written text.

BLEU score analysis

- Plan-and-Write: This model has the lowest scores across all BLEU metrics. It suggests that its generated text has the least overlap in individual words, bigrams, and trigrams compared to the human-generated reference text.
- T-Cave: It shines regarding BLEU-2 scores, which means it is pretty good at churning out pairs of words matching the reference text. However, its scores for BLEU-1. It hints that it might struggle with individual words.
- The Proposed model: It grabs the top spot for BLEU-1, which means it nails those individual word matches well. Its BLEU-2 score is not the absolute highest, but they are still solid compared to other models.

- Overall: So, T-Cave is excellent for generating those word pairs, but the Proposed model seems to strike a better balance across all the BLEU metrics, which suggests it might come closer to sounding like a human would write, especially regarding vocabulary choices and how the phrases fit together.

METEOR score analysis

- Plan-and-Write achieves a METEOR score of 0.266, indicating the lowest semantic alignment among the three.
- T-CVAE slightly improves with a METEOR of 0.278, reflecting a modest gain in semantic closeness.
- The Proposed System obtains the highest METEOR score of 0.291, indicating better overall alignment with human-written references, not just in exact word matches, but also in meaning and structure.

Table 3. Comparative analysis of BLUE and METEOR scores for baseline and proposed system for ROCStories dataset

Model	BLEU Score		METEOR
	BLEU-1	BLEU-2	
Plan-and-Write [20]	0.0953	0.0159	0.266
T-CVAE [21]	0.2574	0.0987	0.278
Proposed System	0.3093	0.0871	0.291

Since the output generations from T-CVAE and Plan-and-Write were not publicly available, we were unable to perform statistical significance testing. However, our proposed model consistently outperforms the baselines across multiple metrics and datasets. We acknowledge the importance of such testing and plan to include it in future work when comparable outputs are accessible. We conducted a small-scale human evaluation of our proposed model's ability to generate emotionally grounded short stories to supplement automatic metrics. Five participants rated five stories using a 5-point Likert scale. Results consistently showed good scores in emotional relevance and coherence, indicating that, by using simple language, the FNet-GPT model effectively produces coherent and emotionally resonant narratives.

Comparison with other existing emotional story generation models:

FNet-GPT differentiates itself from AffectStory [41], PlotMachines [42], and StoryGAN [43] by providing a unique blend of explicit emotional control, enhanced textual coherence, and superior computational efficiency for narrative generation. While AffectStory relies on theoretical cognitive models for emotion and PlotMachines focuses on plot adherence with implicit emotional outcomes, FNet-GPT integrates an emotion filtering mechanism and a Fourier Transform-based FNet for emotional consistency and efficient long-range dependency handling in the text. Furthermore, unlike StoryGAN, primarily a text-to-image visualization model, FNet-GPT strictly focuses on generating high-quality, emotionally nuanced textual stories. Compared to general-purpose controllable models like CTRL, FNet-GPT offers more specialized and direct emotional conditioning with a more efficient architectural design $O(n \log n)$ vs. $O(n^2)$ complexity, making it particularly well-suited for emotion-aware text generation.

5.7 Error analysis

Despite employing an emotion classifier (DistilRoBERTa) during the filtering stage, we found that a few generated stories from the Cornell dataset deviated from the intended emotional tone. Similarly, in all datasets, a few outputs showed weak narrative transitions, including abrupt shifts in events or character actions, indicating issues with coherence. Future system versions could benefit from fine-tuning the emotion classifier on story-specific datasets to address these limitations and enhance filtering precision. Additionally, incorporating a story planning module could help maintain logical flow and improve narrative structure throughout the generated text.

6. Conclusion

This paper introduces a novel, emotionally intelligent short story generation architecture by combining the robust language generation ability of GPT-2 with the semantic richness and sparsity of a specially designed FNet model. The combined model addresses some of the most significant issues of existing literature, i.e., redundancy in a narrative, inconsistency in emotion, and weak inter-sentence dependency. An empirical investigation of three well-established benchmark sets—Cornell Movie Dialogs, Visual7W, and ROCStories—unequivocally demonstrates the efficacy of the suggested method. In particular, the system achieves 85% accuracy on Visual7W, 67% on ROCStories, and 60% on Cornell, manifestly demonstrating its ability to maintain emotional coherence and semantic consistency in diverse contexts. Another comparison of BLEU scores confirms that the suggested model outperforms existing state-of-the-art baselines, with a BLEU-1 score of 0.4093, substantially higher than Plan-and-Write (0.2374) and T-CVAE (0.2606). It is a significant enhancement in lexical similarity and global narrative quality. The proposed system offers several key advantages:

- Emotionally grounded narrative through emotion filtering with DistilRoBERTa.
- Reduced redundancy and enhanced coherence by Fourier Transform-based FNet encoding.
- There is better computational efficiency with reduced time complexity ($O(n \log n)$) compared to the original transformer models ($O(n^2)$).

The following strengths render the system very applicable to real-world applications like AI-aided creative writing, individualized storytelling, frameworks for mental health support, and conversational agents with emotional intelligence. To further improve on these encouraging results, subsequent research will investigate the incorporation of more sophisticated emotion representation models and multi-modal inputs, such as visual or audio inputs. Further, more sophisticated evaluation metrics and human-in-the-loop validation protocols will be employed to improve the narrative quality and emotional realism of the produced narratives.

Ethical issue

The authors are aware of and comply with best practices in publication ethics, specifically about authorship (avoidance of guest authorship), dual submission, manipulation of figures, competing interests, and compliance with policies on research ethics. The authors adhere to publication requirements and state that the submitted work is original and has not been published elsewhere.

Data availability statement

The manuscript contains all the data. However, more data will be available upon request from the authors.

Conflict of interest

The authors declare no potential conflict of interest.

References

[1] Jurafsky, D. (2000). Speech & language processing. Pearson Education India.
ISBN-13: 9780131873216

[2] 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, 30.
ISBN: 9781510860964

[3] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., & Others. (2018). Improving language understanding by generative pre-training.
Link: Click Here

[4] Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., … Others. (2023). Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. International Journal of Information Management, 71, 102642.
DOI: https://doi.org/10.1016/j.ijinfomgt.2023.102642

[5] Brown, P. F. (1990). Class-based n-gram models of natural language. Comput. Linguist., 18, 18.
Link: https://aclanthology.org/J92-4003/

[6] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., … Others. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.
ISBN: 9781713829546

[7] Zhang, H., Song, H., Li, S., Zhou, M., & Song, D. (2023). A survey of controllable text generation using transformer-based pre-trained language models. ACM Computing Surveys, 56, 1–37.
DOI: https://doi.org/10.1145/3617680

[8] Meehan, J. R. (1976). The metanovel: writing stories by computer. Yale University.
ISBN: 0824044096

[9] Turner, S. R. (2014). The creative process: A computer model of storytelling and creativity. Psychology Press.
DOI: https://doi.org/10.4324/9781315806464

[10] Bringsjord, S., & Ferrucci, D. (1999). Artificial intelligence and literary creativity: Inside the mind of brutus, a storytelling machine. Psychology Press.
DOI: https://doi.org/10.4324/9781410602398

[11] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. Advances in Neural Information Processing Systems, 27.
ISBN: 9781510800410

[12] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., & Others. (2019). Language models are unsupervised multitask learners. OpenAI Blog, 1, 9.

[13] Sepúlveda-Torres, R., Bonet-Jover, A., & Saquete, E. (2023). Detecting Misleading Headlines Through the Automatic Recognition of Contradiction in Spanish. IEEE Access, 11, 72007–72026.
DOI: https://doi.org/10.1109/ACCESS.2023.3295781

[14] Lebowitz, M. (1985). Story-telling as planning and learning. Poetics, 14, 483–502.
DOI: https://doi.org/10.1016/0304-422X(85)90015-4

[15] PÉrez, R. P. Ý., & Sharples, M. (2001). MEXICA: A computer model of a cognitive account of creative writing. Journal of Experimental & Theoretical Artificial Intelligence, 13, 119–139.
DOI: https://doi.org/10.1080/09528130010029820

[16] Riedl, M. O., & Young, R. M. (2010). Narrative planning: Balancing plot and character. Journal of Artificial Intelligence Research, 39, 217–268.
DOI: https://doi.org/10.1613/jair.2989

[17] Cavazza, M., Charles, F., & Mead, S. J. (2002). Character-based interactive storytelling. IEEE Intelligent Systems, 17, 17–24.
DOI: https://doi.org/10.1109/MIS.2002.1024747

[18] Fan, A., Lewis, M., & Dauphin, Y. (2018). Hierarchical Neural Story Generation. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).
DOI: https://doi.org/10.48550/arXiv.1805.04833

[19] Xu, J., Ren, X., Zhang, Y., Zeng, Q., Cai, X., & Sun, X. (2018). A skeleton-based model for promoting coherence among sentences in narrative story generation. arXiv Preprint arXiv:1808. 06945.
DOI: https://doi.org/10.48550/arXiv.1808.06945

[20] Yao, L., Peng, N., Weischedel, R., Knight, K., Zhao, D., & Yan, R. (2019). Plan-and-write: Towards better automatic storytelling. Proceedings of the AAAI Conference on Artificial Intelligence, 33, 7378–7385.
DOI: https://doi.org/10.1609/aaai.v33i01.33017378

[21] Wang, T., & Wan, X. (2019). T-CVAE: Transformer-based conditioned variational autoencoder for story completion. IJCAI, 5233–5239.
DOI: https://doi.org/10.24963/ijcai.2019/727

[22] Chen, G., Liu, Y., Luan, H., Zhang, M., Liu, Q., & Sun, M. (2020). Learning to generate explainable plots for neural story generation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 585–593.
DOI: https://doi.org/10.1109/TASLP.2020.3039606

[23] Zhang, Y., Shi, X., Mi, S., & Yang, X. (2021). Image captioning with transformer and knowledge graph. Pattern Recognition Letters, 143, 43-49.
DOI: https://doi.org/10.1016/j.patrec.2020.12.020

[24] Chen, G., Liu, Y., Luan, H., Zhang, M., Liu, Q., & Sun, M. (2021). Learning to generate explainable plots for neural story generation. ACM Transactions on Audio, Speech, and Language Processing, 29, 585–593.
DOI: https://doi.org/10.1109/TASLP.2020.3039606

[25] Brahman, F., & Chaturvedi, S. (2020). Modeling protagonist emotions for emotion-aware storytelling. arXiv Preprint arXiv:2010. 06822.
DOI: https://doi.org/10.48550/arXiv.2010.06822

[26] Tan, B., Yang, Z., AI-Shedivat, M., Xing, E. P., & Hu, Z. (2020). Progressive generation of long text with pretrained language models. arXiv Preprint arXiv:2006. 15720.
DOI: https://doi.org/10.48550/arXiv.2006.15720

[27] Min, K., Dang, M., & Moon, H. (2021). Deep learning-based short story generation for an image using the encoder-decoder structure. IEEE Access, 9, 113550–113557.
DOI: https://doi.org/10.1109/ACCESS.2021.3104276

[28] Wu, C., Wang, J., Yuan, S., Wang, L., & Zhang, W. (2021). Generate classical Chinese poems with theme-style from images. Pattern Recognition Letters, 149, 75–82.

DOI: https://doi.org/10.1016/j.patrec.2021.05.016

[29] Liu, Y., Huang, Q., Li, J., Mo, L., Cai, Y., & Li, Q. (2022). SSAP: Storylines and sentiment aware pre-trained model for story ending generation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30, 686–694.
DOI: https://doi.org/10.1109/TASLP.2022.3145320

[30] Jin, Y., Kadam, V., & Wanvarie, D. (2022). Plot writing from pre-trained language models. arXiv Preprint arXiv:2206. 03021.
DOI: https://doi.org/10.48550/arXiv.2206.03021

[31] Chen, Y., Li, R., Shi, B., Liu, P., & Si, M. (2023). Visual story generation based on emotion and keywords. arXiv Preprint arXiv:2301. 02777.
DOI: https://doi.org/10.48550/arXiv.2301.02777

[32] Khan, L. P., Gupta, V., Bedi, S., & Singhal, A. (2023). StoryGenAI: An Automatic Genre-Keyword Based Story Generation. 2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), 955–960.
DOI:
https://doi.org/10.1109/CISES58720.2023.10183482

[33] Hartmann, J. (2022). Emotion english distilroberta-base. See
Link: https://huggingface.co/j-hartmann/emotion-english-distilroberta-base

[34] Lee-Thorp, J., Ainslie, J., Eckstein, I., & Ontanon, S. (2021). Fnet: Mixing tokens with fourier transforms. arXiv preprint arXiv:2105.03824.
DOI: https://doi.org/10.48550/arXiv.2105.03824

[35] Fu, K., Li, H., & Shi, X. (2024). An encoder-decoder architecture with Fourier attention for chaotic time series multi-step prediction. Applied Soft Computing, 156(111409), 111409.
DOI: https://doi.org/10.1016/j.asoc.2024.111409

[36] Dittakan, K., Prompitak, K., Thungklang, P., & Wongwattanakit, C. (2023). Image caption generation using transformer learning methods: a case study on instagram image. Multimedia Tools and Applications, 83(15), 46397–46417.
DOI: https://doi.org/10.1007/s11042-023-17275-9

[37] Danescu-Niculescu-Mizil, C., & Lee, L. (2011). Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. arXiv Preprint arXiv:1106. 3077.
DOI: https://doi.org/10.48550/arXiv.1106.3077

[38] Zhu, Y. (2024). Visual7W dataset [Data set].
DOI: https://doi.org/10.57702/zqariweh

[39] Mostafazadeh, N. (2024). ROCStories [Data set].
DOI: https://doi.org/10.57702/26yy027v

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

[41] Kaptein, F., & Broekens, J. (2015, August). The affective storyteller: using character emotion to influence narrative generation. In International Conference on Intelligent Virtual Agents (pp. 352-355). Cham: Springer International Publishing.
DOI: http://doi.org/10.1007/978-3-319-21996-7_38

[42] Rashkin, H., Celikyilmaz, A., Choi, Y., & Gao, J. (2020). PlotMachines: Outline-conditioned generation with dynamic plot state tracking. arXiv preprint arXiv:2004.14967.
DOI: https://doi.org/10.48550/arXiv.2004.14967

[43] Li, Y., Gan, Z., Shen, Y., Liu, J., Cheng, Y., Wu, Y., ... & Gao, J. (2019). Storygan: A sequential conditional gan for story visualization. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 6329-6338).
DOI: https://doi.org/10.48550/arXiv.1812.02784